Quantum-Informed Descriptors in Materials Informatics

Dario Massa, M.Sc.

Faculty of Physics

A dissertation submitted to the University of Warsaw
for the degree of Doctor of Philosophy
Under the supervision of

Dr. hab. Stefanos Papanikolaou, prof. NCBJ
Dr. hab. Piotr Sankowski, prof. UW

Warsaw, 2025

Acknowledgments

The results presented here were mainly obtained during my research stay at the NOMATEN Center of Excellence in Multifunctional Materials research, department created in the National Center for Nuclear Research in Otwock, Poland. A portion of the results were obtained during the simultaneous research stay at IDEAS-NCBR, ELLIS (European Laboratory for Learning and Intelligent Systems) research unit in Warsaw, Poland, for collaboration projects with the main institution. The research work was funded:

1. by the European Union Horizon 2020 research and innovation program under Grant Agreement No. 857470 and from the European Regional Development Fund under the program of the Foundation for Polish Science International Research Agenda PLUS, grant No. MAB PLUS/2018/8, and the initiative of the Ministry of Science and Higher Education 'Support for the activities of Centers of Excellence established in Poland under the Horizon 2020 program' under agreement No. MEiN/2023/DIR/3795.

2. by the internship program of IDEAS-NCBR, Warsaw, Poland.

I would like to thank my supervisors, Stefanos Papanikolaou and Piotr Sankowski, for giving me the chance to delve into the exciting and vibrant world opening at the contact points between the Physics of Matter and Machine Learning, where my deepest intellectual curiosities have been nourished. The support of Stefanos has been essential, in particular in shaping the initial idea of the project, and for his willingness and approval to introduce my work to a rich network of scientists, among which are some of the most eminent minds of our research field. In this context, NOMATEN allowed to make each and every opportunity a practical reality, financing my presence and contributions to many international meetings, and I am grateful for this to the former and present Directors. I also thank the NCBJ Director for having allowed my internship at IDEAS-NCBR, where Piotr guided me into advanced machine learning techniques for Physics, and allowed me to have a deep look into the sea of possibilities that the AI revolution is currently opening. Both supervisors patiently drove my efforts in the right directions with their invaluable expertise and, as a Ph.D. candidate, I could not ask for more.

I would also like to thank Prof. Efthimios Kaxiras from Harvard University, Prof. Cesare Franchini from University of Vienna, Prof. Ju Li from MIT and the extraordinary groups of

scientists they are leading, for kindly hosting me during collaboration visits and allowing dynamic scientific exchanges that gave rise to various publications this Thesis is based on.

Finally, my sincere gratitude goes to my wife, for she has always steadied the sails when the winds of life howled fiercest, to my family, for teaching me resilience for my passions, and to my friends, Simone, Alice and Amir, who no matter what, have always been close.

Abstract

Materials Informatics [1] refers to the idea that extensive amounts of data from the fields of Condensed Matter Physics and Materials Science, both experimental and computational, can be used to train machine-learning algorithms for materials design. In particular, examples are direct prediction of properties based on crystalline structures (Crystal Graph Convolutional Neural Networks) [2], improvement of the description accuracy and acceleration of molecular dynamics simulations (e.g., Neural Network Interatomic Potentials) [3], recognition of hidden data patterns in multi-dimensional datasets (Unsupervised Learning) [4], among other optimizations. The key success factor lies in the efficiency of the descriptors used to encode relevant information about properties of the system.

As a first point, in this Thesis, I present the idea of exploiting ab-initio electron charge density profiles from first-principles calculations in Machine Learning (ML) as effective descriptors, starting with the case of study of interstitial defects in bulk crystals [4]. Electron charge density-based descriptors arise from intrinsic interactions, therefore capturing the full landscape of electronic-related behaviors. I first apply this methodology to the context of hydrogen interstitial defects inside FCC crystals and the resulting induced perturbations in electron density fields. Radial Distribution Functions (RDFs) are built in the perturbed electron density fields and compared for a variety of defected FCC metal bulk crystals, exhibiting correlations of their characteristics with an extended dataset of properties, i.e. charge-density properties of hydrogen and metal atoms, structural and atomic properties of the undefected bulk crystals, including large-scale properties like melting point, heat of formation and boiling point. A major finding is the emergence of classes of charge responses, playing a crucial role in hydrite composition search. Through unsupervised learning techniques, I demonstrate the possibility of building density-based composition maps classifying both the charge responses and hydrogen diffusive dynamics. Beyond classification, I also explore how to efficiently deploy electron charge density images in regression tasks, via Deep Learning approaches including multimodal transfer-learning on Electron Charge Density (ECDs) fields from bulk crystals [5]. I demonstrate the highly informative nature of representations extracted from visual (Convolutional Neural Networks on ECDs images) and semantic (Large Language Models on crystallographic textual information)

modalities, exploring a wide set of compositions for the regression of energetics- or structure-related properties. A comparison between a custom solution involving stacked modalities and ready-to-use multimodal models is presented. Such treatment shows potential impact on decision making for new low-cost AI methods in the field of Materials and Chemoinformatics.

As a second point, along the line of direct extraction of atomic representations with Neural Networks, I also investigate the predictive abilities of available foundational graph-based models in Materials Informatics. On the one hand of static structure-property predictions, I propose a substitutional alloying framework to test the transferrability of a pre-trained crystal graph convolutional neural-network model. The question of whether the learned representations of local atomic chemical environments *generalize* towards systems where model-known atomic species are contained in unknown-arrangements (out of training database) is a key question for Materials Discovery tasks, the ultimate scope for which such models have been developed [2]. On the other hand of dynamic simulations, graph-based models for interatomic potential development provide an opportunity to further explore the role of training configurations for accurate predictions. In particular, the importance of dataset building for such potentials is investigated for the modeling of hydrogen diffusion in magnesium through active-learning techniques, with a focus on fine-tuning of pre-trained universal deep-learning potentials [3]. The work reports remarkable agreement with the experimental results for hydrogen diffusion coefficients and transition barriers, at different temperatures and for different potentials.

Streszczenie

Informatyka materiałów [1] odnosi się do koncepcji wykorzystania obszernych ilości danych z dziedziny fizyki materii skondensowanej i nauki o materiałach, zarówno eksperymentalnych, jak w celu obliczeniowych, do trenowania algorytmów uczenia maszynowego w celu projektowania materiałów. Przykładowo, można wskazać bezpośrednie przewidywanie właściwości na podstawie struktur krystalicznych (Crystal Graph Convolutional Neural Networks) [2], poprawę dokładności opisu i przyspieszenie symulacji dynamiki molekularnej (np. Neural Network Interatomic Potentials) [3], rozpoznawanie ukrytych wzorców danych w wielowymiarowych zbiorach danych (Unsupervised Learning) [4], a także inne optymalizacje. Kluczowym czynnikiem sukcesu jest efektywność deskryptorów używanych do kodowania istotnych informacji o właściwościach systemu.

W pierwszym punkcie tej pracy przedstawiam ideę wykorzystania profili gęstości ładunku elektronowego obliczonych metodami pierwszych zasad w Machine Learning (ML) jako efektywnych deskryptorów, zaczynając od przypadku badania defektów międzywęzłowych w kryształach objętościowych [4]. Deskryptory oparte na gęstości ładunku elektronowego wynikają z wewnętrznych interakcji, co pozwala uchwycić pełen obraz zachowań związanych z elektronami. Na początek stosuję tę metodologię w kontekście defektów międzywęzłowych wodoru wewnątrz kryształów FCC i wynikających z nich perturbacji w polach gęstości elektronowej. Radial Distribution Functions (RDF) są tworzone w zaburzonych polach gęstości elektronowej i porównywane dla różnych kryształów metali FCC z defektami, wykazując korelacje ich cech z rozszerzonym zestawem właściwości, tj. właściwościami gęstości ładunku wodoru i atomów metalu, właściwościami strukturalnymi i atomowymi idealnych kryształów objętościowych, w tym właściwościami makroskopowymi, takimi jak temperatura topnienia, entalpia tworzenia i temperatura wrzenia. Kluczowym odkryciem jest wyłonienie się klas odpowiedzi ładunkowych, odgrywających kluczową rolę w poszukiwaniach składu wodorków. Za pomocą technik uczenia nienadzorowanego demonstruję możliwość budowy map składu opartych na gęstości, klasyfikujących zarówno odpowiedzi ładunkowe, jak i dynamikę dyfuzyjną wodoru. Oprócz klasyfikacji eksploruję także sposób efektywnego wykorzystania obrazów gęstości ładunku elektronowego w zadaniach regresji za pomocą podejść głębokiego uczenia, w tym multimodalnego transferu uczenia na

polach gęstości ładunku elektronowego (ECD, ang. electron charge density) w kryształach objętościowych [5]. Demonstruję bardzo informacyjny charakter reprezentacji wyekstrahowanych z modalności wizualnych (splotowe sieci neuronowe na obrazach ECD) i semantycznych (duże modele językowe na tekstowych informacjach krystalograficznych), eksplorując szeroki zestaw składów do regresji właściwości związanych z energetyką lub strukturą. Przedstawiono porównanie między niestandardowym rozwiązaniem złożonym z połączonych modalności a gotowymi modelami wielomodalnymi. Taka analiza ukazuje potencjalny wpływ na podejmowanie decyzji w zakresie nowych wydajnych metod AI w dziedzinie materiałoznawstwa i chemoinformatyki. W drugim punkcie, w ramach bezpośredniego pozyskiwania reprezentacji atomowych za pomocą sieci neuronowych, badam również zdolności predykcyjne dostępnych modeli podstawowych opartych na grafach w informatyce materiałowej. Z jednej strony, w kontekście statycznych przewidywań właściwości struktur, proponuję ramy zastępczego stopowania do testowania transferowalności wstępnie wytrenowanego modelu krystalicznej splotowej sieci neuronowej. Kluczowym pytaniem dla zadania poszukiwania nowych materiałów (głównego celu, dla którego opracowano takie modele) jest to, czy wyuczone reprezentacje lokalnych chemicznych środowisk atomowych *generalizują się* na systemy, w których modelowane gatunki atomowe występują w nieznanych układach (poza bazą danych treningowych) [2]. Z drugiej strony, dynamiczne symulacje i modele oparte na grafach dla rozwoju potencjałów międzyatomowych umożliwiają dalsze badanie wpływu konfiguracji treningowych na dokładne przewidywania. W szczególności badam znaczenie budowy zestawu danych dla takich potencjałów w modelowaniu dyfuzji wodoru w magnezie za pomocą technik aktywnego uczenia, koncentrując się na dostrajaniu wstępnie wytrenowanych uniwersalnych potencjałów głębokiego uczenia [3]. Wyniki pracy wykazują znakomitą zgodność z wynikami eksperymentalnymi dotyczącymi współczynników dyfuzji wodoru i barier przejścia, w różnych temperaturach i dla różnych potencjałów.

Manuscripts composing the Thesis

- Dario Massa, Efthimios Kaxiras, and Stefanos Papanikolaou. "Alloy informatics through ab initio charge density profiles: Case study of hydrogen effects in face-centred cubic crystals", In: Acta Materialia 268 (2024): 119773.
  DOI: 10.1016/j.actamat.2024.119773

- Dario Massa, Grzegorz Kaszuba, Stefanos Papanikolaou and Piotr Sankowski. "Transfer Learning in Materials Informatics: structure-property relationships through minimal but highly informative multimodal input", Accepted in NeurIPS2024 - Machine Learning and the Physical Sciences (2024).
  Public NeurIPS Peer Review Process, Comments and Acceptance: OpenReview
  DOI: arxiv.org/abs/2401.09301

- Dario Massa, Daniel Cieśliński, Amirhossein Naghdi and Stefanos Papanikolaou. "Substitutional alloying using crystal graph neural networks", In: AIP Advances 1 January 2024; 14 (1): 015023
  DOI:10.1063/5.0163765

- Andrea Angeletti, Luca Leoni, Dario Massa, Luca Pasquini, Stefanos Papanikolaou and Cesare Franchini. "Hydrogen Diffusion in Magnesium Using Machine Learning Potentials", In: NPJ Computational Materials (in review) (2024).
  DOI v1: 10.21203/rs.3.rs-4773688/v1
  DOI v2 (after review): 10.48550/arXiv.2407.21088

Contents

Chapter 1

Introduction

The Thesis presents the research work done in the context of Materials Informatics. The title, "*Quantum-Informed Descriptors in Materials Informatics*", underlines the common thread behind the included manuscripts: the quest, development and use of materials descriptors, arising from quantum-level treatment of atomic interactions. While the concept of descriptors will be widely discussed in the next sections, suffice it to say, these are signals (or, in other words, learnable functional forms, fingerprints, representations) widely correlating with the interesting properties of the systems. Their use applies in many contexts, e.g. structure-property predictions, materials classification, characterization of defects, accelerated computational simulations, and, more in general, Materials Discovery. All of these contexts have been explored in this Thesis work, and will be appropriately clarified. The *Informatics* aspect actively operates behind the scenes, as data-driven methods are essential to manage the amount of information resulting from high-throughput quantum simulations, allowing for the construction and use of the descriptors themselves, easing predictions of properties on new data.

In the next subsections of this Introduction, Materials Informatics and various approaches to descriptors modeling for atomistic systems will be first introduced. Density Functional Theory follows in the discussion, being a fundamental instrument in the hands of Computational Materials experts to investigate the properties of atomic systems at the quantum level, and so opening the way to quantum-informed descriptors. Particular focus will be devoted to the use of Electron Charge Densities, as a means to capture and characterize atomic interactions and, therefore, the local atomic chemical environments. In this context, their consideration as efficient quantum level descriptors is proposed. Subsequently, I discuss the direct representations extraction by means of Graph Neural Networks trained on quantum-accuracy data, both for direct regression of crystal properties and development of machine learning interatomic potentials.

Such introductory sections presents the topics both through the example of related works from the literature and from a glance on the included manuscripts, that will be later expanded in detail

during the dedicated commentaries. The Appendix Chpt.(5) supports the discussion with deeper theoretical and practical insights. In Chpt.(2), the main results of the Thesis are presented in a more complete way, with a dedicated commentary for each included manuscript. Chpt.(3) summarizes the research findings under the common thread of quantum-informed descriptors, running through each manuscript, while Chpt.(4) refers the reader to further relevant work performed by the author in the MI context, but with minor contribution.

1.1 Materials Informatics and Descriptors Modeling for Atomistic System

Machine Learning (ML) [6–8] has catalyzed significant advancements in Materials Science, laying the foundation for a new field of computational materials design known as Materials Informatics (MI) [9–17]. This approach, when based on training ML models on large, high-throughput electronic structure databases [18–27], aims to achieve quantum-level prediction accuracy, enabling efficient screening and enhanced material performance in applications spanning multiple technological fields.

Central to MI is the development of effective descriptors that capture essential structural and chemical features of materials. Traditional descriptor models include fixed-length feature vectors of atomic or electronic properties [28–30], with a number of softwares to explore such composition based models [31–34]. Other descriptors have been built on translationally and rotationally invariant functions of atomic coordinates, such as the Coulomb matrix [35,36], atom-centered symmetry functions (ACSFs) [37], Voronoi tasselation [38] and smooth overlap of atomic positions (SOAP) [39], In response, graph neural networks (GNNs) [40] have emerged as a promising approach in Materials Science for learning flexible and powerful representations directly from atomistic data, showing applicability from molecules [41–46] to periodic crystalline structures [47–52]. Once trained on increasingly large datasets, these models can efficiently encapsulate complex atomistic interactions and electronic correlations for a range of materials and structures [53,54]. Allowing a broad spectrum of applications, from direct structure-property prediction models [55–60], to machine-learning force fields development with quantum accuracy [61–66] and accelerated screening of wide composition spaces [67,68], GNNs have surpassed the hand-crafted descriptors solutions [69].

1.2 Quantum-Informed Descriptors

1.2.1 Density Functional Theory and Electron Charge Densities (ECDs)

In Appendix Sec.(5.1), a very detailed introduction to Density Functional Theory (DFT) is reported and advised to the reader who is not familiar with the topic. For this reason, let us here simply state that DFT is a method to compute the ground state energy of a system at the quantum level, accounting for atomic and electronic interactions at different levels of accuracy depending on the chosen level of theory. The method provides an answer to a very non-trivial problem: dealing with a system of N-electrons, the ground-state wavefunction represents an unfeasible target of 3N variables, and the focus is shifted towards the three dimensional ground-state charge density, a much more accessible one which still encodes the properties of the system. It is this crucial role to make of this quantity the ideal candidate for descriptor building, being a natural fingerprint of the fundamental interactions among the atomic constituents.

The range of applications of this method in predicting the properties of materials is remarkably wide, e.g. in surface chemistry and catalysis [70], battery materials [71], magnetic materials [72,73], the study of clusters [74], ionic liquids [75], two dimensional materials [76], superconductivity [77,78] and many others introduced by a number of reviews in the literature [79–83]. A number of DFT codes has been developed and widely used from the Computational Materials community [84–87], and in this Thesis Quantum Espresso (QE) [88–90] and VASP [91–93] have been used.

Appendix Sec.(5.2) explains in details how to extract ECDs with the help of QE and how to post-process the charge information. In the included manuscripts of this Thesis, "*Alloy informatics through ab initio charge density profiles: Case study of hydrogen effects in face-centred cubic crystals*" of Sec.(2.1) and "*Transfer Learning in Materials Informatics: structure-property relationships through minimal but highly informative multimodal input*" of Sec.(2.2), the use of ECDs within ML frameworks is discussed in details. It is of interest for this Section to look at some examples of ECDs from the datasets there proposed. In particular, Fig.(1.1) is a visualization of the differential ECD fields from an FCC supercell of magnesium with an interstitial hydrogen defect.

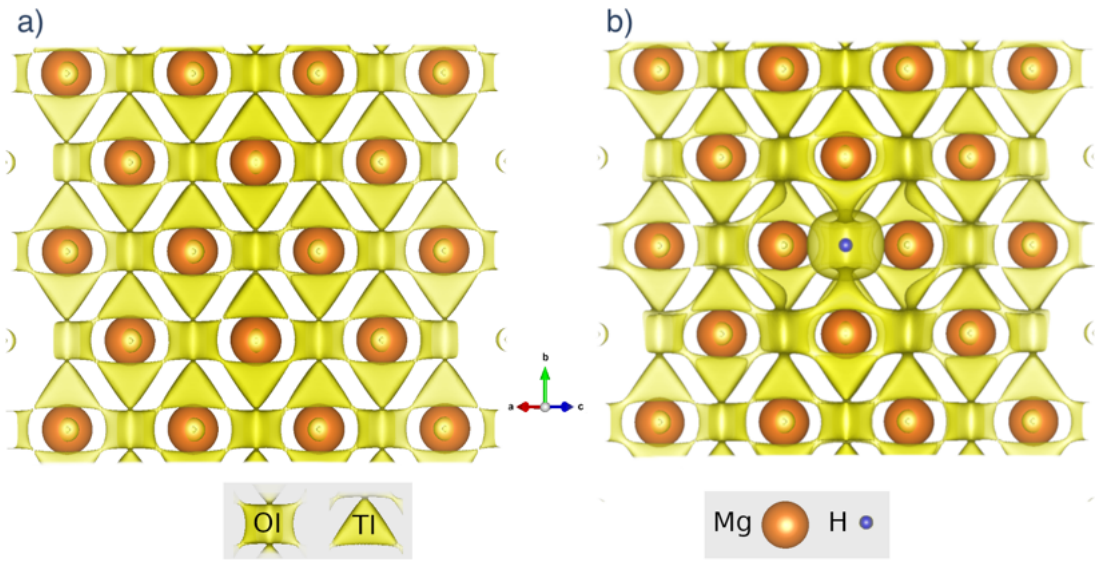


Figure 1.1: Visualization of differential ECD isolines from the manuscript in Sec.(2.1). Panel a) shows the case of a pure Mg supercell, while b) the case of the same supercell with an interstitial H defect. The isoline levels in the ECD fields are selected to be the same.

As it can be observed in panel a), in absence of defects the density isolines highlight specific high-symmetry interstitial regions of electron accumulation, which correspond to the sites where hydrogen most likely sits after structure optimization, namely the octahedral (OI) and tetrahedral interstitial sites (TI). Panel b) instead shows the same isoline levels in the presence of hydrogen, and the resulting perturbation in its neighboring regions. Such perturbations naturally carry information about hydrogen, its interactions in the magnesium environment, and therefore about magnesium as well. In the second proposed manuscript of this context, I exploit an existing database, the Materials Project [22, 94], which includes both total and differential ECDs for a wide variety of single and multicomponent systems, together with the associated properties computed at DFT-level. Fig.(1.2), reports the visualization of an example system from the database.

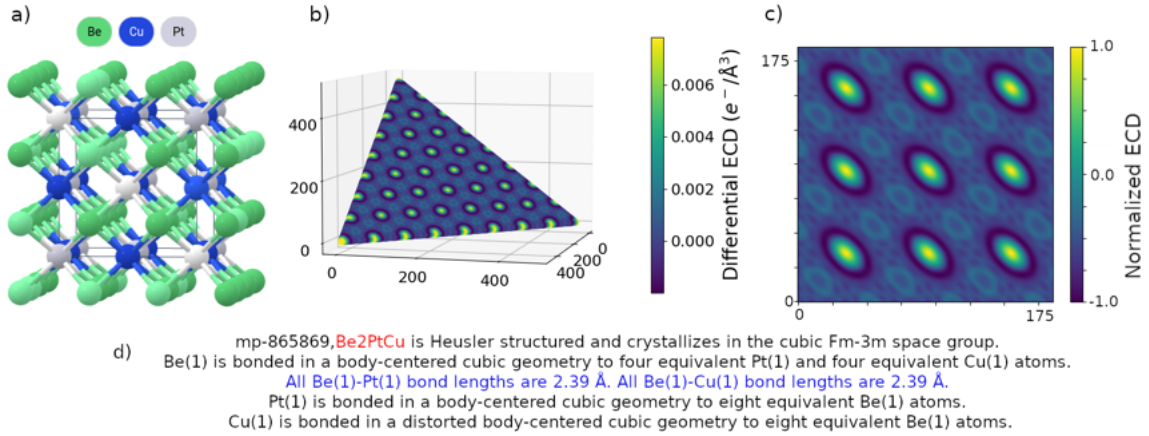


Figure 1.2: a) An example system from the database built in the context of the manuscript presented in Sec.(2.2), Be_2PtCu. b) A visualization of the extracted planar ECD field from the $\langle 111 \rangle$ plane. c) A squared section from the previous plane. d) Crystallographic information about the material: in red, the composition; in blue, the bond lenghts information.

While the full tridimensional field could be used at higher computational cost, I propose a solution to downscale the size of the problem preserving meaningful information: given that the differential ECDs allow to focus on the electronic contributions arising from atomic interactions, and that symmetrical crystals like FCC are characterized by specific high-symmetry planes, one could restrict the information to such planes with largest atomic density and still obtain a comprehensive representation of the local atomic chemical environments.

1.2.2 ECDs in Machine Learning

A brief literature research on the use of ECDs in machine learning reveals how the recent years, e.g. since this Thesis work on the topic was started in early 2022, have seen great simultaneous efforts in the same direction. For example, in a very recent pre-print (2024), Febrer *et. al* introduce Graph2Mat [95], a method to predict electron density fields of molecular systems. The authors show also how integrating the predictions into DFT calculations, as better charge density guess than the initial superposition of isolated atomic charges (see Appendix Sec.(5.1)), the number of iterations for convergence can be reduced up to 40%. On a similar line is a recently published work from T. E. Koker *et al.*, proposing an equivariant graph neural network predicting ECDs with remarkable efficiency and scalability, using higher order tensor representations rather than scalars and vectors. The model seems to outperform other approaches not yet mentioned here [96,97] on different databases [94,98–101]. It is worth mentioning an earlier work (2023) of T. Lv *et al*, presenting a deep-learning model for ECD predictions through a symmetry-preserving method that allows remarkable efficiency in training during a single DFT calculation [102]. The authors applied the method on amorphous Si, an Al surface, an Al-Mg alloy and water, achieving

small errors in comparison to the DFT estimates. The mentioned approaches from the literature all aim at predicting electron densities after having trained deep learning models on the same data structure. While this allows to expand the approaches towards interesting and impactful hybrid DFT implementations, improving computational efficiency and scalability, it might not be considered a direct bridging between the learned ECD-based representations and target materials properties.

This allows to expand, again briefly, the reference made in the previous Section to the ECD datasets built for the investigations in the included manuscripts of Sec.(2.1) and Sec.(2.2), before going into further details during their presentation in the relative Sections. One of the differences in the approach of the first, "*Alloy informatics through ab initio charge density profiles: Case study of hydrogen effects in face- centred cubic crystals*", with respect to the proposed literature works, is that descriptors are built through a post-processing of the perturbed ECD fields of the host crystals in which an interstitial hydrogen defect is inserted. The dimensionality of the signal is reduced, up to a one-dimensional radial distribution function of its values in the matrix, but strong evidence is reported for its correlation with an extensive set of properties of the host crystal. Furthermore, a pure unsupervised learning approach (see Appendix Sec.(5.3.1)) on a dataset of charge density related properties provides a composition map where classification of the defect-host interaction behavior emerges. In the second work included in the Thesis, "*Transfer Learning in Materials Informatics: structure- property relationships through minimal but highly informative multimodal input*", the representations extracted with Convolutional Neural Networks (CNNs) (see Appendix Sec.(5.3.2)) from ECDs images are used for direct regression of scalar target properties, e.g. the bulk modulus, with a focus on adapting open-source pretrained models on such input data, as well as on the possibility to combine visual with semantic representations extracted with Large Language Models (LLMs) (see Appendix Sec.(5.3.3)) from crystallographic information.

1.3 Descriptors from Graphs: Models and Interatomic Potentials

Graphs (G), in their being composed by a set of vertices $v \in V$ and edges $e_{vw} \in E$ connecting them, allow for an intuitive mapping with the atomic realm. Indeed, whether the GNN models are applied for direct structure-property predictions or for the development of machine-learning interatomic potentials, they usually share the following common features:

- nodes represent atoms and edges represent chemical bonds;
- nodes and edges are equipped with attributes vectors encoding relevant information;
- a sequence of update operations, including node update U, message generation M, and final

graph-readout R learnable functions on the graph, which respectively allow an internal flow of information and a final extraction of a graph-level representation.

The last step frames the Message Passing protocol [103], schematized in Fig.(1.3), in which the v-th node-level representation h_v^{t+1} is obtained by the generation of informative messages m_v^{t+1} through edges to neighboring nodes $w \in N(v) = \{z \in V | (v,z) \in E\}$. Gathering all incoming messages, all nodes representations get updated during $t = 1\,..\,C$ cycles, and within such process a representation of the local atomic chemical environment is learned.

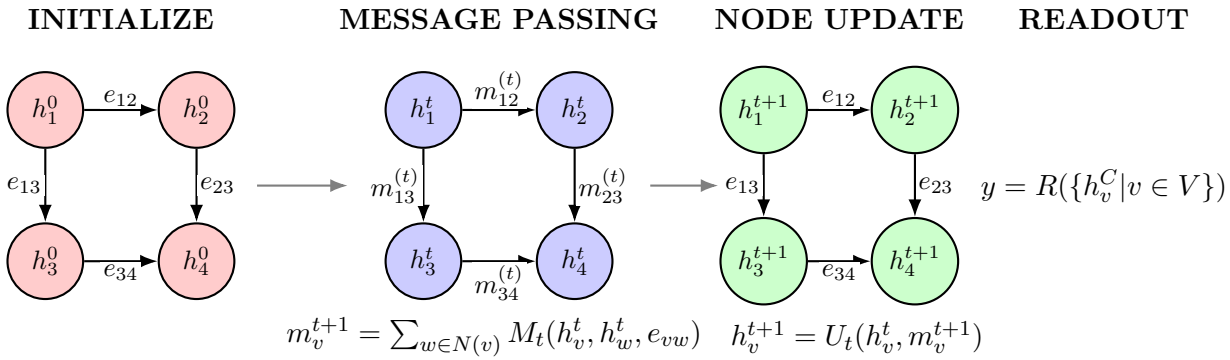


Figure 1.3: Scheme of the Message Passing protocol. The graphs are defined by nodes and edges with their attribute vectors. In order to extract representations of the atoms' local chemical environment, the node information gets updated via informative messages from neighboring nodes for each step of the cycle.

For a complete table of GNN models for structure-property predictions, the review work of Reiser *et al* is recommended [54]. In particular, some of the 3D geometric message passing models include MEGNet [57], DimeNet [60], SchNet [58] and ALIGNN [59].

In Sec.(2), I present the use of one of such models, MEGNet 2, for direct structure-property predictions in substitutionally defected crystals. Functionalization of materials through defecting or alloying is a common practice to alter the properties of the host material, leveraging a wide compositional space of possibilities to which corresponds a wide variability of properties. Such process is certainly part of the MI realm, due to the need of data-driven methods for efficient explorations, and overlaps the so-called Materials Discovery. However, the use of Deep Learning (DL) models leads to a great loss of interpretability and to the risk of not being able to discern the crucial but blurry boundary between interpolation and extrapolation regimes. For this aim, two concepts should be always kept in mind: i) the quality of the training data; ii) the diversity in the distribution, classes, or in other words, composition of the training data samples. In the context of this Thesis, the first point could refer to the level of theory and accuracy with which the data-points have been obtained: training on high-quality DFT data is essential if the aim of the predictions is to be as coherent as possible with such quality. The second point hints instead at the need of curated datasets which include, as much as possible, representative samples for the systems of interest. The problem of substitutional alloying represents a relevant example:

if a GNN-based model has been pre-trained to predict the properties of a wide compositional space of bulk crystals, will it be able to *generalize* to the prediction of properties of those systems with a substitutional defect? In other words, the problem corresponds to test the learned local atomic chemical environment in systems made of known atomic species, in known host crystalline symmetries, with new unknown arrangements. The included paper "*Substitutional alloying using crystal graph neural network*" tries to propose an example protocol to test such generalization capabilities, reporting both encouraging error levels for the prediction of structural properties and the difficulty to accurately predict energy related ones, without a proper fine-tuning on the class of substitutionally defected systems. However, the vibrant MI community has also produced other systematic ways to test GNN models for direct structure-property prediction. An example is the MatBench platform [26], proposing a rigorous set of related ML tasks with the autonomous production of a baseline model.

A similar discussion on the importance of *complete* training databases touches the use of GNN-based models to learn the Potential Energy Surface (PES) for interatomic potential (IPs) development. As nicely explained by T. W. Ko and S. P. Ong [104], graph-based models have a clear advantage on traditional or alternative ML models for IPs, in their ability to deal with arbitrary complex systems in terms of distinct elements belonging to the training structures, allowing the access for wider composition spaces, even if at an intense computational cost. However, proper sampling for dataset building is always a crucial aspect in relation to the prediction tasks to be performed. In Sec.(2.4), I present the study of hydrogen diffusion through various machine-learning interatomic potentials (MLIPs). In particular, the predictions of hydrogen diffusion coefficients and barriers are compared to literature results, both experimental and computational. The potentials compared in such task are: a Bayesian on-the-fly interatomic potential developed with VASP from the authors, the pre-trained universal IPs (UIPs) MACE [105] and CHGNet [106], and their fine-tuned versions on the on-the-fly VASP database associated with the potential development. The study reveals the inadequacy of the pre-trained UIPs, again due to the absence of crucial high-temperature metastable states in their training databases, which are instead explored during an accurate ab-initio MD (AIMD) trajectory. Fine-tuning of the models on AIMD configurations allows to recover DFT accurancy in the estimate of forces and energies, and to reach remarkable agreement with experimental results. While comparing with experiments is always a good practice to test the quality of a model, it must be noted that, similarly to the case of direct structure-property predictions, the MatBench Discovery [107,108] Leaderboard has been designed to assess the performances of available GNN models in Materials Discovery, and allows evaluation on 13 prediction tasks including optical, electronic, thermal, elastic and tensile properties. Beyond the UIPs presented in the included manuscript, a large variety of models has been developed for the same scope, like Orb [109], MatterSim [110] or

M3GNet [111], among others [107].

1.4 Actively Updated Descriptors: Optimization and Active Learning in DFT

In the previous Section, the development of MLIPs with the help of GNNs has been introduced, together with the crucial problem of dataset building ensuring its diversity and quality. One of the answers to such problem is the use of Active Learning (AL) methods, sometimes also called *on-the-fly* trainings. The technique can be generally described through the scheme proposed in Fig.(1.4). With the start of an AIMD trajectory, a database is built with the initially explored structures, including information on their atomic positions, forces, stresses and energies. Such data is used to construct atomic descriptors and train a force field (FF) which will be called to update the positions and velocities, in place of the computationally expensive AIMD, whenever a probability model automatically judges its reliability to be sufficient basing on prediction errors associated to the new structure. A way to achieve such uncertainty estimation has been proposed via query-by-committee [112,113] methods, where the difference in the predictions from a collection of on-the-fly trained neural-network potentials is considered. Such set of predictors usually differs in the random initialization and data splitting for training, and the resulting prediction variance is found to be a good estimate of the prediction error [114,115]. However, such solution is highly demanding from a computational point of view due to the simultaneous generation of multiple models [116]. Other methods use Bayesian inference [117]: a multidimensional Gaussian distribution is assumed for the prior distribution of the PES, and the MLIP[1] uncertainty is evaluated as the variance of the posterior distribution inferred on the basis of the AIMD database. R. Jinnouchi *et al.* [116] explained how such technique allows to skip more than 99% of the AIMD steps, while keeping coherent accuracy.

Through the explained process, the AL methodology enables to focus resources on structures that, along the explored trajectory, are found to be *dissimilar* from the previous ones beyond a certain measure. Indeed, when it happens, the AIMD model is used to evaluate the properties of such structures, which enrich the database and improve, during training, the MLIP.

[1] note that MLIP and MLFF can and will be used interchangeably.

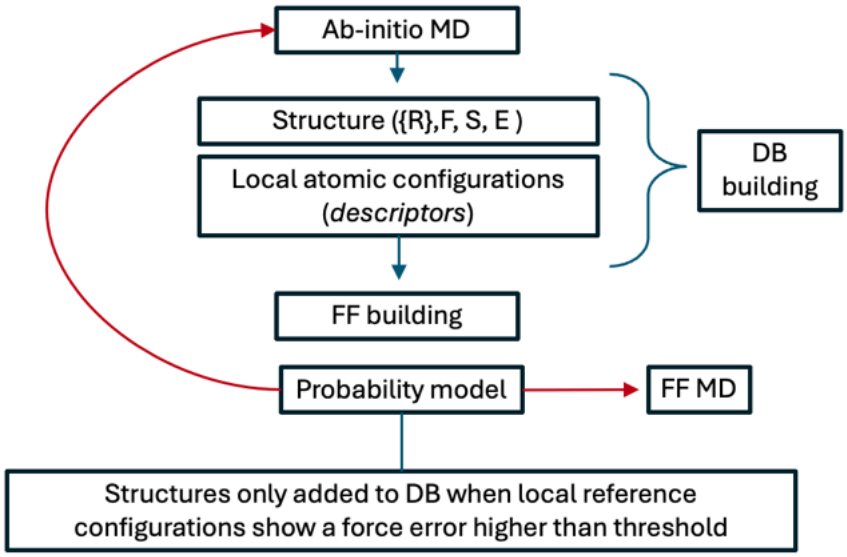


Figure 1.4: Example of Active Learning procedure used in the manuscript "*Hydrogen Diffusion in Magnesium Using Machine Learning Potentials*" presented in Sec(2.4).

In the previous Section, the work from the included manuscript "*Hydrogen Diffusion in Magnesium Using Machine Learning Potentials*" was briefly mentioned, focusing on the use of pre-trained UIPs. The present Section gives the opportunity to focus on another aspect of the manuscript, which is the application of the VASP AL procedure to build a MLFF, and the fine-tuning of the UIPs on the AL database. In Fig.(1.5), a scheme of the on-the-fly training during AIMD trajectories of a $MgH_{0.06}$ system is explained. The MLFF is trained during a three stages temperature ramping from $0K$ to $673K$, with further sampling at interesting target temperatures, within the Bayesian AL framework. While such VASP-MLFF can be used to predict the diffusive characteristics of interstitial hydrogen, its training database is a precious source of diverse configurations that can be exploited to improve existing pre-trained models, like the UIPs. In the proposed work of Sec.(2.4), the fine-tuning procedure of UIPs shows remarkable improvements in the predictions of such models, allowing recovery of ab-initio accuracy and exceptional agreement with experimental results for the hydrogen diffusion coefficients and barriers.

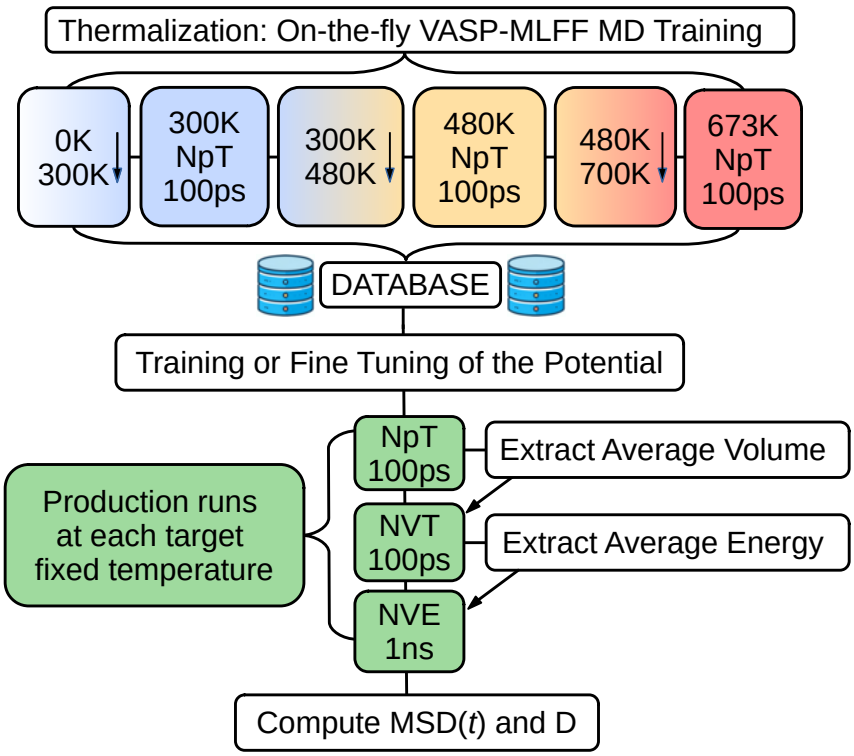


Figure 1.5: Example of application of the Active Learning procedure from the manuscript "*Hydrogen Diffusion in Magnesium Using Machine Learning Potentials*" presented in Sec 2, and deployment of the related databse for finetuning of pre-trained UIPs.

Trying to resume and compare the capabilities of the techniques presented in this Thesis for ML-accelerated MD simulations, one could consider the following:

- Universal Interatomic Potentials:
  - Are usually pre-trained on a rich chemical composition space;
  - Are fast and versatile;
  - Are most probably pre-trained on a limited configurational sampling with respect to the one needed for accurate predictions of defect or high-temperature dynamics;
  - Can be fine-tuned on a properly designed database, including sparse, diverse and relevant configurations from the class of systems to be explored along the target trajectories;
- Interatomic Potentials via Active Learning (on-the-fly optimization and training):
  - Are characterized by an error oriented sampling which allows a natural exploration of relevant states;
  - Are slower, but more accurate than graph-based pre-trained models;
  - Allow for the building of valuable and re-usable on-the-fly databases for the fine-tuning of pre-trained models.

Chapter 2

Main results of the Thesis

The results of the Thesis are presented in four manuscripts, characterized by the common thread of the extraction and use of quantum-informed descriptors in the field of MI and Computational Materials Physics. In particular, the first two presented manuscripts explore the building of efficient bulk crystal representations from their ECDs: in one case, via simple unsupervised learning on charge-related features in the presence of hydrogen defects in a variety of host metallic crystals; in the second case, via multi-modal supervised learning techniques in a variety of stable undefected crystals, through CNNs on their electron charge density images, and LLMs on their crystallographic information. Along the lines of direct extraction of atomic descriptors with DL, the third manuscript tackles the use of pre-trained CGNNs to predict the properties of unseen substitutionally defected crystals. The last manuscript contains the use of well established descriptors to build and deploy MLIPs for the study of defect dynamics.

Manuscript I

The manuscript, titled "*Alloy informatics through ab initio charge density profiles: Case study of hydrogen effects in face-centred cubic crystals*", coauthored by Efthimios Kaxiras (Harvard University) and Stefanos Papanikolaou (NOMATEN) and published in Acta Materialia, presents a novel method to investigate materials properties through electron charge densities. In particular, the method is demonstrated in the context of a variety of bulk FCC crystals with interstitial hydrogen defects, and exploits the defect-induced charge density perturbations as fingerprints for descriptor building. The resulting descriptors are proved to correlate strongly with atomic and structural properties of the host metal crystals, and support the construction of composition maps that help understanding similarities in the charge responses, as well as to hypothesize a connection among the density-based classes and defect migration energy barriers.

Manuscript II

The manuscript, titled "*Transfer Learning in Materials Informatics: structure-property relationships through minimal but highly informative multimodal input*", coauthored by Grzegorz Kaszuba (IDEAS-NCBR, Poznań University of Technology), Stefanos Papanikolaou (NOMATEN) and Piotr Sankowski (University of Warsaw) has been accepted in the notorious peer-reviewed 2024 NeurIPS conference, for the workshop ML4PS - "Machine Learning and the Physical Sciences". Aim of the research is to prove the possibility of extracting descriptors directly from ECDs with DL techniques, and to use them for regression of the associated physical properties. Particular focus is devoted to the use of transfer-learning, adapting and combining pre-trained available image models (Convolutional Neural Networks) and text models (Large Language Models) to relatively small datasets of respectively ECD images and crystallographic information from FCC bulk stable crystals. The work shows promising results towards the prediction of computationally expensive quantities, e.g. bulk moduli, and stands as an example of low-cost AI approaches in MI.

Manuscript III

The manuscript, titled "*Substitutional alloying using crystal graph neural networks*", coauthored by Daniel Cieslinski, Amirhossein Naghdi (NOMATEN) and Stefanos Papanikolaou (NOMATEN) and published in AIP Advances, aims at investigating the capabilities of a CGCNN model in predicting the properties of substitutionally defected bulk crystals, when pre-trained on their respective undefected phases. In this sense, the work proposes a procedure to test and deploy the model transferability for the purpose of materials property functionalization through defects, or, in other words, alloy discovery. Even if with a small DFT validation set, having considered the zero-shot nature (no fine-tuning) of the task, the prediction error on the structural properties (e.g. bulk modulus) shows interesting results, but is visibly worse on energetics.

Manuscript IV

The manuscript, titled "*Hydrogen Diffusion in Magnesium Using Machine Learning Potentials*", coauthored by Andrea Angeletti (University of Vienna), Luca Leoni (University of Bologna), Luca Pasquini (University of Bologna), Stefanos Papanikolaou (NOMATEN), and Cesare Franchini (University of Vienna, University of Bologna) and currently in the last stages of review in NPJ Computational Materials, advances the computational modeling of defects diffusion in bulk crystals using ML-accelerated molecular dynamics (ML-MD), demonstrating its methods in the case of hydrogen in magnesium. In particular, the work highlights the crucial role of tailored datasets construction for training and deployment of such models. By training and

comparing MLIPs, including VASP-MLFF (on-the-fly or active-learning), MACE, and CHGNet (universal pre-trained and our fine-tuned models), we show how the fine-tuning procedure on the VASP-MLFF database adapts UIPs to systems with diffusing defects. Fine-tuned models achieve DFT-level accuracy, exceptionally approaching the experimental temperature dependence of hydrogen diffusion coefficients and transition energy barriers, whereas general-purpose UIPs fall short due to the absence of high-temperature and transition state configurations in their original training databases. Such tailored dataset construction approach offers a reliable and cost-efficient framework for modeling defect dynamics, extending ML applications in Materials Science while maintaining ab-initio accuracy.

2.1 Manuscript I

"Alloy informatics through ab initio charge density profiles: Case study of hydrogen effects in face-centred cubic crystals"

Dario Massa, Efthimios Kaxiras and Stefanos Papanikolaou

Commentary

In Materials- and Cheminformatics, a common key problem is the identification or extraction of efficient descriptors, capable of grasping the essential character of the atomic chemical environment in a wide compositional and configurational space.

This work introduces a predictive framework, termed "Alloy Informatics", based on ML techniques applied to first-principle ECD profiles. Due to their natural encoding of the information about the system and the quantum-level interactions involved in it, charge densities are ideal candidates for descriptors building. By focusing on hydrogen interstitial defects in various face-centered cubic metal crystals, the approach demonstrates the use of defect-induced ECD perturbations (charge responses) as descriptors to sample atomic and structural properties of host materials. Radial Distribution Functions (RDFs) of the charge density responses in different metallic environments represent a fingerprint of the host-defect interaction and, therefore, of the host environment itself as well. For this reason, such signals, along with other charge density properties, e.g. hydrogen Bader charge metrics, strongly correlate with key atomic and structural properties. Unsupervised ML reveals non-local, compositional, and configurational similarities across different crystals, suggesting potential in classifying and predicting defect behaviors. In particular, the study highlights the emergence of response classes, each correlating with the increase or decrease of a given charge-density based property (e.g. valence, or charge extensions). Such composition maps also support a link from static to dynamical properties of the defect, such as hydrogen migration energy barriers, as Nudge-Elastic-Band calculations in DFT prove for outliers in the map. The generality and simplicity of the method provides a versatile tool for materials design, particularly in the context of functionalization through defects, setting a foundation for more advanced materials informatics approaches that could facilitate multiscale modeling and enhance predictive power for catalysis, hydrogen storage, and other applications.

In this study, the PhD candidate: performed high-throughput DFT simulations on HPC cluster with the QE code to extract the structural properties of the involved bulk crystals, the differential ECDs of the systems with and without hydrogen, as well as the transition energy

barriers of hydrogen in two representative samples along different paths with the NEB method; he performed Bader charge analysis on the resulting ECD fields, collecting the relevant charge properties; he designed a Python-based routine to compute the defect-induced perturbations and to characterize them as descriptors, including the complete unsupervised ML analysis after creating a proper database; he prepared all the figures, the manuscript draft, and all its subsequent versions according to the suggestions of the co-authors.

Acta Materialia 268 (2024) 119773

Contents lists available at ScienceDirect

Acta Materialia

journal homepage: www.elsevier.com/locate/actamat

Full length article

Alloy informatics through ab initio charge density profiles: Case study of hydrogen effects in face-centred cubic crystals

Dario Massa [a,b,*], Efthimios Kaxiras [c], Stefanos Papanikolaou [a]

[a] NOMATEN Centre of Excellence, National Centre for Nuclear Research, Warsaw, Poland
[b] IDEAS-NCBR, Warsaw, Poland
[c] Department of Physics, Harvard University, Cambridge, MA, USA

ARTICLE INFO

Keywords:
Material informatics
Ab-initio
Electronic charge density
Machine learning
Density functional theory
DFT

ABSTRACT

Materials design has traditionally evolved through trial-error approaches, mainly due to the non-local relationship between microstructures and properties such as strength and toughness. We propose 'alloy informatics' as a machine learning based prototype predictive approach for alloys and compounds, using electron charge density profiles derived from first-principle calculations. We demonstrate this framework in the case of hydrogen interstitials in face-centred cubic crystals, showing that their differential electron charge density profiles capture crystal properties and defect-crystal interaction properties. Radial Distribution Functions (RDFs) of defect-induced differential charge density perturbations highlight the resulting screening effect, and, together with hydrogen Bader charges, strongly correlate to a large set of atomic properties of the metal species forming the bulk crystal. We observe the spontaneous emergence of classes of charge responses while coarse-graining over crystal compositions. Nudge-Elastic-Band calculations show that RDFs and charge features also connect to hydrogen migration energy barriers between interstitial sites. Unsupervised machine-learning on RDFs supports classification, unveiling compositional and configurational non-localities in the similarities of the perturbed densities. Electron charge density perturbations may be considered as bias-free descriptors for a large variety of defects.

1. Introduction

Hydrogen has, nowadays, a dual importance, both for renewable energy solutions [1,2], as well as for the fundamentals of hydrogen-embrittlement effects [3–7]. Strict criteria for hydrogen storage systems automotive applications [8] on capacities, working temperatures, reversibility of cycles, safety and price require further improvements in the deployed materials, beyond hydrides [9–11] and other solutions. For constrained optimization problems like these, the use of Machine Learning (ML) techniques appears essential for addressing the required complexity and navigating the optimal compositional space.

Indeed, in materials design and development, the role of ML [12–14] methods has been established as crucial in the context of data-driven approaches. These types of methods have evolved into a new field of computational materials design, referred to as Materials Informatics [15–19]. Existing open-access simulation and experiment datasets, have been used to develop ML models and predictions for electrocatalysts materials [20,21], carbon-capture materials [22,23], Metal Oxide Frameworks (MOFs) [24], organic materials [25,26], and many other classes, with the common aim of maximizing the screening efficiency and materials performances. Conventionally, the foremost challenge, besides the presence of an efficient and well-tested ML algorithm, is the specification of the set of effective system descriptors.

In the context of ML approaches with atomistic input, i.e. coming from simulation data, such as density functional theory (DFT) [27–33], a large variety of descriptors have been proposed, including fixed-length feature vectors of elemental or electronic properties [34–36], as well as structural descriptors, based on rotationally and translationally invariant functions of atomic coordinates, like the Coulomb matrix [37], the atom-centred symmetry functions (ACSFs) [38], the social permutation invariant coordinates (SPRINT) [39], the smooth overlap of atomic positions (SOAP) [40] and the global minimum of root mean-square distance [41]. However, given that the number of DFT calculations is naturally limited, the challenge in selecting the appropriate set of descriptors stems from the necessity for them to effectively encompass element-specific microstructural features while also accounting for potential electron–electron correlations across various length scales.

In the search of hydrogen-based energy solutions, as well as in studies of embrittlement effects, hydrogen may be regarded as a defect

* Corresponding author at: NOMATEN Centre of Excellence, National Centre for Nuclear Research, Warsaw, Poland.
E-mail address: dario.massa@ncbj.gov.pl (D. Massa).

https://doi.org/10.1016/j.actamat.2024.119773
Received 31 October 2023; Accepted 16 February 2024
Available online 21 February 2024

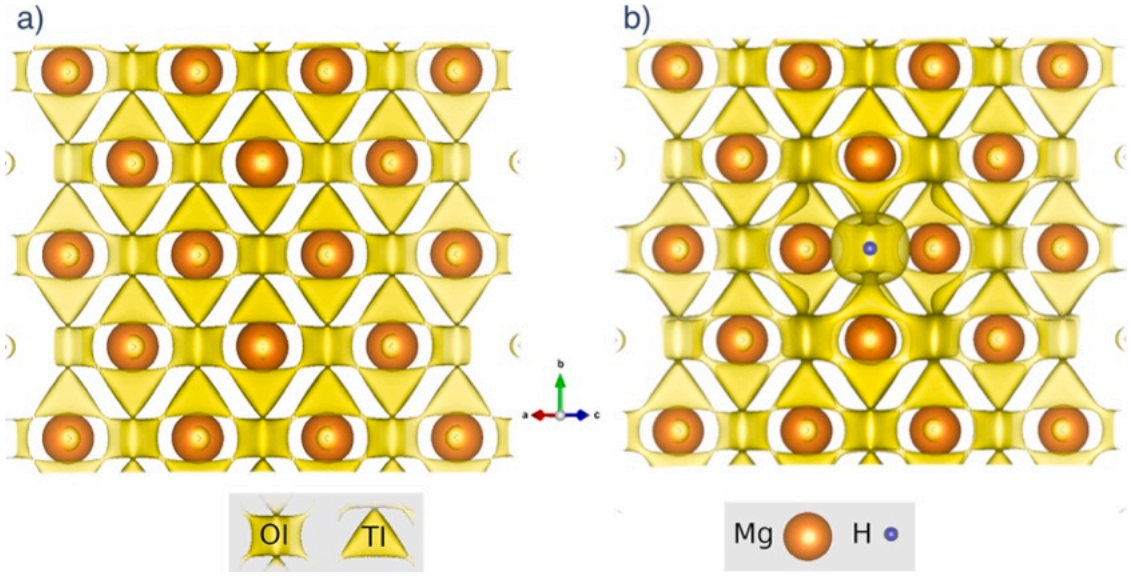


Fig. 1. Three dimensional differential electron charge accumulation isolines in a bulk Mg supercell, (a) in the absence of H and (b) in the presence of an H atom sitting at a high symmetry site. The isoline level, of 0.00183 e/bohr^3 in both visualizations, has been set to visibly highlight the two high-symmetry interstitial regions in which hydrogen could stably sit, namely OI and TI.

inside the host material, thus the classification of fundamental interactions among the two for constructing optimal ML descriptors represents the critical challenge. In the present work, we use the defect-induced electron charge density perturbations in order to illuminate classification features of local crystalline defects and their properties. We consider hydrogen interstitial defects in a variety of bulk face-centred cubic (FCC) metal crystals as a key demonstrative example of a high-throughput procedure [42,43], where our descriptors are constructed on the base of hydrogen differential electron charge density profiles, and of their local quantities, such as radii and extrema. In particular, in Section 3.1 we show how the building of a Radial Distribution Function (RDF) in the density fields allows for a fair comparison between the profiles, and the distinctive features of the effects of hydrogen are discussed in terms of charge density perturbations. In Section 3.3, we prove that the electron density related features show remarkable correlations with both atomic and structural features of the host bulk crystals, and that they can also be exploited to infer hydrogen mobility in the surrounding matrix. After the building of a density dataset, in Section 3.4 we perform unsupervised ML, while in Section 3.2 we investigate over the possibility of a mean-field like treatment of the density perturbations, averaging out compositional peculiarities. Basing on preliminary tests, we discuss interesting trends between the profile features and the properties of the embedding crystals, uncovering compositional relationships (among different crystals) and non-compositional ones (among different hydrogen interstitial equilibrium positions in the same crystals). The results of this study might pave the way to novel material informatics methods for prediction of defects behaviour in materials and their related effects, by interpretation of the perturbed differential charge density profiles (PChDP) in its surroundings.

2. Methods

We focus on the hydrogen-induced changes in the differential [44] electron charge density profile of an otherwise periodic crystal structure. For this aim, we develop 35 simple FCC supercells of pure metallic crystals, listed in Tab.S.(I) of the associated Supplementary Material (SM), in which a hydrogen atom is inserted. An example of such considered structures is contained in Fig. 1.

2.1. Density functional theory methods

We perform Density-Functional-Theory (DFT) simulations in the `QUANTUM ESPRESSO` (QE) package [45–47] of $2 \times 2 \times 2$ supercells of pure FCC metallic bulk crystals, whose list is collected in Tab.S.(I) of SM. Pseudo-Potentials for all involved atomic species are taken to be Ultra-Soft with an additional implementation of the Perdew–Burke–Ernzerhof (PBE) [48] functional.

In order to better model dispersion forces in the possible presence of weakly interacting states, the non-local exchange–correlation density functional rvv10 [49] from Vydro and Van Voorhis has been also implemented. Further, smearing [50] has been introduced to correctly deal with metals, and the calculations have been set to be spin-polarized for possible non-trivial magnetic effects. The convergence with respect to the total energy on the number of k-points, plane-waves cutoff energy and smearing spreading is checked, for each case, after a preliminary variable-cell relaxation of the pure metals, and the final equilibrium bulk structures are then obtained via a further fixed-cell relaxation with the optimal parameters. The common acceptance threshold in the variation of the total energy upon parameter change is set to 10^{-5} Ry. This process allows to ensure the correct optimization of both parameters and structures for each specific case, which might need different optimal parameters sets. The parameter sets for each atomic species are contained in Tab.S.(I) of SM. The forces and total energy convergence thresholds for ionic minimization are set to a common value of 10^{-5}a.u. and 10^{-6}a.u. respectively. The structural properties like the bulk, Young and shear modulii and the Poisson ratios of the pure optimized samples are then extracted via the `THERMO_PW` [51] driver of QE, and are reported in Tab.S.(II) of SM, together with the crystals lattice constants and other relevant structural information resulting from this optimization process.

In the relaxed, converged crystalline supercells, we insert one hydrogen atom, and let a relaxation minimize the forces in the system, confirming the well known two possible equilibrium interstitial positions of hydrogen in FCC structures: either as an octahedral interstitial (OI) or a tetrahedral interstitial (TI) [52].

In Fig. 1, there is a demonstration of differential electron charge density isolines in the absence and presence of H in a Mg supercell. The accumulation isoline level has been set to the same one in both

Table 1
Density-related properties promoted as describing features of hydrogen interaction in different materials. (M) stands for Metal and (H) stands for Hydrogen.

Feature	Description
Ch(M) ex.(e^-)	exchanged charge from the M atom
Ch(H) ex.(e^-)	exchanged charge from the H atom
R(M)(Å)	Bader radius of the M charge
R(H)(Å)	Bader radius of the H charge
V(M)(Å^3)	Bader volume of the M charge
V(H)(Å^3)	Bader volume of the H charge
RDF peak width	half-height width of the RDF peak
RDF MAX	maximum of the RDF peak

cases, and in such a way to visibly highlight the two high-symmetry states interstitial regions in which hydrogen could stably sit.

We extract the PChDP around hydrogen atoms, by subtracting the differential density of pure bulk structures from the ones of defected samples, both for OI and TI cases. At the same time, for all samples and configurations, the electron (pseudo-)charge densities and the all-electron ones are also stored and used in Bader charge analysis [53]; the latter provides useful information regarding Bader charge exchanges, volumes and radii [54]. In Tab.S.(III) and Tab.S.(IV) of SM, these quantities are documented for each studied crystal. In particular, for building a dataset with atomic, structural and density properties, we consider a set of elemental properties from the `Mendeleev` package, a Python resource [55], and we maintain Bader Analysis [53] quantities related to the hydrogen atom and to one of its first nearest neighbour. Technically, in an FCC crystal in which hydrogen lives as OI, there are six nearest neighbours, but the extracted Bader information is nearly identical for symmetry. For simplicity and consistency, once the choice of which metal atom should be selected for its Bader properties is done in the OI case, the same metal atom is chosen for the TI case as well.

In order to compute the hydrogen binding energies (BEs) in different crystals, we consider the following formula

$$E_{\mathrm{BE}} = E_{\mathrm{MH}} - E_{\mathrm{M}} - \frac{E_{\mathrm{H2}}}{2} \tag{1}$$

where E_{MH} is the total energy of the bulk supercell with the hydrogen interstitial atom, E_{M} the total energy of the supercell without hydrogen, and E_{H2} the energy resulting from a fixed-cell relaxation of an hydrogen molecule in a sufficiently large cubic simulation box ($a = 24$Å). The BE is computed with the same QE parameters used for the specific metal bulk system each time in consideration, for computational consistency.

2.2. Isolines and materials' comparisons

In order to access to a more complete description of the effects of a systematic change in composition of the embedding FCC matrices, and the dependence on the different hydrogen interstitial positions, we investigate the contours of the PChDP around hydrogen. All the density grids are brought to an identical resolution by using `mp-pyRho`[56]. Then, identifying a unique radius R in the unit cell to which each value of the density profile can be associated, a Radial Distribution Function (RDF) is calculated [57], representing the basis of our comparison between different profiles. The task is parallelized using the `Joblib` Python package. Finally, shifting the obtained signals by the hydrogen position in the grid is a key step for efficiently comparing the profiles. In Table 1 are reported the relevant density-related properties which have been further considered.

2.3. Validation strategies and procedures

2.3.1. Diffusive properties

The key aspects related to hydrogen for industrial applications, especially for renewable energy such as hydrogen storage or catalysis of energy reactions, ultimately rely on its diffusive characteristics in different materials [58–61]. In order to identify validation routes based on the natural relationship between the extension of a charge profile and its mobility in the surrounding environment, we utilize DFT Nudge-Elastic-Band (NEB) calculations to quantify energy barriers for hydrogen migration along highly-symmetric sites in bulk metals. In particular, $2 \times 2 \times 2$ supercells with one hydrogen atom have been considered as starting and ending images, and the total number of points to discretize the path has been set to 14, with quasi-Newton Broyden optimization scheme and optimization step length of 1.0 a.u.. The threshold controlling the error (the norm of the force orthogonal to the path) in the climbing-image simulation is set to 0.05 eV/Å.

2.3.2. Size effects in charge density calculations

The charge density perturbations due to an interstitial hydrogen defect shall display size effects, that may gain crucial importance for the possibility of undesired artifacts from spurious interactions in small supercells. As sanity checks, in Fig.S.(1) and Tab.S.(V) of the SM we report the comparison of a three dimensional perturbation induced by an hydrogen interstitial defect for a $2 \times 2 \times 2$ and $3 \times 3 \times 3$ Mo supercell, proving that in typical metals, and thus, for the aims of this study, the expected size effects are negligible.

2.3.3. Effects of different crystalline symmetries in charge density considerations

For the applicability of the proposed method, we mainly consider the important example of FCC pure bulk crystals, materials that are predominantly metals. Clearly, not all the considered elements form FCC crystals at Standard Temperature and Pressure (STP) conditions, but we opted for a single-phase study for the sake of a simpler comparison among the profiles, avoiding possible differences arising from different crystalline backgrounds and interstitial positions which could complicate the origin of charge density fluctuations. However, the methods discussed in this manuscript are general and shall become applicable to other crystalline structures. For this purpose, the SM contain examples of hydrogen-induced PChDP for non-metallic HCP supercells, including C, Si and Hg, and we promote specific relevant considerations and comparisons with the results presented in the main text.

2.4. Machine learning

2.4.1. Principal Component Analysis (PCA)

s In the present work, we follow the Principal Component Analysis [62,63] method, probably the most popular choice for developing unsupervised learning, allowing the simultaneous preservation of as much information (variability) as possible. Let us consider a dataset represented by a $n \times p$ matrix, $\mathbf{D}$, where n is the number of samples and p is the number of features in the dataset. If we define a linear combination of features as

$$\sum_{j=1}^{p} c_j \mathbf{d}_j = \mathbf{Dc} \tag{2}$$

where $\mathbf{c}$ is a p-dimensional vector of constants, the aim is to search the combination such that its variance is maximized, which means maximizing the quadratic form

$$var(\mathbf{Dc}) = \mathbf{c}^T \mathbf{Sc} \tag{3}$$

where $\mathbf{S}$ is the $p \times p$ covariance matrix. Requiring with a Lagrange multiplier the vectors of constants to have unit norm, the actual problem to solve becomes the following

$$max\left[\mathbf{c}^T \mathbf{Sc} - \lambda(\mathbf{c}^T \mathbf{c} - 1)\right] \tag{4}$$

and the consequent differentiation with respect to the vector of constants leads to an eigenvalue problem which is at the base of PCA

$$\mathbf{Sc} = \lambda \mathbf{c} \tag{5}$$

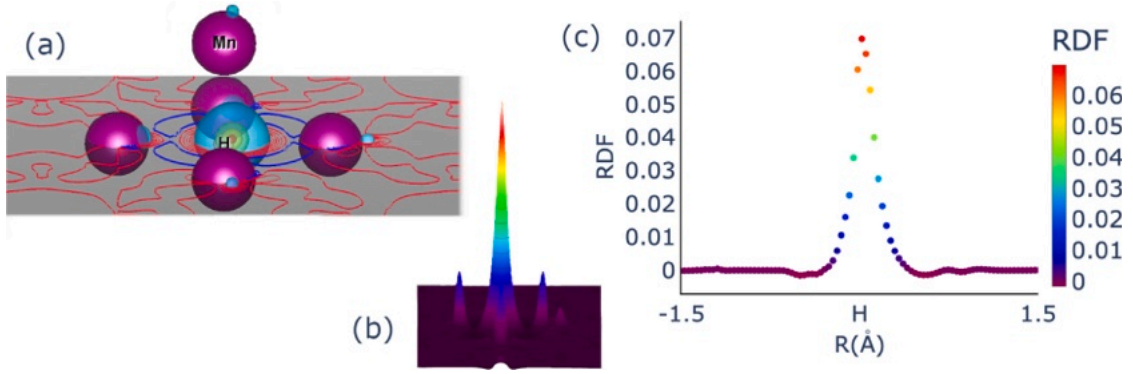


Fig. 2. (a) Three dimensional visualization of the PChDP of Mn (OI case), resulting from the subtraction of the differential profiles of the pure crystals from the defected ones; multiple 3d-isolines are highlighted, starting from the highly positive (accumulation) in red (0.07 e/bohr^3), to positive in yellow and green (0.05 and 0.03 e/bohr^3 respectively), to faintly positive, in blue (0.009 e/bohr^3). On the highlighted grey sectional plane, some accumulation (red) and depletion (blue) isolines are represented, and (b) displays the corresponding bird-eye visualization of the perturbation, with a colour gradient for the charge value mapping the isolines values displayed in the 3D visualization panel, ranging from −0.0043 e/bohr^3 to 0.0759 e/bohr^3, and with out-of-plane extension basing on the intensity. (c) The RDF profile obtained from the PChDP. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Therefore, λ are the eigenvalues of the covariance matrix **S** corresponding to the unit norm eigenvectors **c**. Being **S** a squared real symmetric matrix, it has exactly p eigenvalues, corresponding to eigenvectors $\mathbf{a}_k$ that are the solution of the problem of finding p new linear uncorrelated combinations (Principal Components, PCs) $d_{\mathbf{a}_k} = \sum_{j=1}^{p} a_{jk} d_j$ that maximize the variance [64]. In a geometric interpretation of PCA, we can interpret the eigenvalues of the covariance matrix as associated to the magnitude of the variance captured by the corresponding PC (axis) in the multivariate cloud space, and the eigenvectors as the orientation of the PC (axis). The so-called loadings, eigenvectors scaled up by variances, are the carriers of the combined information on the captured variance and direction of the PCs, and, representing the covariances between the latter and the original dataset variables, tell us how much of the variation in a variable is explained by the component.

2.4.2. K-Means clustering

Unsupervised learning can be accomplished by combining PCA and K-Means clustering. The K-Means algorithm was proposed by Lloyd in 1982 [65] to solve the problem of finding the minimum total squared distance of each point $x \in X \subset \mathbb{R}^d$ from its closest centre $m \in M$ of cluster $c \in C$.

$$\phi_X = \sum_{x \in X} \min_{m \in M} \|x - m\|^2 \tag{6}$$

A survey from ten years ago [66] reports it already was, and probably still is, the most used clustering algorithm. The problem is solved in the following way:

- Define arbitrary (typically random) centres $m_1^{(t=0)} \dots m_k^{(t=0)}$;
- Define a cluster c_i as the set of points closer to m_i than m_j, for each $j \neq i$;
- Define the new k-centres (k-means) as $m_i^{(t+1)} = \frac{1}{|c_i|} \sum_{x \in c_i} x$
- Repeat the previous two steps until the set of means is converged

Even though the target total squared distance is a monotonic decreasing function of the number of clusters and, since there are k^n possible clusters, the algorithm is proved to converge, bad clustering can be obtained without a proper initialization. For this reason, Arthur and Vassilvitski introduced a seeding algorithm [67], `K-Means++`, able to outperform the results of usual `K-Means` just by weighting the selected data-points according to their squared distance from the already randomly selected centre.

A part from a proper initialization, whenever no ground truth is available about the number of cluster in which the data is expected to classify in, there is the need of finding the optimal number of clusters. A variety of methods is available in order to perform the cluster number analysis, and we here focus on the Elbow analysis [68], based on the sum of squared distances [69] in Eq. (6): when its decrease with the number of clusters starts to be negligible, the optimal number is found, above which overfitting occurs.

3. Results

3.1. The RDFs

In order to better explain the building and meaning of the OI and TI RDF profiles, let us focus on one specific example, the case of an hydrogen atom as an octahedral interstitial defect in a Mn supercell, shown in Fig. 2.

The PChDP isolines can be visualized in the three dimensional structure of the supercell, as in panel (a): the PChDP has its positive maximum (accumulation) close to the hydrogen atom, and fades out to a zero level oscillation with increasing distance from it, as the progressively weaker isolines report. In panel (b) we remove one dimension from the visualization, in order to appreciate the perturbation on a plane that crosses the positive maximum. The colour gradient for the charge value is mapping the isolines values displayed in the 3D visualization panel, but is also strictly related to the RDF peak height of panel (c): it leads to the removal of the orienteering dependence and is built as explained in Section 2. If one now repeats the extraction for both the OI and TI cases of different samples, as described in Section 2.2, obtains the result shown in Fig.S.(7) of the SM. For completeness, some relevant quantities from the plots are reported in Table(III) and Table(IV) of SM, which include peaks heights and widths.

It is important to notice how even at this level, without the support of any ML model, one is able to make some compositional-based comparison between the investigated cases, by looking at the peaks properties. An intuitive but strong relation providing a first insight on the presented curves could be found in the charge densities themselves. If one considers the obtained maxima values for the RDF profiles and the properties of the charge density profiles coming from Bader charge analysis, as in Fig. 3, good linear relations are found, especially with the hydrogen Bader charges. The finding suggests the tendency of having higher RDF peaks for larger accumulated charges on hydrogen, which stands as reasonable thinking about the RDF perturbation as a representative of the charge density response (screening) to the added impurity, and given that the RDF profiles themselves are built on the hydrogen PChDPs. However, a clearer picture requires the exploration of correlations with the entire spectrum of atomic-structural and density properties, as will be proposed in Section 3.3. Here we would like to stress that the RDF itself is providing an isotropic measure

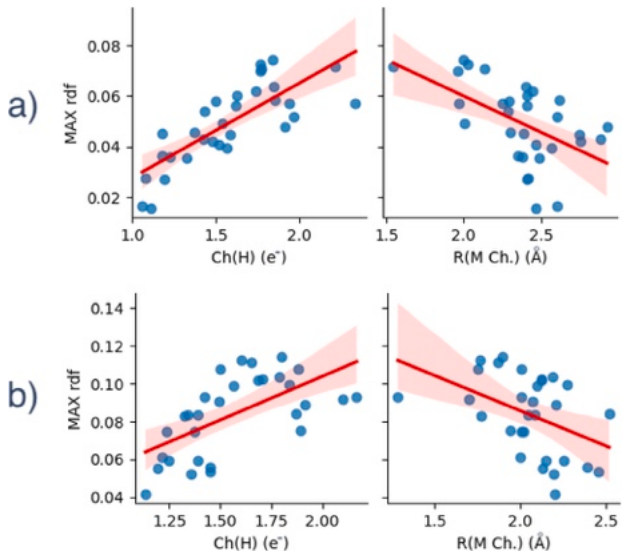


Fig. 3. Linear fits and 95% confidence regression intervals on the relation among RDF peaks properties (MAX rdf) and the charge density ones of the hydrogen (H) and metal atoms (M), both for OI (a) and TI (b) configurations.

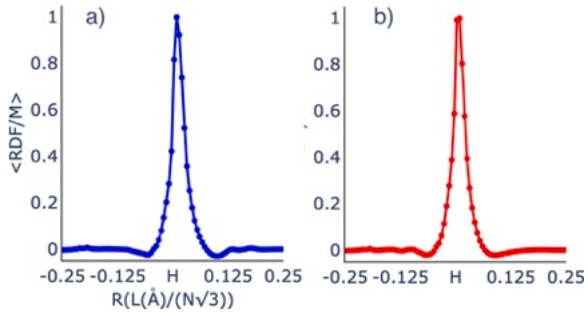


Fig. 4. RDFs averaged over all the available configurations (Eq. (7)), for both OI (a) and TI (b) states. The axes share the same labels.

of the PChDP in the crystal due to the insertion of the H atom, and its comparison for different metallic crystals can clearly provide a measure of similarity of the charge response. For this reason, it can be considered as the descriptor of the perturbed profiles, being a unique signal and fingerprint of the defect-crystal interaction.

In the next paragraph, we question the possibility of a unique average descriptor, on the basis of which a first classification of the system-specific defect-induced perturbations can be demonstrated.

3.2. Composition specific fluctuations from an average oscillation

It is of interest the possibility of considering a H-*universal* RDF curve for our problem, capable of grasping the average behaviour of all the system specific charge density responses that we have observed, assuming crystal composition as a degree of freedom. In order to do so, one could consider the average over N different bulk crystals

$$\langle \frac{\mathrm{RDF}_i}{M_i} \rangle \Big|_{i=1..N} \quad , \quad M_i = max(\mathrm{RDF}_i) \tag{7}$$

which is proposed in Fig. 4. The resulting averaged curves of OI and TI cases are very similar, a part from an inverted symmetry noticeable in the short-range rippling of the profiles.

In order to be able now to quantify the capabilities of a coarse-grained perturbation of grasping the peculiarities of the system specific ones, let us consider the deviations from the average

$$\frac{\mathrm{RDF}_i}{M_i} - \langle \frac{\mathrm{RDF}_i}{M_i} \rangle \Big|_{i=1..N} \quad , \quad M_i = max(\mathrm{RDF}_i) \tag{8}$$

which are reported in Fig. 5, both for OI and TI states.

Interestingly, again without any ML support, the deviations can be easily grouped in three behavioural classes basing on the minima and maxima positions, which also show a conservation in their overall characteristics going from the OI to TI state, a part, as for the averaged RDFs, from an inverted symmetry along the axis of distances from their centre (hydrogen). As it can be noticed from the periodic table plots, the majority of the group members are conserved going from the OI to TI state, except for few changes involving Na, Mo, Rh, In, Sn, Os and Ir.

Going from Group 1 to Group 2, which together include almost all the non-noble metals, the local deviations from the average charge perturbation gain two global minima which are almost symmetric with respect to their centre: thinking in terms of the original characteristics of the signals appearing in the operation of Eq. (7), this should imply a change in the width of the crystal specific one, with respect to the averaged perturbation. Interestingly, we notice how noble-like metals, namely Au, Pd, Pt, Rh and Ir, are classified inside Group 3 (OI), and are characterized by a considerable increase in the number of maxima and minima of the presented deviations. This special class of metals is generally known for their stability, resistance to corrosion and oxidation in moist air, as well as being effective catalysts for the hydrogen evolution reaction (HER) [70]. Other transition metals from Group 2, such as Ti, V and Cr, when combined to form an high-entropy alloy (HEA), have been proved to absorb hydrogen at room temperature and with moderate equilibrium pressures for the dihydride formation [71].

Even though verifying the actual catalytic activity or hydrogen storage performances of the proposed crystals, or their combinations, in the presence of hydrogen molecules is beyond the scope of the present work, we believe these observations inspire ourselves, and possibly the readers, for further work in this direction. It is worth to point out that the proposed element-specific deviations are a sophisticated form of characterization of the metal-hydrogen interaction, inheriting the richness of the presented descriptors, and delivering a bias-free classification of the local peculiarities of the charge density responses due to characteristic atomic (therefore, electronic) environments. One may argue the inter-classes variations in the number of stationary points to be related to a change in the energy of the electrons providing the screening of the impurity: a longer rippling profile shall be associated to lower energies, and indeed, as we will show in the next sections, the accumulated charge on H is significantly minimized in the noble metals.

In the next paragraph, the capabilities of our descriptors to correlate with the physical properties of the systems are quantitatively investigated, and will shed light over the intertwined relations driving the presented classification. In Section 3.4, an alternative and more traditional classification method will be presented basing on a set of density-related features, investigating their actual influence on the samples distribution.

3.3. The density dataset: correlations

In order to prove the key role of the PChDP as effective descriptors, a dataset has been built on the base of the density information, from the Bader charge and RDF analysis, as well as on the structural and atomic one, collected for each metallic crystal here considered. The Pearson correlation matrix for the dataset can guide our attention on the most relevant features, and, given that the relationships between atomic and structural properties are well known, in Fig. 6 we focus on some specific sub-blocks.

Let us focus first on the density related features. Some expected behaviour is obtained: (i) M bader charge and H bader charge exchanges show a remarkable positive correlation (0.7 in scale), implying the exchange from M to H (Tab.S.(III,IV) in S.M.) (ii) the larger is the number of valence electrons of the metals (M), the smaller the charge

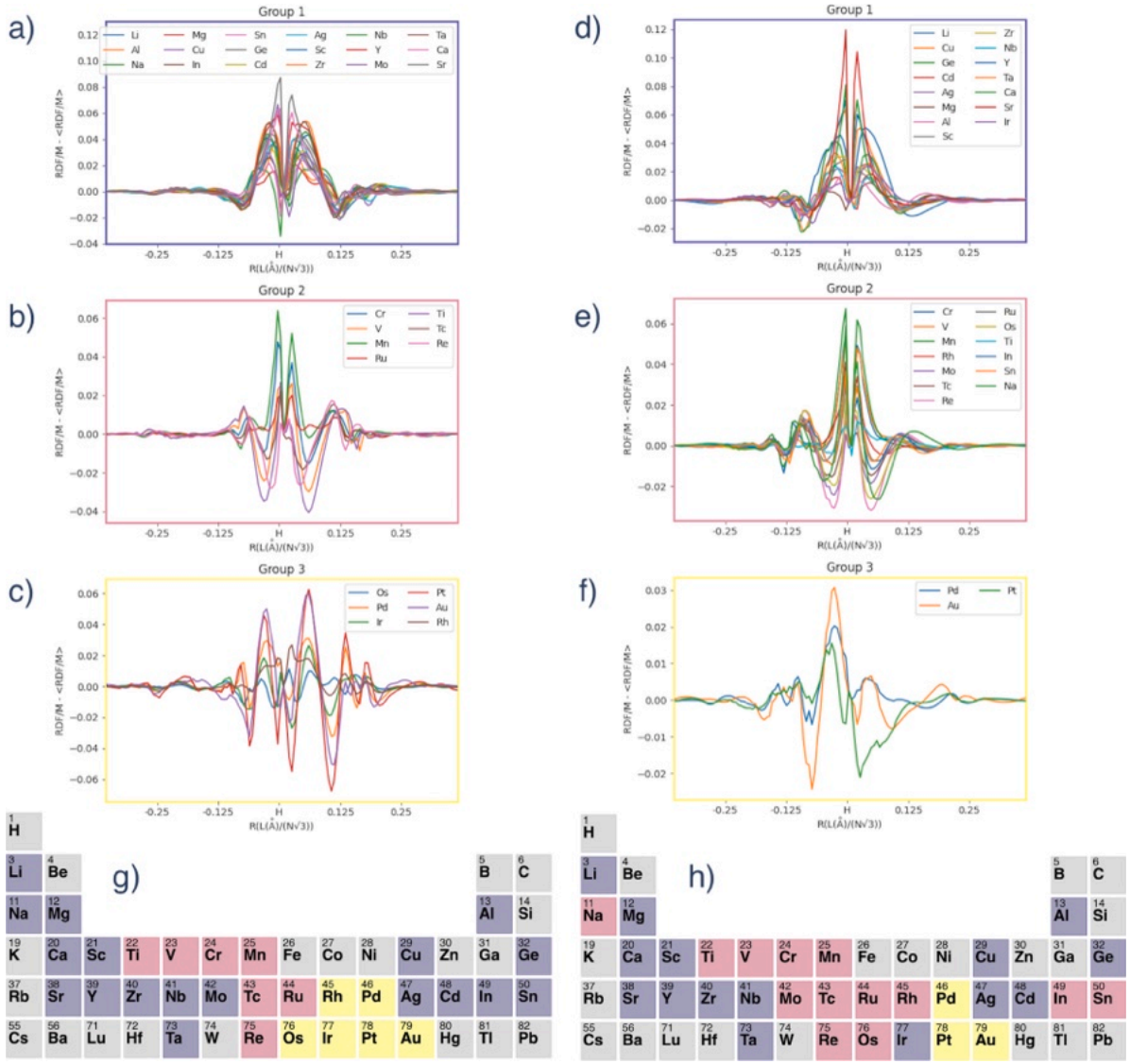


Fig. 5. Crystal-specific oscillations from the rescaled and averaged RDF (Eq. (8)), both for OI (a: Group1, b: Group2, c: Group3) and TI (d: Group1, e: Group2, f: Group3) states. The three behavioural groups are highlighted by a coloured plot frame, which helps highlighting them in the periodic tables below (g: OI, h: TI). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

it exchanges with (to) hydrogen (H); (iii) the two definitions of bader charge extension in terms of radius and volume are largely correlated; (iv) low correlation is present between valence and radius of the M charges, due the expected correlation of the latter with the definition of the atomic radius [72], and therefore the periodic behaviour along the periodic table. Moreover, the maximum value and the width of the computed RDF peaks in the grid of the hydrogen (H) PChDP is strongly anti-correlated with the Bader charge extension of the metal (M) both in terms of radius [73] (−0.6 and −0.5 respectively in scale) and volume (−0.5 and −0.7 respectively in scale), and hydrogen charges properties are affecting the RDF peaks in an overall similar way. From the similar behaviour of M and H charge extensions, the similar influence of the latter on the RDF peaks properties can now be understood.

Let us now consider the correlations between the density related features and the atomic features of the involved metals. Interestingly, the RDF peaks widths show strong anti-correlations with the metal atomic radius, atomic volume and any other definition of bonding radius, and the reason lies in the close relationship that the RDF peaks have with the metal and hydrogen charge features: the bader charge extensions of the metals are fully correlated with the definitions of atomic volume and radius, and very tight relationships also hold with hydrogen charge extensions.

Coming back to Fig. 5, the reader can now understand the sensitivity of the descriptors properties to changes in the host crystalline matrix composition, which is at the base of the previously emerged classification: for example, discrepancies in widths between crystal-specific and a-specific perturbations (averaged) drive the oscillatory behaviour of their difference, and give meaning both to an average screening, as the result of average atomic properties, as well to interstitial hydrogen as a probe of the composition-specific properties, through the induced electronic perturbation. It is also interesting to notice how non-negligible correlations are also found with larger-scale properties like melting points, heat of formation, boiling point.

Concerning the correlations between the density related features and the structural ones [74], again the PChDP information grasps very efficiently the properties of the host crystals: the Poisson ratio and shear modulus are both encoded in the assumed hydrogen bader charge with a strong anti-correlation (−0.6 in scale), while the RDF peaks properties are closely related to the lattice constants. Visualizations of the discussed intertwined relations of atomic and structural properties with the density descriptors are shown in the example of Fig. 7.

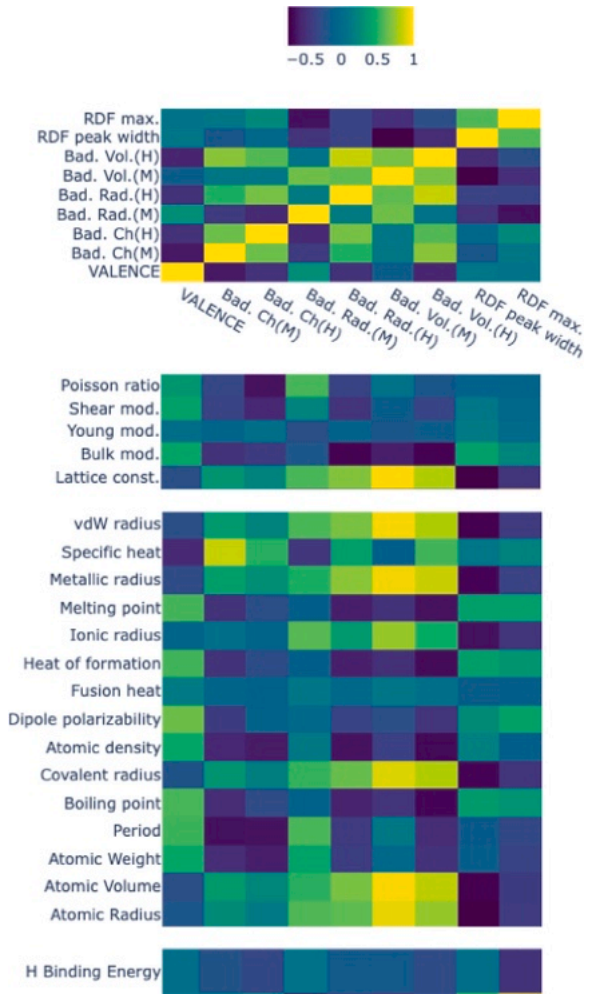


Fig. 6. Correlation matrix between the density-related features (x-axis) with themselves (higher y-block), with the structural (second y-block) and atomic (third y-block) features, and with the computed hydrogen binding energy (lower y-block).

The remarkable correlations that Fig. 6 highlights among the hydrogen BEs and charge related properties, like the RDFs maxima or the hydrogen Bader chargers, have led the focus over this quantity and its variations among the different crystals, and corroborates the thesis of effectiveness of the proposed electron density fields as descriptors of the system interactions.

Fig. 8 reports the computed BE values on periodic tables for both OI and TI states. Overall, a part from few outliers, the binding tends to gradually increase along the periods, with negative values among the alkali, alkaline earth and early transition-metals, where it crosses the zero level to become positive among the late transition-metals, increasing up to the crystallogens. It is to notice that the possibility of obtaining composition maps that could drive the design of systems basing on key properties like the hydrogen binding energy is highly desirable in the context of materials for hydrogen storage: the U.S. D.O.E. has fixed indeed a specific range of interest for this value [75] in hydrogen applications. However, a reliable descriptor is needed to capture the composition induced variations, basing on its underlying power to encode local environments and interactions. In Fig.S.(8) of SM, we propose a visualization of the relation between the hydrogen Bader charges and its binding energy, which is found to be almost linear.

As a last point of this discussion, we want to prove that charge density can help in directly accessing also to the diffusion properties of the hydrogen atom in the different crystals, by looking for a relationship between its migration barriers and charge spatial extensions. In order to prove it, following the procedure explained in Section 2.3.1, Nudge-Elastic-Band calculations have been applied for two cases: Al (high volume of hydrogen Bader charge), Mn (low volume of hydrogen Bader charge).

Table 2
Results from the Nudge-Elastic-Band analysis of the energy barrier for the hydrogen migration along a path in different crystals. For the OI-TI path, two values are present because the barrier evaluation is different depending on whether is computed relatively to the OI (first value) or TI (second value) extremes of the path.

	OI-OI E(eV)	OI-TI E(eV)
Al	0.369	0.085 , 0.122
Mn	1.241	0.701 , 0.291

The results in Fig. 9 and Table 2 show a much smaller energy barrier for the high charge volume case (in Al) with respect to the low charge volume one (in Mn). The full migration energy barrier profiles along the proposed paths also show a similar behaviour in the transition states: the favourite path is OI-TI, as also highlighted in other works for the case of Al [76]. The results also lead to an interesting comparison with a specific result in the literature, which, to our opinion, makes both findings more complete: L. Messina et al. [77] have taken into account a complementary approach for the investigation of the diffusion of impurities in bulk FCC iron; instead of systematically changing the species forming the pure FCC bulk while keeping the same impurity, as done in the present work, they have been keeping the same crystal, systematically varying the interstitial impurity. Even if hydrogen was not involved, among their findings is a counter-intuitive but largely justified correlation between the migration barriers of a solute in iron and the solute compressibility: the larger solutes diffuse faster because of lower migration barriers (a part from P and Si, in their study); moreover, the migration barriers (diffusion coefficients) are displaying a parabolic trend for 4d and 5d elements and an M-shaped (W-shaped) trend for the 3d ones.

In the present problem of systematically changing the hosting matrix around interstitial hydrogen, the extracted density information (i.e. its Bader charge volume) is closely related to the mobility of the interstitial in the system and to the compressibility of the latter, in its pure conditions: the larger volumes (smaller transition barriers) are found for larger compressibilities of the pure hosting crystals, in systems where the number of valence electrons of the metals tend to be smaller.

After having assessed the strong descriptive power of the electron density features in grasping a wide spectrum of properties of the systems of interest, in the next section we restrict to the former features and apply unsupervised ML techniques to test their ability to uncover hidden patterns and behavioural classes.

3.4. Unsupervised Machine Learning on a pure density dataset

Let us consider the application of unsupervised ML techniques to the previously presented dataset, in order to possibly reveal hidden patterns in the data that can lead to deeper understanding and characterization. At this stage, we report the result of the analysis on the pure density features dataset. As explained in Section 2.4.1, PCA is a powerful procedure to reduce the dimensionality of the dataset down to its most relevant components. The histogram in Fig.S.(2) of SM provides the explained variance ratios by each of the resulting principal components: the first two, are able to account for around 70% of the total variance, but, being the result of a transformation of the original dataset components, the individual role of the original features is not clear at this stage.

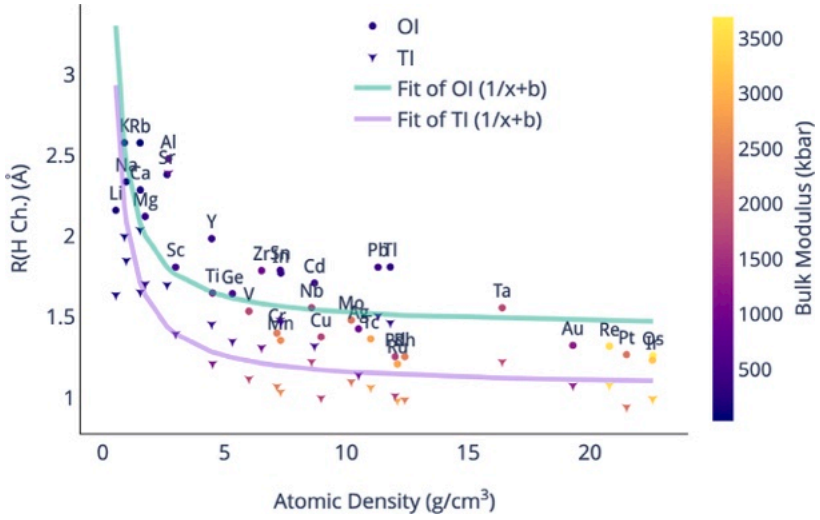


Fig. 7. Plots of specific relationships involving atomic (atomic density), structural (bulk modulus) and electron density (H Bader radius) features for both OI and TI samples, with the superimposed visualization of fitted trend-lines.

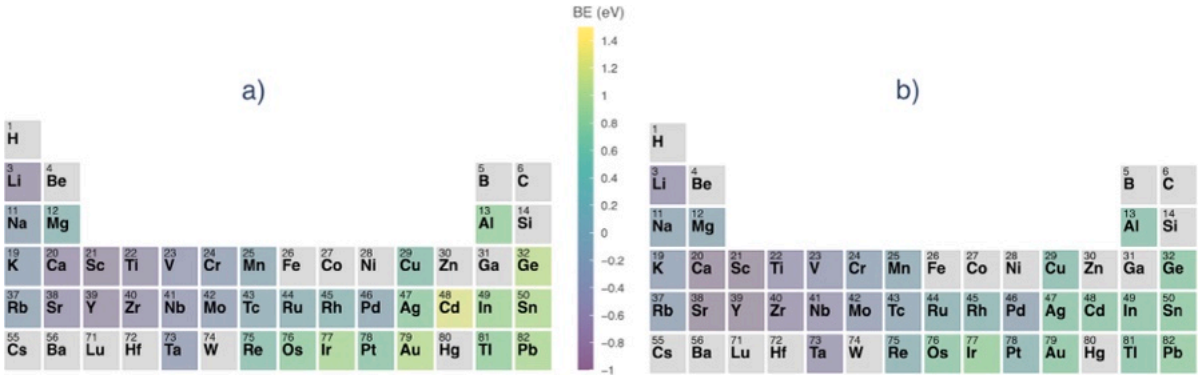


Fig. 8. Periodic table plots of the hydrogen binding energy (BE) in the OI (a) and TI (b) states of the considered FCC crystals. The colorscales are unified to the one of the single colorbar displayed. The bulk crystals that have not been considered in this study have been either removed (lanthanoids, actinoids, groups 15–18, period 7) or left in grey. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

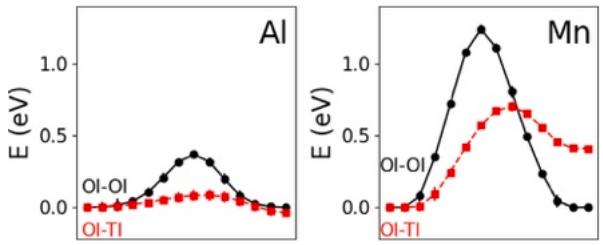


Fig. 9. Hydrogen migration energy barriers in Al (more extended Bader charge volumes) and in Mn (less extended). The barriers are the result of climbing-image NEB calculations, as explained in Section 2.3.1.

In order to try to understand whether an overall classification could emerge in the form of behavioural classes among the investigated samples, clustering methods can be applied. Since we know that the samples are characterized by two different hydrogen interstitial positions in the bulk host crystals, namely OI and TI, the most naive way of tentative clustering could be to highlight the two in the reduced PCA space and evaluate their separation, as shown in Fig. 10. As it can be noticed, the two clusters have a similar distribution in the first two principal components space, but slightly separated. The reason for the shifting lies in the electron density information, the only one differentiating the OI and TI samples at this stage. However, resorting

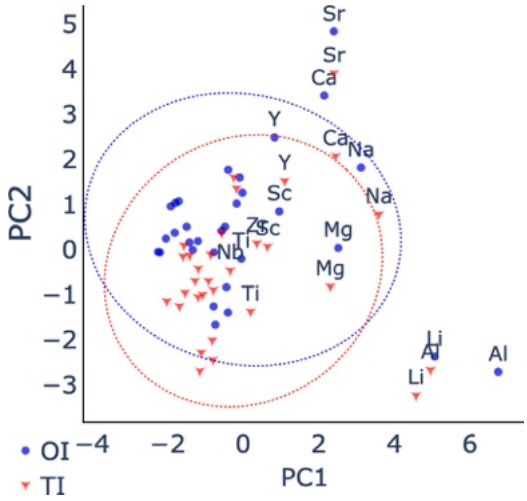


Fig. 10. Visualization of OI and TI samples in the first two-PCA components space, together with the corresponding 95% confidence ellipses. Labelling has been limited to samples out of the crowded region, in order to ease the reading of the plot.

to separation in intuitive classes does not necessarily help with the emergence of meaningful local clusters in our dataset, or with the

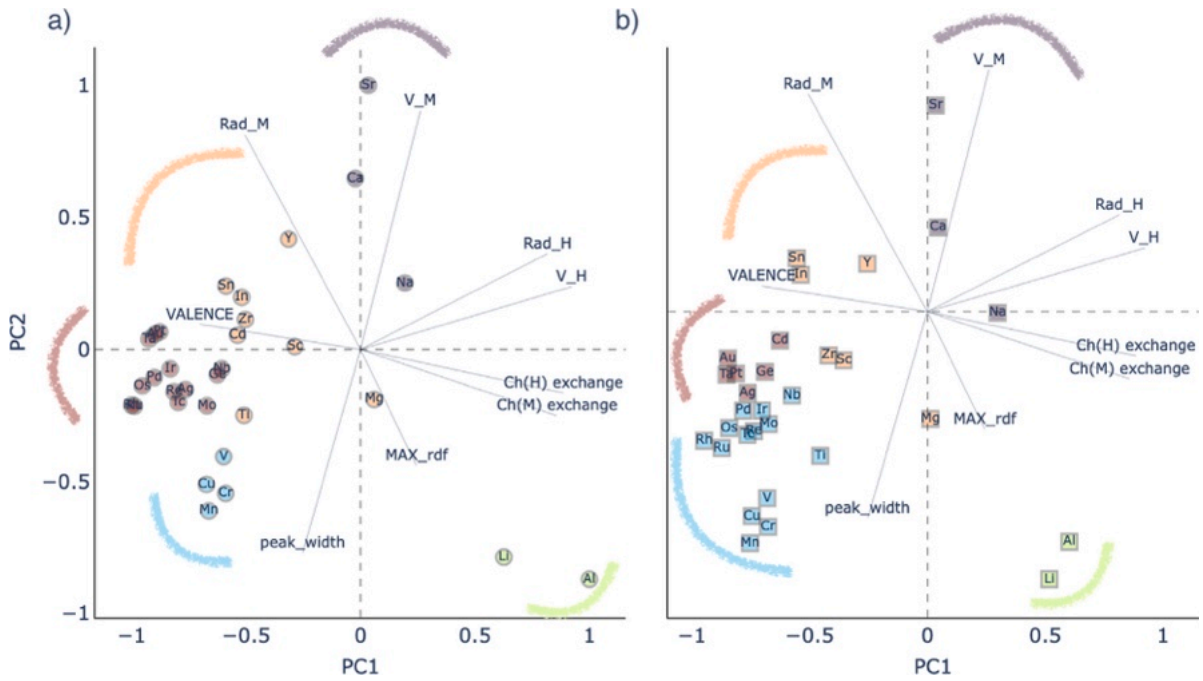


Fig. 11. The biplot for the first and second principal components, both in OI (a) and TI (b) configurations. The plot is a representation of the loadings on the two-components subspace, allowing to recognize the influence of each feature on the components, and thus their driving on the clusters distribution. In order to allow for a comprehensive visualization of the biplot, the PCs are rescaled to the $[-1, 1]$ interval. The colours for the scattered dots come from the K-Means clustering labels for the optimal number of clusters, and coloured arcs highlight the plane sectors where the corresponding clusters are found. The loadings values are presented in the Supplementary Material in Tab.S.(VI). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

understanding of the role of each specific electron density feature in driving the samples distributions.

In absence of prior knowledge on the number of classes that our dataset could distribute among, the K-Means method, introduced in Section 2.4.2, is among the ones that can be adopted to perform cluster analysis on the reduced space of the principal components. A detailed evaluation of the optimal number of clusters should be performed in the absence of ground truth, and, as shown in Fig.S.(3) of SM, the Elbow method indicates $k = 5$ as the number of clusters to optimally group our data.

A quantification of the influence of each feature is necessary and can be accessed through the so called *biplots*, in which the loadings are displayed in the reduced spaces of the principal components. As discussed in Section 2.4.1, the loadings represent the covariances between the PCs and the original variables and therefore are informative on the variables variation explained by the components. In the biplots of Fig. 11, the clusters have been highlighted following the optimal K-Means clustering previously analysed, in order to be able to associate the influence of the original variables also on the clusters distribution, besides the overall one. As it can be noticed, out of the optimal clusters, some are noticeably smaller: the one of Al and Li, i.e. maximizing the charge exchange from M (in terms of a loss) to H (in terms of a gain), or the one of Sr, Ca and Na, maximizing the measures of the metal charge extensions, minimizing at the same time the RDF peak properties. Noticeably, transition metals form three close clusters, all in favour of minimizing the amount of charge exchanged to hydrogen, and so its charge extensions. As seen in the previous subsection, charge extensions are related to migration energy barriers between high-symmetry interstitial states, and this analysis represents an example of direct access to configurational and compositional subspaces exploration for tuning the defect influence on the electron density local landscape, which, as shown, grasps a large variety of physical properties of the host matrix, and of the defect-host interaction, as we believe a general effective descriptor should do.

4. Final remarks

Through a detailed correlation analysis involving the properties of the ab-initio defect-induced charge density perturbations, we prove the ability of the latter to effectively grasp atomic and structural properties of their host metal systems. The expected strong correlations between the induced charge perturbation profiles and the underlying captured atomic charge properties, suggests the role of the latter in the long-bridging abilities of the former with a wide spectrum of crystal properties. The power of the proposed approach lies in its simplicity and generality: any defecting or functionalization of a system introduces a perturbation in the charge density profile, which, embodying the guest-host interaction, naturally inherits features of both on a pure density basis. Further reasoning comes from the brief analysis of the correlation between the hydrogen charge extensions and its transition energy barriers, for candidate systems in which the former quantity presented large differences: where hydrogen displays larger charge extensions, there are found smaller energy barriers for transitioning along different paths. This interesting result shall be further investigated, but clearly defines a strategy of predicting defects mobility just on a density-based argument. The power of the density related features in driving the distribution of the samples in a classification task has been addressed, in turn highlighting the importance of the density-aided classification task itself in studies of defects. The resulting maps, containing behavioural classes spanning different compositions and configurations, can effectively drive material searches basing on maximization/minimization of density properties, knowing their quantified strong influence on the "phenotypical" atomic/structural properties, being the density the well known associated "genotype", in this biological analogue. A detailed analysis of crystal-specific charge density response variations from a compositionally coarse-grained one has also been performed, leading to the spontaneous emergence of behavioural groups showing the same oscillatory peculiarities. Further investigation is needed in order to fully correlate the gradual changes in the charge responses of the different crystals with complex phenomena like catalysis or storage performances. However, the power of the

approach and of the new proposed density-based descriptors gives hope and inspiration for deeper and larger scale investigations along these important directions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data supporting the findings of this study are available upon reasonable request. Please contact the corresponding author for access to the data.

Acknowledgements

We acknowledge support from the European Union Horizon 2020 research and innovation program under grant agreement no. 857470 and from the European Regional Development Fund via the Foundation for Polish Science International Research Agenda PLUS program grant No. MAB PLUS/2018/8. We also acknowledge fruitful conversations with Prof. Ju Li, as well as the MISTI grant, funded by Lockheed-Martin, bridging the contact.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.actamat.2024.119773.

References

[1] J.O. Abe, A.P.I. Popoola, E. Ajenifuja, O.M. Popoola, Hydrogen energy, economy and storage: Review and recommendation, Int. J. Hydrogen Energy 44 (29) (2019) 15072–15086.
[2] R. Moradi, K.M. Groth, Hydrogen storage and delivery: Review of the state of the art technologies and risk and reliability analysis, Int. J. Hydrogen Energy 44 (23) (2019) 12254–12269.
[3] X. Li, X. Ma, J. Zhang, E. Akiyama, Y. Wang, X. Song, Review of hydrogen embrittlement in metals: hydrogen diffusion, hydrogen characterization, hydrogen embrittlement mechanism and prevention, Acta Metall. Sin. (English Letters) 33 (2020) 759–773.
[4] P. Gong, A. Turk, J. Nutter, F. Yu, B. Wynne, P. Rivera-Diaz-del Castillo, W.M. Rainforth, Hydrogen embrittlement mechanisms in advanced high strength steel, Acta Mater. 223 (2022) 117488.
[5] G. Lu, E. Kaxiras, Hydrogen embrittlement of aluminum: the crucial role of vacancies, Phys. Rev. Lett. 94 (15) (2005) 155501.
[6] M.B. Djukic, G.M. Bakic, V. Sijacki Zeravcic, A. Sedmak, B. Rajicic, The synergistic action and interplay of hydrogen embrittlement mechanisms in steels and iron: Localized plasticity and decohesion, Eng. Fract. Mech. 216 (2019) 106528.
[7] V.A. Polyanskiy, A.K. Belyaev, Advances in Hydrogen Embrittlement Study, Springer, 2021.
[8] US DOE, Target explanation document: onboard hydrogen storage for light-duty fuel cell vehicles, https://www.energy.gov/sites/prod/files/2017/05/f34/fcto_targets_onboard_hydro_storage_explanation.pdf.
[9] Jose Bellosta von Colbe, Jose-Ramón Ares, Jussara Barale, Marcello Baricco, Craig Buckley, Giovanni Capurso, Noris Gallandat, David M. Grant, Matylda N. Guzik, Isaac Jacob, Emil H. Jensen, Torben Jensen, Julian Jepsen, Thomas Klassen, Mykhaylol V. Lototskyy, Kandavel Manickam, Amelia Montone, Julian Puszkiel, Sabrina Sartori, Drew A. Sheppard, Alastair Stuart, Gavin Walker, Colin J. Webb, Heena Yang, Volodymyr Yartys, Andreas Züttel, Martin Dornheim, Application of hydrides in hydrogen storage and compression: Achievements, outlook and perspectives, Int. J. Hydrogen Energy 44 (15) (2019) 7780–7808.
[10] Shin-ichi Orimo, Yuko Nakamori, Jennifer R. Eliseo, Andreas Züttel, Craig M. Jensen, Complex hydrides for hydrogen storage, Chem. Rev. 107 (10) (2007) 4111–4132.
[11] Morten B. Ley, Lars H. Jepsen, Young-Su Lee, Young Whan Cho, José M. Bellosta Von Colbe, Martin Dornheim, Masoud Rokni, Jens Oluf Jensen, Mikael Sloth, Yaroslav Filinchuk, Complex hydrides for hydrogen storage–new perspectives, Mater. Today 17 (3) (2014) 122–128.
[12] R.S. Michalski, J.G. Carbonell, T.M. Mitchell, Machine Learning: An Artificial Intelligence Approach, Springer Science & Business Media, 2013.
[13] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[14] S. Papanikolaou, Learning local, quenched disorder in plasticity and other crackling noise phenomena, npj Comput. Mater. 4 (1) (2018) 27.
[15] S. Ramakrishna, T.-Y. Zhang, W.-C. Lu, Q. Qian, J.S.C. Low, J.H.R. Yune, D.Z.L. Tan, S. Bressan, S. Sanvito, S.R. Kalidindi, Materials informatics, J. Intell. Manuf. 30 (2019) 2307–2326.
[16] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: Recent applications and prospects, npj Comput. Mater. 3 (1) (2017) 54.
[17] K. Takahashi, Y. Tanaka, Materials informatics: a journey towards material design and synthesis, Dalton Trans. 45 (26) (2016) 10497–10499.
[18] L. Ward, C. Wolverton, Atomistic calculations and materials informatics: A review, Curr. Opin. Solid State Mater. Sci. 21 (3) (2017) 167–176.
[19] K. Rajan, Materials informatics, Mater. Today 8 (10) (2005) 38–45.
[20] J. Greeley, T.F. Jaramillo, J. Bonde, I.B. Chorkendorff, J.K. Nørskov, Computational high-throughput screening of electrocatalytic materials for hydrogen evolution, Nature Mater. 5 (11) (2006) 909–913.
[21] W.T. Hong, R.E. Welsch, Y. Shao-Horn, Descriptors of oxygen-evolution activity for oxides: a statistical evaluation, J. Phys. Chem. C 120 (1) (2016) 78–86.
[22] C.E. Wilmer, M. Leaf, C.Y. Lee, O.K. Farha, B.G. Hauser, J.T. Hupp, R.Q. Snurr, Large-scale screening of hypothetical metal–organic frameworks, Nature Chem. 4 (2) (2012) 83–89.
[23] L.-C. Lin, A.H. Berger, R.L. Martin, J. Kim, J.A. Swisher, K. Jariwala, C.H. Rycroft, A.S. Bhown, M.W. Deem, M. Haranczyk, et al., In silico screening of carbon-capture materials, Nature Mater. 11 (7) (2012) 633–641.
[24] M. Fernandez, N.R. Trefiak, T.K. Woo, Atomic property weighted radial distribution functions descriptors of metal–organic frameworks for the prediction of gas uptake capacity, J. Phys. Chem. C 117 (27) (2013) 14095–14105.
[25] M. Rupp, A. Tkatchenko, K.-R. Müller, O.A. Von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, Phys. Rev. Lett. 108 (5) (2012) 058301.
[26] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O.A. Von Lilienfeld, Machine learning of molecular electronic properties in chemical compound space, New J. Phys. 15 (9) (2013) 095003.
[27] W. Kohn, A.D. Becke, R.G. Parr, Density functional theory of electronic structure, J. Phys. Chem. 100 (31) (1996) 12974–12980.
[28] A.J. Cohen, P. Mori-Sánchez, W. Yang, Challenges for density functional theory, Chem. Rev. 112 (1) (2012) 289–320.
[29] J.K. Nørskov, F. Abild-Pedersen, F. Studt, T. Bligaard, Density functional theory in surface chemistry and catalysis, Proc. Natl. Acad. Sci. 108 (3) (2011) 937–943.
[30] R.G. Parr, W. Yang, Density-functional theory of the electronic structure of molecules, Annu. Rev. Phys. Chem. 46 (1) (1995) 701–728.
[31] J.W.D. Connolly, A.R. Williams, Density-functional theory applied to phase transformations in transition-metal alloys, Phys. Rev. B 27 (8) (1983) 5169.
[32] D. Kwak, M. Wang, K.J. Koski, L. Zhang, H. Sokol, R. Maric, Y. Lei, Molybdenum trioxide (α-MoO3) nanoribbons for ultrasensitive ammonia (NH3) gas detection: integrated experimental and density functional theory simulation studies, ACS Appl. Mater. Interfaces 11 (11) (2019) 10697–10706.
[33] Q. He, B. Yu, Z. Li, Y. Zhao, Density functional theory for battery materials, Energy Environ. Mater. 2 (4) (2019) 264–279.
[34] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, I. Tanaka, Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization, Phys. Rev. Lett. 115 (20) (2015) 205901.
[35] D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nature Commun. 7 (2016) 11241.
[36] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, Universal fragment descriptors for predicting properties of inorganic crystals, Nature Commun. 8 (1) (2017) 15679.
[37] M. Rupp, A. Tkatchenko, K.-R. Müller, O.A. Von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, Phys. Rev. Lett. 108 (5) (2012) 058301.
[38] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. 134 (7) (2011).
[39] F. Pietrucci, W. Andreoni, Graph theory meets ab initio molecular dynamics: atomic structures and transformations at the nanoscale, Phys. Rev. Lett. 107 (8) (2011) 085504.
[40] S. De, A.P. Bartók, G. Csányi, M. Ceriotti, Comparing molecules and solids across structural and alchemical space, Phys. Chem. Chem. Phys. 18 (20) (2016) 13754–13769.
[41] A. Sadeghi, S.A. Ghasemi, B. Schaefer, S. Mohr, M.A. Lill, S. Goedecker, Metrics for measuring distances in configuration spaces, J. Chem. Phys. 139 (18) (2013).
[42] Stefano Curtarolo, Gus L.W. Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, Ohad Levy, The high-throughput highway to computational materials design, Nature Mater. 12 (3) (2013) 191–201.

[43] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, Christopher Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD), Jom 65 (2013) 1501–1509.
[44] *differential*, is here addressing charge density minus superposition of atomic densities.
[45] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L. Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al., QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials, J. Phys. Condens. Matter 21 (39) (2009) 395502.
[46] Paolo Giannozzi, Oliviero Andreussi, Thomas Brumme, Oana Bunau, M. Buongiorno Nardelli, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Matteo Cococcioni, et al., Advanced capabilities for materials modelling with quantum ESPRESSO, J. Phys. Condens. Matter 29 (46) (2017) 465901.
[47] Paolo Giannozzi, Oscar Baseggio, Pietro Bonfà, Davide Brunato, Roberto Car, Ivan Carnimeo, Carlo Cavazzoni, Stefano De Gironcoli, Pietro Delugas, Fabrizio Ferrari Ruffino, et al., Quantum ESPRESSO toward the exascale, J. Chem. Phys. 152 (15) (2020).
[48] John P. Perdew, Kieron Burke, Matthias Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (18) (1996) 3865.
[49] Riccardo Sabatini, Tommaso Gorni, Stefano De Gironcoli, Nonlocal van der Waals density functional made simple and efficient, Phys. Rev. B 87 (4) (2013) 041108.
[50] M.P.A.T. Methfessel, A.T. Paxton, High-precision sampling for Brillouin-zone integration in metals, Phys. Rev. B 40 (6) (1989) 3616.
[51] https://dalcorso.github.io/thermo_pw/.
[52] D. Ikuta, E. Ohtani, A. Sano-Furukwa, Y. Shibazaki, H. Terasaki, L. Yuan, T. Hattori, Interstitial hydrogen atoms in face-centered cubic iron in the Earth's core, Sci. Rep. 9 (2019) 7108.
[53] W. Tang, E. Sanville, G. Henkelman, A grid-based bader analysis algorithm without lattice bias, J. Phys.: Condens. Matter 21 (8) (2009) 084204.
[54] Here we call *Bader radii* the minimum distances between each atom and the boundary of its Bader volume.
[55] M. Łukasz, Mendeleev – A Python resource for properties of chemical elements, ions and isotopes, 2014, https://github.com/lmmentel/mendeleev, 0.11.0.
[56] https://github.com/materialsproject/pyrho.
[57] M.S. Shell, Thermodynamics and Statistical Mechanics: An Integrated Approach, Cambridge University Press, 2015.
[58] Q. Li, X. Lin, Q. Luo, Y. Chen, J. Wang, B. Jiang, F. Pan, Kinetics of the hydrogen absorption and desorption processes of hydrogen storage alloys: A review, Int. J. Miner. Metall. Mater. 29 (2022) 32–48.
[59] K. Koizumi, K. Nobusada, M. Boero, Hydrogen storage mechanism and diffusion in metal–organic frameworks, Phys. Chem. Chem. Phys. 21 (15) (2019) 7756–7764.
[60] Z. Chen, Z. Ma, J. Zheng, X. Li, E. Akiba, H. Li, Perspectives and challenges of hydrogen storage in solid-state hydrides, Chin. J. Chem. Eng. 29 (2021) 1–12.
[61] J. Čermák, L. Král, Hydrogen diffusion in Mg–H and Mg–Ni–H alloys, Acta Mater. 56 (12) (2008) 2677–2686.
[62] K. Pearson, LIII, On lines and planes of closest fit to systems of points in space, Lond. Edinb. Dublin Philos. Mag. J. Sci. 2 (11) (1901) 559–572.
[63] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417.
[64] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. a Math. Phys. Eng. Sci. 374 (2065) (2016) 20150202.
[65] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory 28 (2) (1982) 129–137.
[66] P. Berkhin, Survey of Clustering Data Mining Techniques, Technical Report, Accrue Software San Jose CA, 2002.
[67] D. Arthur, S. Vassilvitskii, K-means++ the advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007, pp. 1027–1035.
[68] R.L. Thorndike, Who belongs in the family? Psychometrika 18 (4) (1953) 267–276.
[69] it is to notice how this quantity is related to compactness of the clusters.
[70] C. Li, J.-B. Baek, Recent advances in noble metal (Pt, Ru, and Ir)-based electrocatalysts for efficient hydrogen evolution reaction, ACS Omega 5 (1) (2020) 31–40.
[71] F. Marques, M. Balcerzak, F. Winkelmann, G. Zepon, M. Felderhoff, Review and outlook on high-entropy alloys for hydrogen storage, Energy Environ. Sci. 14 (10) (2021) 5191–5227.
[72] it will be shown explicitly soon in the same section.
[73] The radius of the M charge here behaves differently from the volume, and it may be due to its definition in the bader extraction program, as the minimum distance from the Bader surface. In this view, volume can be seen as a more reliable estimate of charge extensions due to anisotropies and bondings. We believe this to be reason also for the very high correlation between V(M) and V(H) and less pronounced one between R(M) and R(H).
[74] These are properties of the metal bulk crystals before being defected with interstitial hydrogen.
[75] S. Satyapal, J. Petrovic, C. Read, G. Thomas, G. Ordaz, The U.S. Department of Energy's National Hydrogen Storage Project: Progress towards meeting hydrogen-powered vehicle requirements, Catalysis Today 120 (3) (2007) 246–256.
[76] G. Lu, D. Orlikowski, I. Park, O. Politano, E. Kaxiras, Energetics of hydrogen impurities in aluminum and their effect on mechanical properties, Phys. Rev. B 65 (6) (2002) 064102.
[77] L. Messina, M. Nastar, N. Sandberg, Pär Olsson, Systematic electronic-structure investigation of substitutional impurity diffusion and flux coupling in BCC iron, Phys. Rev. B 93 (18) (2016) 184302.

Supplemental materials to the manuscript

SIZE EFFECTS

In Fig.S.(1) we propose a comparison between the hydrogen-induced electron charge density perturbation isolines in $2 \times 2 \times$ and $3 \times \times 3$ supercells, considering Mo as an example. As shown in the main results of the paper, the Radial Distribution Functions in the density grids are dominated by positive (accumulation) peaks, coming from a dominance of charge accumulation, rather than deplition, in the different matrices. Fig.S.(1) proves high compatibility among the two supercell sizes results, therefore we believe that a studied carried on $2 \times 2 \times 2$ supercells would not suffer from relevant size artifacts.

UNSUPERVISED LEARNING

In the paper to which this material is supplementary, it has been shown the use of density information coming from different crystals in dataset to be used for unsupervised learning tasks. Among these tasks, a standard technique is Principal Component Analysis, and the evaluation of the explained variance ratio captured by each resulting component, visible in Fig.S.(2). As visible from the histogram, the first two principal components are able to account for more than 70% of the total variance. In the space of the two main principal components, clustering methods can be applied to highlight the emergence of behavioural classes, and we have decided to proceed with K-Means. In absence of ground-truth on the number of clusters, the Elbow method can be applied to find out the optimal one, as in Fig.S.(3).

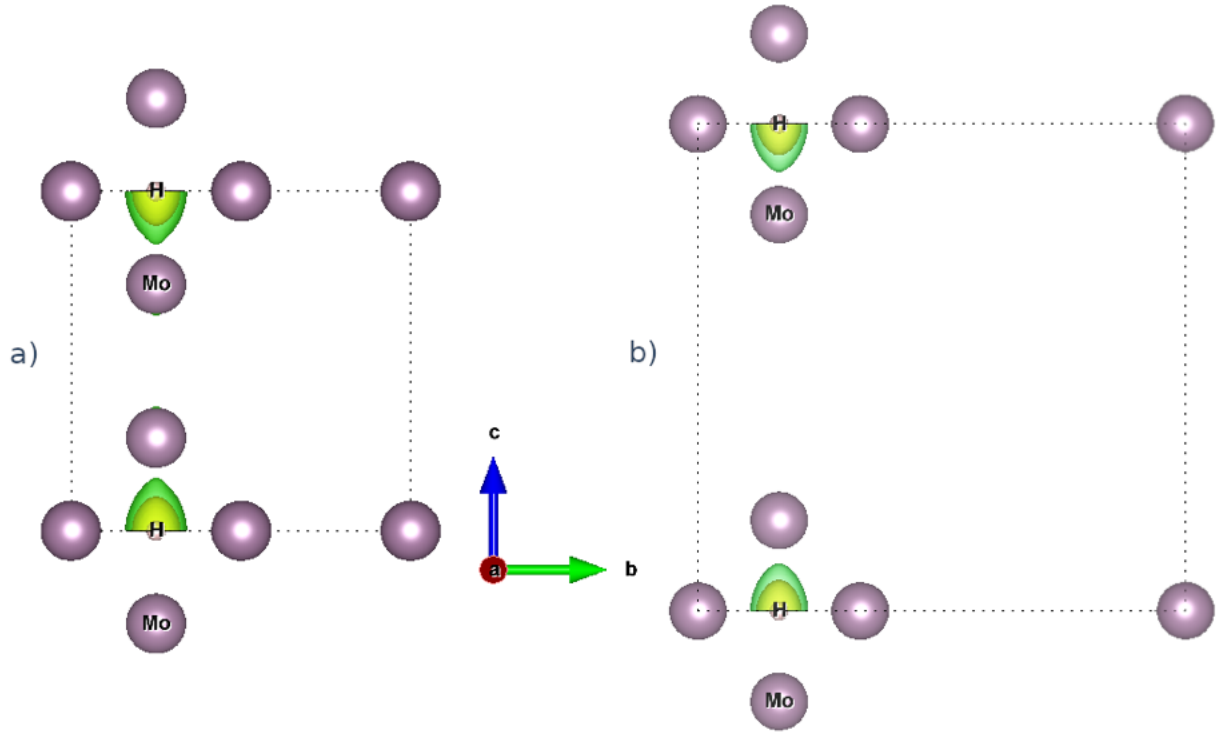


Fig.S. 1: Visualization of the three dimensional differential electron charge perturbation induced by an hydrogen interstitial defect in a $2 \times 2 \times 2$ (a) and $3 \times 3 \times 3$ (b) supercell of Mo. Two same isoline levels, in green and yellow, are visible in the plots, and their values are reported in Tab.S.(V), together with the minimum and maximum isolines values of the two grids. For simplicity of visualization, just the H-neighbouring Mo atoms and their periodic replicas are reported in the FCC supercells.

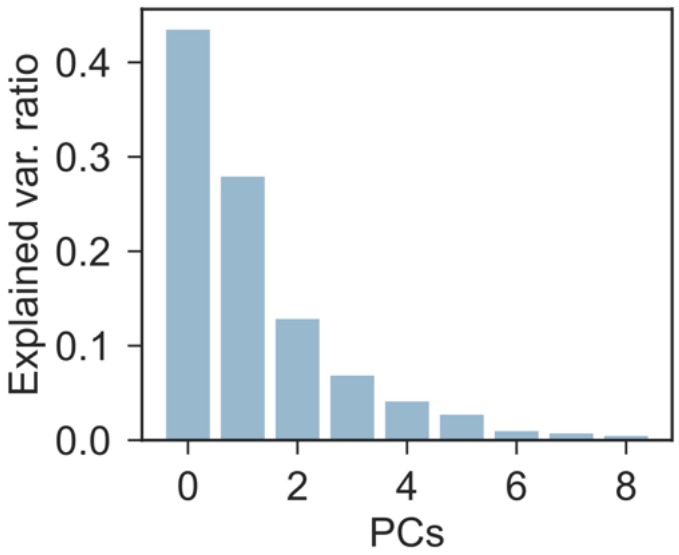


Fig.S. 2: Histogram of explained variance ratios by each of the resulting principal components after applying PCA: the first two, are able to account for about 70% of the total variance.

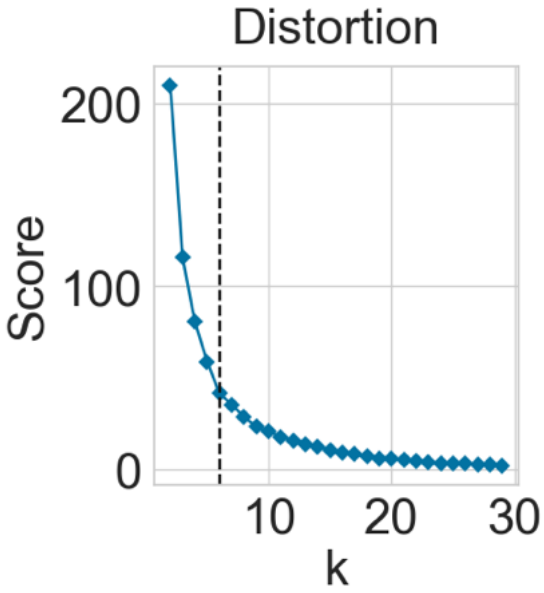


Fig.S. 3: The Elbow method score as a function of the number k of K-Means clusters. The test indicates $k = 5$ as the number of clusters to optimally group our data.

DIFFERENT SYMMETRIES

In the main text of the work, we have presented results regarding hydrogen-induced electron charge density perturbations in a variety of FCC metallic bulk systems. We found interesting the idea of providing few examples of the same densities in non-metallic bulks of different symmetries, like HCP. In Fig.S.(4,5,6) we provide these examples for one interstitial hydrogen atom living in $2 \times 2 \times 2$ supercells of C, Si and Hg, respectively. Intentionally, we selected interstitial sites without paying attention to them being highly symmetric, in order to explore the characterization of the perturbations in various conditions. In HCP C, the hydrogen atom sits in between two layers of the bulk structure. The pertubation symmetry is locally spherical and resembles the ones examined in FCC crystals. In HCP Si, the perturbation is still predominantly characterized by charge accumulation with slight elongated distorsion towards the Si atoms highlighting bonding. In HCP Hg, hydrogen-induced charge depletion gains importance and reaches the

same order of magnitude as accumulation. In this latter case, the hydrogen fingerprint on the density field has multiple peaks and a much less trivial landscape, but would still represent a unique descriptor of the system properties.

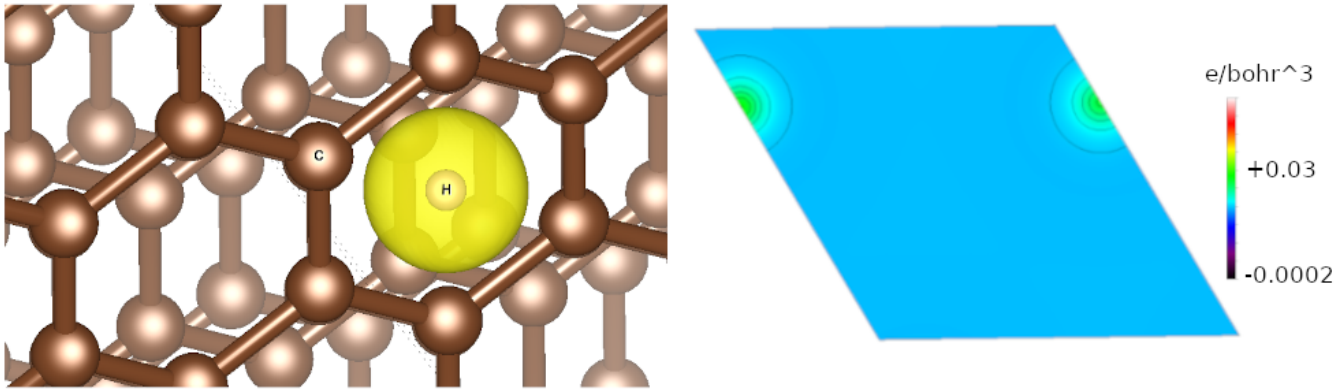


Fig.S. 4: Three dimensional (left) and two dimensional visualization (right) of the differential electron charge density perturbation induced by the hydrogen defect in a HCP C $2 \times 2 \times 2$ supercell. The isoline value on the left is 0.0008 e/bohr^3.

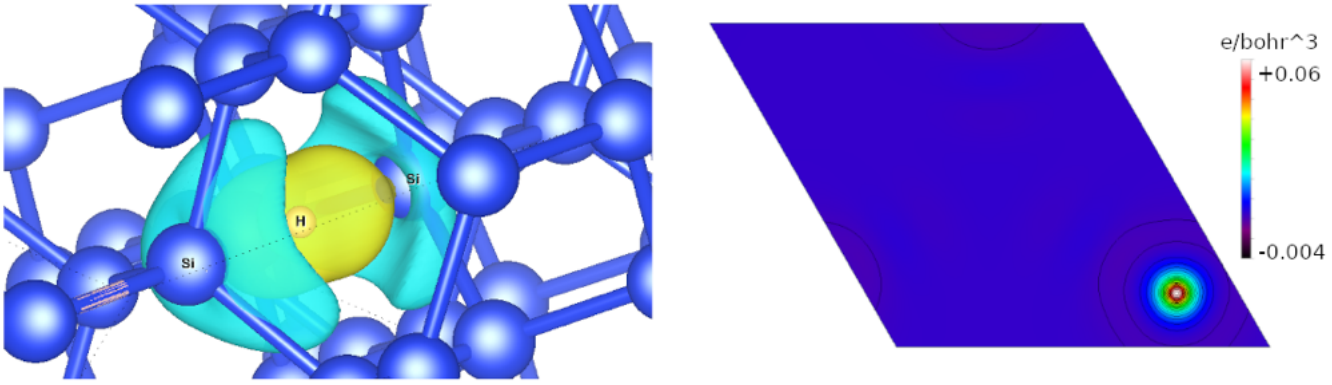


Fig.S. 5: Three dimensional (left) and two dimensional visualization (right) of the differential electron charge density perturbation induced by the hydrogen defect in a HCP Si $2 \times 2 \times 2$ supercell. The isoline value on the left is 0.0020 e/bohr^3.

ADDITIONAL FIGURES

TABLES

[htb]

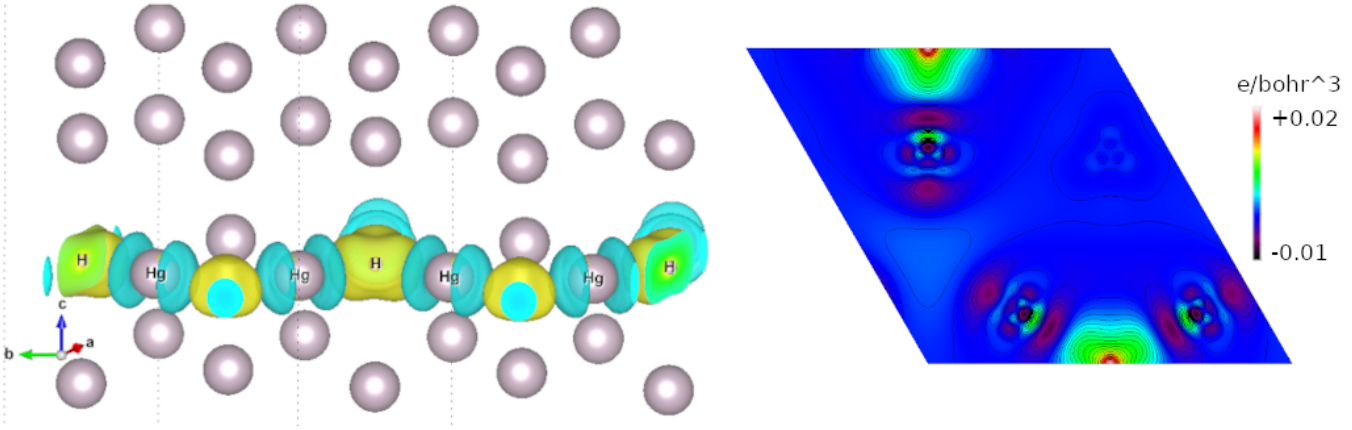


Fig.S. 6: Three dimensional (left) and two dimensional visualization (right) of the differential electron charge density perturbation induced by the hydrogen defect in a HCP Hg $2 \times 2 \times 2$ supercell. The isoline value on the left is 0.0010 e/bohr^3.

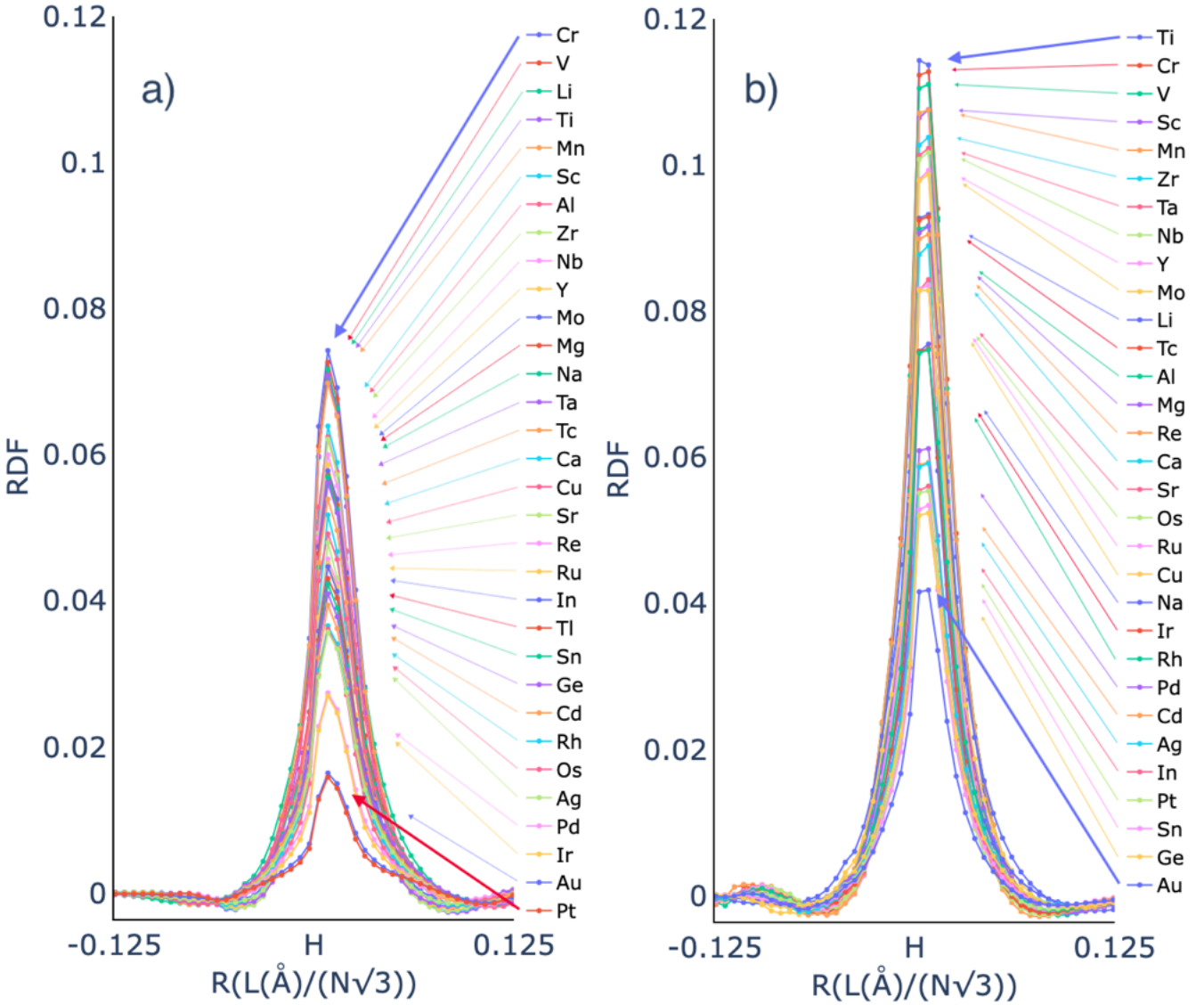


Fig.S. 7: Density Radial Distribution Function profiles for the different metal species forming the host crystals, both for the OI case (a) and TI (b). The profiles are all centered around the relative hydrogen atom position in the grid box. Coloured arrows point towards the maximum of the corresponding RDF curve. The unique radius R for the building of the RDFs comparison is here reported in units normalized by the number of grid points, $N = N_x = N_y = N_z = 100$, and taking into account the three-dimensional nature of the distances. For completeness, some relevant quantities from the plots are reported in Tab.S.(III) and Tab.S.(IV) of SM.

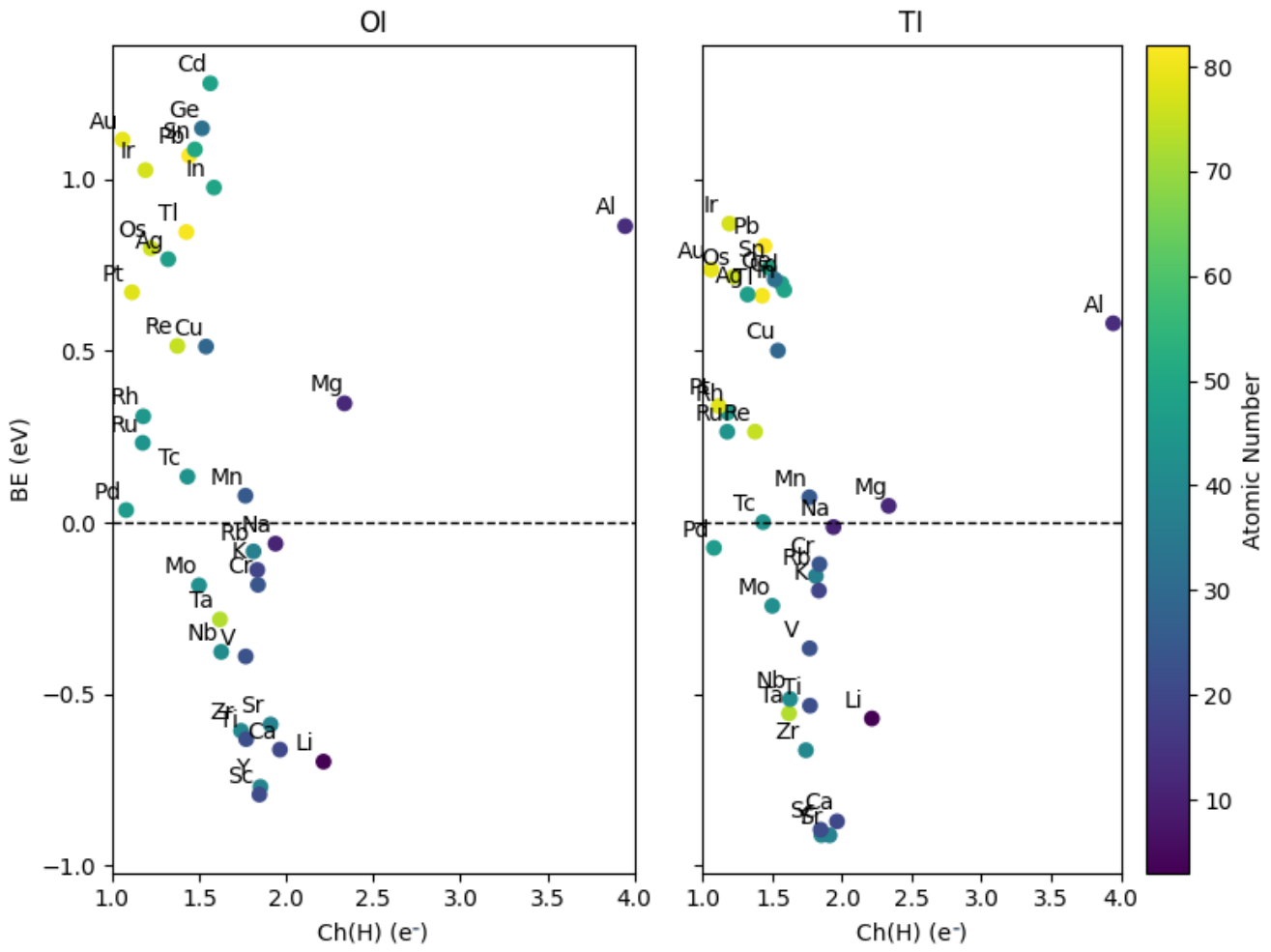


Fig.S. 8: Visualization of the relation between the hydrogen binding energies and its Bader charge extensions in the different considered crystals.

Crystal	E_{cut}(Ry)	$k_{pts.}$	degauss(Ry)
Thallium	120	8	0.05
Lead	120	12	0.02
Tin	100	16	0.025
Rubidium	80	10	0.025
Indium	100	14	0.04
Zirconium	100	10	0.02
Yttrium	80	10	0.02
Tantalum	120	10	0.04
Strontium	80	10	0.015
Rhenium	120	14	0.03
Niobium	120	8	0.05
Sodium	100	10	0.04
Potassium	80	10	0.03
Cadmium	80	16	0.03
Silver	120	12	0.05
Rhodium	120	14	0.03
Osmium	120	16	0.03
Molybdenum	120	12	0.001
Scandium	100	8	0.03
Magnesium	100	10	0.04
Vanadium	100	10	0.05
Titanium	100	8	0.03
Technetium	80	10	0.03
Ruthenium	100	14	0.03
Platinum	80	14	0.03
Palladium	80	12	0.03
Iridium	80	16	0.03
Germanium	80	16	0.03
Chromium	100	8	0.005
Calcium	80	10	0.01
Gold	80	14	0.001
Aluminum	80	10	0.04
Manganese	100	8	0.03
Copper	80	12	0.05
Lithium	80	10	0.03

Tab.S. I: Table with the converged DFT parameters implemented in the simulations of the pure FCC metallic samples. The kinetic energy cutoff for the charge density and the potential is taken as $E_{cut}^{rho} = 6 \cdot E_{cut}$, while the number of k-points in the table is intended to represent the selected 3-dimensional mesh $k_{pts.} \times k_{pts.} \times k_{pts.}$ in k-space. During the convergence tests, the common acceptance threshold in the variation of the total energy upon parameter change has been set to 10^{-5} Ry.

Crystal	Lattice Const. M (Å)	B(kbar)	Poisson R.	Young M.	Shear M.
Thallium	5.002120	296.44047	0.40522	41.28377	14.68947
Lead	5.069333	407.73666	0.78395	-233.36752	-65.40753
Tin	4.872865	439.46883	0.40446	216.13070	76.94449
Rubidium	6.896321	33.59768	0.34958	26.82293	9.93748
Indium	4.824002	343.24574	0.43818	95.77966	33.29902
Zirconium	4.543623	930.20851	0.34802	823.87718	305.58726
Yttrium	5.045803	392.16302	0.25913	562.49297	223.36647
Tantalum	4.240945	1740.99067	0.09708	2258.63998	1029.38703
Strontium	5.934284	130.09873	0.27180	168.15819	66.11013
Rhenium	3.936890	3501.73236	0.28170	4582.32578	1787.59034
Niobium	4.242892	1620.01587	0.66029	-1462.08151	-440.30960
Sodium	5.210005	84.89276	0.32649	77.96226	29.38663
Potassium	6.466959	41.39434	0.33615	35.33335	13.22206
Cadmium	4.525227	436.67101	0.66442	-131.12809	-39.39151
Silver	4.183577	913.21303	0.37763	660.02634	239.55201
Rhodium	3.865649	2436.38085	0.25757	3541.13412	1407.92280
Osmium	3.884664	3698.84839	0.23318	5906.41516	2394.79156
Molybdenum	4.033093	2304.21667	0.79723	-3992.55717	-1110.75225
Scandium	4.596587	506.92217	0.29678	608.82277	234.74411
Magnesium	4.493598	365.42375	0.26644	505.05471	199.40004
Vanadium	3.832294	1757.16501	0.81350	-211.67221	-58.36018
Titanium	4.116707	1097.47681	0.36308	857.83362	314.66815
Technetium	3.902914	2850.00899	0.30875	3263.65295	1246.86241
Ruthenium	3.844112	2884.20345	0.25197	4280.15149	1709.37112
Platinum	4.026684	2280.07686	0.37831	1662.58253	603.12330
Palladium	3.988079	1627.68208	0.36285	1322.69354	485.26651
Iridium	3.917226	3231.09305	0.23926	5053.83387	2039.05015
Germanium	4.352847	526.74907	0.41439	217.70530	76.96105
Chromium	3.640683	2416.34647	0.94407	-6413.44601	-1649.49145
Calcium	5.446610	186.12162	0.29113	229.70885	88.95667
Gold	4.207946	1246.19358	0.35609	181.33488	66.85944
Aluminum	4.051156	753.26295	-2.42950	6628.79311	-2318.56397
Manganese	3.524944	2610.34410	0.24555	3979.29705	1597.40238
Copper	3.644979	1713.15436	0.36665	1364.14460	499.08408
Lithium	4.288954	148.32103	0.34126	128.74396	47.99367

Tab.S. II: The structural information of the pure bulk metals. The Bulk Moduli (B), Poisson Ratios, Young Moduli and Shear Moduli have been computed via Thermo_PW in Quantum Espresso.

Crystal	VALENCE	Ch(M) ex.(e^-)	Ch(H) ex.(e^-)	R(M)(Å)	R(H)(Å)	V(M)(Å^3)	V(H)(Å^3)	RDF peak width	RDF MAX
Thallium	14.0002	1.090700	0.426291	2.887141	1.810707	202.957874	52.687930	5.123690	0.043096
Lead	14.9998	1.092029	0.442550	2.925526	1.809477	212.174845	50.747294	3.311369	0.000320
Tin	14.9998	1.116130	0.475228	2.754140	1.790510	186.371487	47.363754	5.076671	0.042318
Rubidium	9.9999	1.169677	0.811846	3.233507	2.579271	481.020725	172.325983	2.308308	0.000124
Indium	14.0002	1.114555	0.584974	2.745684	1.772557	180.425142	59.032840	5.124412	0.044712
Zirconium	13.0001	0.995388	0.740299	2.443752	1.788788	157.555295	43.461600	4.817026	0.062148
Yttrium	11.9998	1.030338	0.851696	2.619778	1.986491	212.747840	65.069771	4.632992	0.058647
Tantalum	28.0003	1.000756	0.619039	2.409388	1.558314	128.334835	30.065419	4.868236	0.056160
Strontium	11.0000	1.113777	0.910118	2.928166	2.383004	333.896394	117.093181	4.451627	0.047993
Rhenium	15.9999	1.065569	0.374854	2.299586	1.320261	100.843211	19.611020	4.959082	0.045747
Niobium	14.0001	1.083274	0.626436	2.412199	1.559032	125.868122	31.868896	4.942415	0.060012
Sodium	9.9999	1.198209	0.938265	2.251995	2.338292	200.104177	152.026014	4.928493	0.056993
Potassium	10.0001	1.190132	0.833606	2.908772	2.579914	388.794096	168.589937	3.080109	0.000342
Cadmium	13.0001	1.109309	0.563714	2.571604	1.710287	149.213605	54.083418	5.149464	0.039453
Silver	12.0000	1.042793	0.321617	2.490956	1.427438	120.711034	33.179946	5.376127	0.035751
Rhodium	18.0001	0.999350	0.178495	2.349289	1.255547	96.555521	18.211878	5.245036	0.036602
Osmium	17.0001	1.014469	0.221748	2.379844	1.261711	97.883616	16.598398	5.056293	0.036162
Molybdenum	15.0001	1.127255	0.499084	2.293173	1.481930	106.406886	24.481625	5.053025	0.057802
Scandium	11.9998	1.208514	0.845836	2.400477	1.809642	150.205482	57.297644	4.871119	0.063933
Magnesium	11.0002	1.207869	1.334277	1.964853	2.122914	138.399775	101.149970	5.170782	0.057169
Vanadium	14.0002	1.004913	0.767163	2.024360	1.536250	94.678433	32.810298	5.229013	0.072671
Titanium	13.0002	1.031297	0.769750	2.131976	1.649858	115.736830	39.986057	5.090211	0.070894
Technetium	16.0001	1.038577	0.432690	2.285423	1.365815	98.687362	21.985729	5.014717	0.053937
Ruthenium	17.0001	0.999019	0.175542	2.389375	1.210719	95.050740	16.939753	5.184028	0.045108
Platinum	10.9999	0.985363	0.113874	2.467398	1.268213	109.542614	19.348826	4.794035	0.015905
Palladium	10.9999	0.979355	0.080380	2.418034	1.256054	106.356794	19.482662	5.118990	0.027426
Iridium	10.0000	1.014016	0.190940	2.406424	1.233746	100.230725	17.751418	4.917427	0.027035
Germanium	15.0000	1.121465	0.516142	2.466538	1.645134	132.218509	44.834682	5.305638	0.040993
Chromium	14.9997	1.108383	0.837709	1.999145	1.401443	78.341578	29.980409	5.330796	0.074272
Calcium	11.0002	1.070980	0.963679	2.604790	2.287240	262.507311	109.599336	4.575987	0.051780
Gold	12.0002	1.023319	0.059576	2.605179	1.325308	123.650005	22.275065	5.052181	0.016455
Aluminum	3.9999	1.476572	2.946357	0.703848	2.479957	93.710980	119.706033	5.284540	0.062454
Manganese	15.9997	1.099465	0.765646	1.957536	1.356901	71.415206	26.929691	5.461539	0.069813
Copper	11.9998	1.061955	0.539289	2.005434	1.377594	79.417147	30.107248	5.477359	0.049192
Lithium	3.9999	1.778538	1.213930	1.544476	2.161298	36.699317	133.519060	5.296808	0.071685

Tab.S. III: The density information coming from the Bader Charge and RDF Analysis on the OI-defected supercell samples.

Crystal	VALENCE	Ch(M) ex.(e^-)	Ch(H) ex.(e^-)	R(M)(Å)	R(H)(Å)	V(M)(Å^3)	V(H)(Å^3)	RDF peak width	RDF MAX
Thallium	14.0002	1.093567	0.356148	2.545377	1.465705	203.888848	48.841286	1.151142	0.000797
Lead	14.9998	1.140725	0.413141	2.584940	1.508106	210.276691	50.398776	1.110279	0.000782
Tin	14.9998	1.146056	0.447967	2.450224	1.480478	185.886156	49.321816	4.709087	0.053449
Rubidium	9.9999	1.348649	0.778730	2.809091	2.040784	419.691058	165.045977	1.382962	0.000286
Indium	14.0002	1.136333	0.450868	2.389979	1.465630	180.368311	50.313394	4.759745	0.056079
Zirconium	13.0001	1.159861	0.787938	2.190511	1.311776	152.153758	41.626603	4.696433	0.103792
Yttrium	11.9998	1.180116	0.836165	2.271268	1.456762	205.789103	56.499096	4.305434	0.099258
Tantalum	28.0003	1.056152	0.705935	2.123984	1.224397	127.370855	31.141105	4.842529	0.102275
Strontium	11.0000	1.269420	0.869590	2.516584	1.699554	312.108926	95.041288	3.878054	0.084323
Rhenium	15.9999	1.113671	0.495037	2.068914	1.077999	100.422245	21.926330	5.064890	0.090493
Niobium	14.0001	1.088275	0.684128	2.120602	1.224953	126.613851	32.139568	4.896473	0.101722
Sodium	9.9999	1.463217	0.893035	1.943787	1.850117	145.656724	148.928045	4.479028	0.075516
Potassium	10.0001	1.422602	0.806801	2.539632	1.999094	311.017251	166.239056	1.172665	0.000790
Cadmium	13.0001	1.102413	0.391133	2.251143	1.321713	150.053184	41.895034	5.018768	0.059318
Silver	12.0000	1.060015	0.248673	2.149160	1.141106	120.659403	27.813010	5.193334	0.059168
Rhodium	18.0001	1.023506	0.238498	2.019039	0.988487	96.107919	18.714500	5.284418	0.074714
Osmium	17.0001	1.077913	0.389592	2.083358	0.993345	96.824240	19.239292	5.083399	0.083542
Molybdenum	15.0001	1.110887	0.565290	2.097791	1.100053	108.221570	25.653460	5.010381	0.098712
Scandium	11.9998	1.062693	0.879424	2.007437	1.396542	160.374054	52.697116	4.722296	0.107582
Magnesium	11.0002	1.314572	1.097658	1.699838	1.705034	131.015764	82.999180	4.963581	0.091603
Vanadium	14.0002	1.013847	0.650536	1.869766	1.119319	94.868342	29.128173	5.442587	0.111030
Titanium	13.0002	1.148948	0.800109	1.893797	1.212688	112.279150	38.441382	5.067758	0.114307
Technetium	16.0001	1.078196	0.423641	2.005951	1.064550	98.172688	22.278736	5.045718	0.092920
Ruthenium	17.0001	1.061316	0.339811	2.047080	0.983592	93.881196	19.946138	5.250330	0.083500
Platinum	10.9999	1.041885	0.192874	2.131596	0.942638	108.417780	19.844089	4.839737	0.055412
Palladium	10.9999	1.042516	0.213853	1.998600	1.013624	104.660523	21.921053	5.062282	0.061210
Iridium	10.0000	1.054041	0.374941	2.008881	0.995622	99.705215	20.216678	4.879931	0.074878
Germanium	15.0000	1.120860	0.355051	2.196150	1.350224	132.143784	39.080898	5.234845	0.052370
Chromium	14.9997	1.094787	0.603630	1.767076	1.071638	79.731303	24.965407	5.461745	0.112767
Calcium	11.0002	1.342536	0.916390	2.204972	1.654794	230.761276	89.605791	4.166815	0.088972
Gold	12.0002	1.025995	0.133964	2.197147	1.076686	124.290160	23.280690	5.042500	0.041849
Aluminum	3.9999	1.392441	1.738636	0.655547	2.392302	96.574625	80.546765	5.118417	0.091678
Manganese	15.9997	1.098617	0.496781	1.751445	1.037575	71.685497	21.893630	5.599272	0.107534
Copper	11.9998	1.092595	0.325503	1.769494	1.000538	78.698192	23.330764	5.554652	0.082863
Lithium	3.9999	1.738300	1.168284	1.288919	1.637778	41.523050	129.875995	5.351738	0.093254

Tab.S. IV: The density information coming from the Bader Charge and RDF Analysis on the TI-defected supercell samples.

Mo	$2 \times 2 \times 2$ (e/bohr^3)	$3 \times 3 \times 3$ (e/bohr^3)
Green isol.	0.005	0.005
Yellow isol.	0.01168	0.01168
Max isol.	0.0804678	0.07499216
Min isol.	-0.00838115	-0.00889763

Tab.S. V: Isoline values for the size effects test reported in Fig.S.(1).

Features	PC1	PC2
VALENCE	-0.703	0.096
Ch(M) exchange	0.857	-0.251
Ch(H) exchange	0.885	-0.165
Radius of M charge	-0.506	0.813
Radius of H charge	0.816	0.362
Volume of M charge	0.261	0.904
Volume of H charge	0.922	0.238
RDF peak width	-0.253	-0.770
Max of RDF	0.243	-0.440

Tab.S. VI: Table of the computed loadings on PC1 and PC2 on the pure density dataset.

2.2 Manuscript II

"Transfer Learning in Materials Informatics: structure-property relationships through minimal but highly informative multimodal input"

Dario Massa, Grzegorz Kaszuba, Stefanos Papanikolaou and Piotr Sankowski

Commentary

Knowing that material properties are closely linked to their electronic structure, we propose supervised DL solutions that directly extract descriptors from ab-initio ECDs profile images through CNNs, and from textual crystallographic information through LLMs. With such aim, particular focus is devoted to minimizing the complexity of architecture modeling stage, of data dimensionality and model training, for a cost-effective but high-impact solution in the field of materials design and cheminformatics. Transfer learning (TL) approaches are exploited, so to adapt the knowledge of existing pre-trained and open-source models to new data from the physics domain. In particular, the regression of energy and structural properties in the example of a very diverse but limited set of FCC stable crystals' compositions is presented.

We first demonstrate the application of OpenAI's Contrastive Language-Image Pre-training (CLIP) model. While allowing to skip the architecture modeling step, such a ready-to-use pre-trained multimodal solution does not allow to directly inspect the contributions of each modality to the training process, or freedom in the choice of the single modality models to be used. Therefore, the study presents also a protocol for training a custom stacked model, combining pre-trained single-modalities at the level of their high-dimensional representations, yielding an interpretable, flexible solution for various applications. Key findings include the overall ability of the presented approaches to predict materials properties from limited data inputs. The bulk modulus prediction represents a nice example in which a predictive tool is highly desirable, due to the need of multiple DFT calculations for its estimate. The inclusion of semantic crystallographic data can enhance the predictive accuracy in the regression of energetics, while does not really affect the prediction of the bulk modulus. However, future work should include multiple experimental runs to quantify variability in the predictions, especially for the formation energy property for which contrastive trends of improvement are found when different amount of text is involved.

The promising results lay a foundation for further development of multimodal AI methods in MI, and hint at future possibilities for inverse design in Materials Science, integrating low-cost AI and generative models to facilitate data-driven materials discovery.

In this study, the PhD candidate: conceptualized the work; created a database by retrieving the properties of a subset of materials from the Materials Project Database with the dedicated API in Python (electron densities, the associated materials properties and textual crystallographic information), and curating all its aspects (best slicing of the ECDs, outliers identification, different text amount); with the support of his collaborators, the candidate also identified valuable ways of including the textual information in the supervised DL schemes, applying transfer learning techniques on the CLIP pre-trained multi-modal model, as well as on models for separated modalities (visual and semantic); the candidate trained the models on HPC, and also took care of preparing part of the figures, the manuscript draft and all its subsequent versions according to the suggestions of the co-authors.

Transfer Learning in Materials Informatics: structure-property relationships through minimal but highly informative multimodal input

Dario Massa
IDEAS-NCBR Chmielna 69 00-801 Warsaw, Poland
NOMATEN Center of Excellence in Multifunctional Materials,
ul. Andrzeja Sołtana, 05-400 Otwock, Świerk, Poland
dario.massa@ncbj.gov.pl

Grzegorz Kaszuba
IDEAS-NCBR Chmielna 69 00-801 Warsaw, Poland
Poznań University of Technology, Poland

Stefanos Papanikolaou
NOMATEN Center of Excellence in Multifunctional Materials,
ul. Andrzeja Sołtana, 05-400 Otwock, Świerk, Poland

Piotr Sankowski
IDEAS-NCBR Chmielna 69 00-801 Warsaw, Poland
MIM Solutions, Poland
piotr.sankowski@mim.ai

Abstract

In this work we propose simple, effective and computationally efficient transfer learning approaches for structure-property relation predictions in the context of materials, with highly informative input from different modalities. As materials properties stand from their electronic structure, representations are extracted directly from datasets of electronic charge density profile images using Neural Networks. We demonstrate transferability of the existing pre-trained Convolutional Neural Networks and Large Language Models knowledge to physics domain data, exploring a wide set of compositions for the regression of energetics- or structure-related properties, and the role of semantic crystallographic information in the context of multimodal approaches. We test the applicability of the CLIP multimodal model, and employ as well a training protocol for building a more interpretable and versatile stacked custom solution from different pre-trained modalities. The study offers a promising avenue for enhancing the effectiveness of descriptor identification in physical systems, shedding light on the power of multimodal transfer learning for materials property prediction. Instead of using the well-established GNN-based approaches, we explore the transfer learning of image- and text-based architectures, which can impact decision making for new low-cost AI methods in the field of Materials and Chemoinformatics.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

Machine Learning (ML) [1, 2] has initiated a remarkable transformation in Materials Science, with the emergence of Materials Informatics (MI) playing a pivotal role in pushing the boundaries of innovation in the context of multiple technological research challenges [3, 4, 5, 6, 7, 8, 9, 10, 11]. A key aspect is the identification of efficient descriptors, capable of capturing the essential features of complex physical systems. In the realm of materials simulations, a wide variety of descriptors have been proposed, both based on translationally and rotationally invariant functions of atomic coordinates [12, 13, 14, 15, 16], or involving fixed-length feature vectors of atomic or electronic properties [17, 18, 19], as well as more advanced feature extraction approaches through Graph Convolutional Neural Networks (GCNNs) [20, 21, 22, 23, 24, 25, 26]. Wide open-source databases [27, 28, 29, 30], originating from high-throughput [31, 32] ab-initio Density-Functional-Theory (DFT) [33, 34] calculations, represent the data source for the development of such models and MI in general.

In real-life research settings, it is very common to come across low-quantity but highly informative data about the systems of interest: in the field of Molecular Science, an example can be found in molecular orbitals, encoding reactivity and optical properties; in the field of Materials Science, for example, phonon dispersion relations encode information on thermal conductivity. Recent works have demonstrated the power of deploying Electron Charge Density (ECD) profiles in ML for unsupervised studies of defects-related effects in metals [35], or prediction of density fields in unseen samples via graph-based models [36]. The interest in such quantity lies in its encoding of the necessary information regarding the properties of the system – in terms of constitutive atoms and their interactions - regardless of dimensionality, stability or geometry. In this view, the ECD profiles represent an example of highly informative data, which could be efficiently exploited for structure-property relations also with the simplest available solutions.

In this work, we aim to propose to utilize the computational efficiency of multimodal transfer learning solutions to predict materials properties from ECD profiles [37]. We build an image and text dataset based on the Materials Project (MP) [28], which could stand as a representative example of a relatively small-sized but very diverse problem to be learned. We successfully test the transferability of the known multimodal Contrastive Language-Image Pre-training (CLIP) model from OpenAI [38]. Furthemore, we employ a protocol for building a custom, more interpretable and versatile stacked multimodal solution starting from single pre-trained modalities. The proposed setting stands as an impactful solution involving minimal information and computational effort which could lay the basis for novel low-cost AI models for inverse materials design and databases in the field of Materials and Cheminformatics.

2 Methods

2.1 Multimodal dataset

We build our multimodal dataset as a subset of the Materials Project [28], restricting it to stable face-centered cubic (FCC) crystals. Their MP-IDs, the associated target properties and ECD fields are collected, resulting in 781 samples. Data-cleaning is performed to filter out outliers in the distribution of maxima and minima of the collected ECD fields, through definition of acceptance boundaries with the interquartile range (IQR), resulting in 592 samples. As shown in Fig.(3) of the Appendix, the majority of the collected samples is represented by ternary crystals, and their quantum-level characterization is very demanding from a computational point of view. We use the dedicated Py-Rho [39] to unify the grid dimensions to $60 \times 60 \times 60$. To further simplify the problem without losing the descriptive power of the data, we slice the three-dimensional data along specific bi-dimensional plane, characterized by the highest atomic density. Having restricted ourselves to FCC crystals, such plane corresponds to the $\langle 111 \rangle$ Miller indices. As shown in Fig.(1), we perform the slicing in a supercell made by an $8 \times 8 \times 8$ repetition of the unit-cell field data, and focus the extraction of a squared image from such slice. Even though both total and differential ECDs are available, we restrict the dataset on the latter type for the enhanced variability within and among samples originating from the highlighting of interatomic interactions, rather than local atomic electron densities. A comparison between the slices along the z-axis of the total and differential ECD fields is presented in Appendix in Fig.(4). For regression experiments, we target energetics, with the formation (E_F) and Fermi energies (E_{fermi}), as well as a structural property, the bulk modulus (K_{vrh}). Before application of the models,

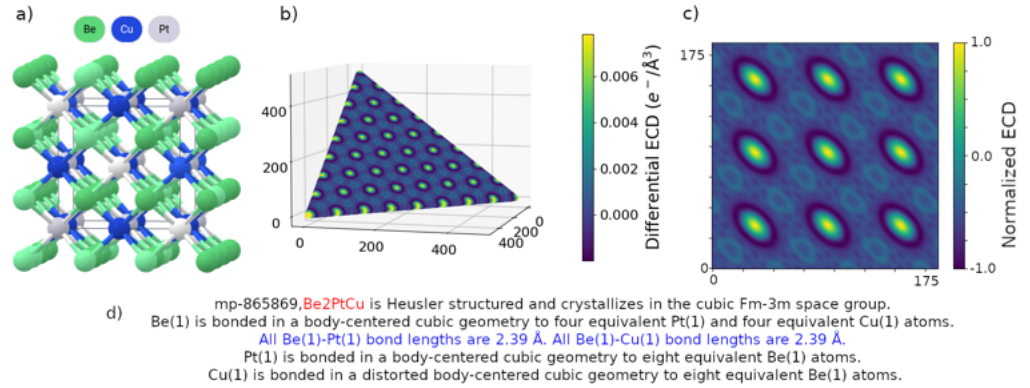


Figure 1: Visualization of the creation steps of the multimodal dataset. a) A bulk crystal is considered together with its electron charge density field, its target properties and its structure file (.cif). b) The differential electron charge density field is sliced along the maximum atomic density plane from an 8x8x8 repetition of the sample. c) A squared focus from the slice is extracted, and properly normalized. d) The crystallographic textual information about the sample is produced. During experiments, we also consider fractions of text as: 'Formula' (only the composition name), 'Keywords' (the composition name and the bond lengths information), and 'All' (the entire text).

the target properties undergo standardization, and the ECD isoline values a z-score normalization and rescaling in the $[-1; 1]$ range: this allows to preserve peculiar patterns of zero-level oscillations across electron accumulation and depletion isolines. Finally, with the help of the Robocrystallographer [40], a tool capable of generating concise text descriptions of crystalline structures by analyzing their symmetry, local environment, and extended connectivity from their Pymatgen structures, we build a set of text attributes associated to each MP-ID, an example of which is reported in Fig. (1).

2.2 Multimodal models

We use two foundational models to incorporate the textual and image information about the crystal structures, resulting in a stacked custom multimodal model. The schematic representation of the architecture is reported in Appendix in Fig.(5) To process the images, we use InceptionV3, a model pre-trained on ImageNet. For text, we use Roberta, trained on language modeling task. We use the same procedure to adapt either of the models. First, we remove model's classification head and replace it with a readout network composed of 4 dense layers, whose width gradually decreases from pre-trained model's latent dimension to 1 output channel. We first fit the new dense readout, to then unfreeze the rest of the model. After training the image- and text-based property regressors, we load them both to create a multimodal model. We remove the final layer of both readout dense networks. We concatenate the outputs of penultimate layers of our models, each having 256 channels, and create two new readout layers. Once again, we first fit the newly created layers to then unfreeze the remainder of the model. All dense readout layers, apart from the ultimate single-channel layer, are followed by ReLU activation, BatchNorm and dropout. To contrast this approach, we also conduct experiments with CLIP, a transformer-based model pre-trained on a text-image multimodal task. We add a single linear layer on top of the pre-trained model to get the regression output.

3 Experiments

First, we verify the transferability of the CLIP multimodal model to our physics domain problem. To do so, we use 85% of the dataset for cross-validation experiments with five folds, and keep the remaining 15% for further testing. Preliminary experiments have been conducted to optimize the hyperparameters of the model, and are reported in the Appendix Table(2). In Table(1), we report the performances of the CLIP model averaged over various trials. The higher and more stable performances are found for the regression of the bulk modulus, but overall we observe promising generalization capabilities in terms of R^2, on average around 0.8. We report examples of learning curves over different folds in the Appendix Fig.(6). Additional experiments have been conducted by

Table 1: Statistical results for regression of the target properties with the CLIP multimodal model showing the mean R^2 across validation folds, as well as the test set R^2 for each run, along with the average values. CLIP used the full textual description available ('All').

		E_F	E_{fermi}	K_{vrh}
		R^2/RMSE[eV/atom]	R^2/RMSE[eV]	R^2/RMSE[GPa]
#1	Val. Folds Average	0.91/0.21	0.90/0.89	0.91/21.9
	Test	0.72/0.36	0.87/0.90	0.88/24.4
#2	Val. Folds Average	0.88/0.23	0.90/0.88	0.88/24.1
	Test	0.89/0.24	0.82/1.25	0.80/35.3
#3	Val. Folds Average	0.77/0.35	0.90/0.86	0.92/21.1
	Test	0.67/0.35	0.51/2.07	0.80/36.5
$\langle\#\rangle$	**Val. Folds Average**	0.85/0.26	0.9/0.86	0.90/22.4
	Test set	0.76/0.31	0.73/1.41	0.83/32.1

taking the Fourier Transform of ECD images to test the importance of periodicity. The Appendix Fig(9) and Table(3) respectively report examples of such data and the related performances of the model, which however did not underline specific improvements.

Fig.(2) reports the regression performance of the stacked custom multimodal model during the different stages of its implementation. Already at the stage of InceptionV3 (image-only) finetuning, the bulk modulus prediction reaches high R^2 values in generalization, comparable to the CLIP model, and the subsequent finetuning of the RoBERTa (text-only) model or the composition of the semantic and visual embeddings from the finetuned models do not seem to provide visibile improvements. The prediction of the Fermi energy based on ECD images only, instead, does not generalize well, and remarkably benefits from the use of textual crystallographic information at all its levels (Formula, Keywords and All). The formation energy is a challenging prediction target, though it shows slight improvement as text input is added.

4 Discussion, Limitations and Future Work

In this work we propose computationally efficient approaches to property prediction in materials through the use of multiple modalities in transfer learning, with ECDs as highly informative input data. In particular, we show the dual approach of a ready-to-use pre-trained multimodal model like CLIP and the creation of a custom stack of single-modality pre-trained models. Even though the performances of the two approaches are similar, the latter stands as i) a more interpretable solution, because the single-modality stages convey the importance of each model and data type in the final combined performance; ii) more versatile solution, in it being composed of arbitrary single-modality models, which might be chosen appropriately for the planned application. We underline the highly informative nature of the data: the image of a single slice out of a three-dimensional field could be enough to build expressive features for the prediction of computationally expensive material properties properties like the bulk modulus, which generally requires calculating the total energy of the system under different volumes and then fitting to an equation of state. Furthermore, we observe remarkable regression improvements through inclusion of semantic embeddings, as opposed to only using visual information. It is an encouraging sign of the possibilities that multimodal approaches offer in such context.

The work presents some limitations and potential for future work. As we run independent trials for different variants of textual descriptions, we notice that multiple instances of image-only finetuning show palpable variance in performance, even though they are not influenced by varying text input (Fig. (2, Image-FT). This points at the need for multi-run experiments to quantify the variance in performance. An exciting avenue already under consideration could be the implementation of multimodal feature estimation in an inverse design procedure, which will be able to employ ECD fields and basic crystallographic information in a search for desired set of properties. Such approach would represent a remarkable leap in the field, connecting low-cost approaches, generative models and ECD prediction.

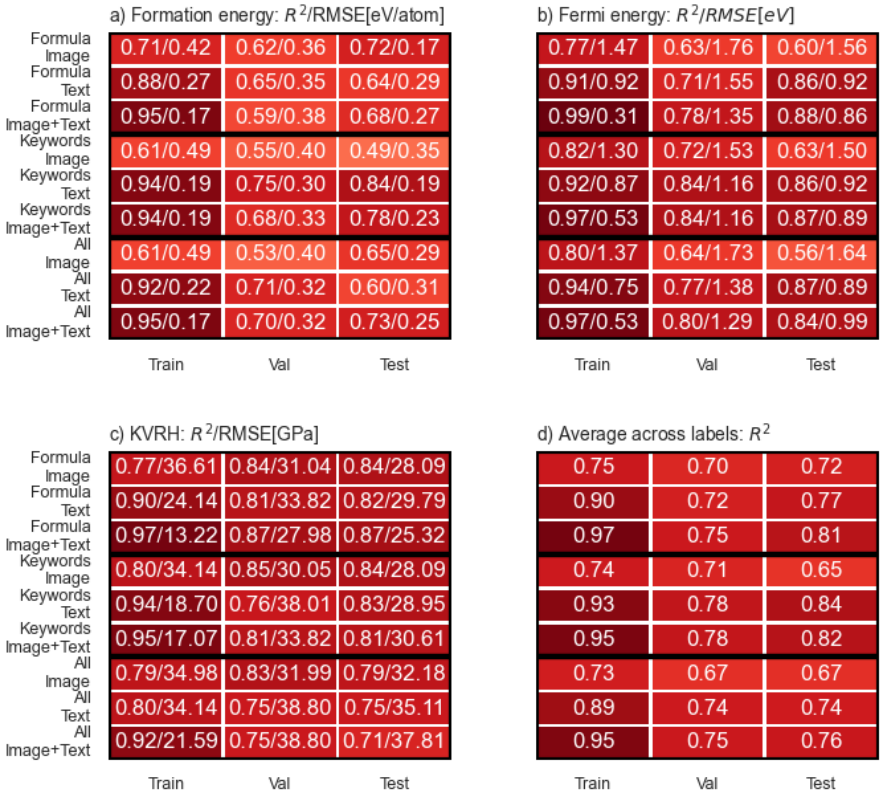


Figure 2: Regression performance of the different stages of a custom stacked multimodal model towards the three target properties: the formation energy, E_F, the Fermi energy, E_{fermi}, and the bulk modulus, K_{vrh}. The 'Image' only stage represents the fine-tuning (FT) of the InceptionV3 model. The 'Text' only stage represents the fine-tuning of the RoBERTa-Base model. The 'Image+Text' stage represents the fine-tuning of the embedding concatenation from the two modalities. Each of the stages involving text, also report the extent of textual information used, being it only the 'Formula', the 'Keywords' (composition name and bond lenghts information) or 'All' available text.

Acknowledgments and Disclosure of Funding

The Author(s) D. Massa and S. Papanikolaou were supported by the European Union Horizon 2020 research and innovation program under Grant Agreement No. 857470 and from the European Regional Development Fund under the program of the Foundation for Polish Science International Research Agenda PLUS, grant No. MAB PLUS/2018/8, and the initiative of the Ministry of Science and Higher Education 'Support for the activities of Centers of Excellence established in Poland under the Horizon 2020 program' under agreement No. MEiN/2023/DIR/3795. Piotr Sankowski was partially supported by ERC consolidator grant TUgbOAT No. 772346 and NCN No. 2020/37/B/ST6/04179.

References

[1] Ryszard Stanislaw Michalski, Jaime Guillermo Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[3] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3(1):54, 2017.

[4] Logan Ward and Chris Wolverton. Atomistic calculations and materials informatics: A review. *Current Opinion in Solid State and Materials Science*, 21(3):167–176, 2017.

[5] Krishna Rajan. Materials informatics. *Materials Today*, 8(10):38–45, 2005.

[6] Li-Chiang Lin, Adam H Berger, Richard L Martin, Jihan Kim, Joseph A Swisher, Kuldeep Jariwala, Chris H Rycroft, Abhoyjit S Bhown, Michael W Deem, Maciej Haranczyk, et al. In silico screening of carbon-capture materials. *Nature materials*, 11(7):633–641, 2012.

[7] Christopher E Wilmer, Michael Leaf, Chang Yeon Lee, Omar K Farha, Brad G Hauser, Joseph T Hupp, and Randall Q Snurr. Large-scale screening of hypothetical metal–organic frameworks. *Nature chemistry*, 4(2):83–89, 2012.

[8] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):095003, 2013.

[9] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

[10] Wesley T Hong, Roy E Welsch, and Yang Shao-Horn. Descriptors of oxygen-evolution activity for oxides: a statistical evaluation. *The Journal of Physical Chemistry C*, 120(1):78–86, 2016.

[11] Jeff Greeley, Thomas F Jaramillo, Jacob Bonde, IB Chorkendorff, and Jens K Nørskov. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nature materials*, 5(11):909–913, 2006.

[12] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

[13] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7), 2011.

[14] Fabio Pietrucci and Wanda Andreoni. Graph theory meets ab initio molecular dynamics: atomic structures and transformations at the nanoscale. *Physical review letters*, 107(8):085504, 2011.

[15] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

[16] Ali Sadeghi, S Alireza Ghasemi, Bastian Schaefer, Stephan Mohr, Markus A Lill, and Stefan Goedecker. Metrics for measuring distances in configuration spaces. *The Journal of chemical physics*, 139(18), 2013.

[17] Atsuto Seko, Atsushi Togo, Hiroyuki Hayashi, Koji Tsuda, Laurent Chaput, and Isao Tanaka. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. *Physical review letters*, 115(20):205901, 2015.

[18] D Xue, PV Balachandran, J Hogden, J Theiler, D Xue, and T Lookman. Accelerated search for materials with targeted properties by adaptive design. *Nat Commun*, 7:11241, 2016.

[19] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications*, 8(1):15679, 2017.

[20] T. Xie and J. C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, 2018.

[21] C. Chen et al. Graph net- works as a universal machine learning framework for molecules and crystals. *Chem. Mater.*, 31(9):3564, 2019.

[22] S. Back et al. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *J. Phys. Chem. Lett.*, 10:4401–4408, 2019.

[23] J. Jorgensen et al. Neural message passing with edge updates for predicting properties of molecules and materials. *arXiv.org e-Print archive*, page arXiv:1806.01261, 2018.

[24] K. Schütt et al. Schnet: a continuous-filter convolutional neural network for modeling quantum interactions. *NeurIPS*, 30:991–1001, 2017.

[25] D. K. Duvenaud et al. Convolutional networks on graphs for learning molecular fingerprints. *NeurIPS*, page 2224–2232, 2015.

[26] Z. Wu et al. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530, 2018.

[27] E. Saal et al. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd). *JOM*, 65(1501), 2013.

[28] A. Jain et al. The materials project: A materials genome approach to accelerating materials innovation. *APL Mater*, 1(011002), 2013.

[29] S. Curtarolo. A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.*, 58(227), 2012.

[30] J. Hachmann et al. The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.*, 2(2241), 2011.

[31] Stefano Curtarolo, Gus LW Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, and Ohad Levy. The high-throughput highway to computational materials design. *Nature materials*, 12(3):191–201, 2013.

[32] James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65:1501–1509, 2013.

[33] Walter Kohn, Axel D Becke, and Robert G Parr. Density functional theory of electronic structure. *The journal of physical chemistry*, 100(31):12974–12980, 1996.

[34] Aron J Cohen, Paula Mori-Sánchez, and Weitao Yang. Challenges for density functional theory. *Chemical reviews*, 112(1):289–320, 2012.

[35] Dario Massa, Efthimios Kaxiras, and Stefanos Papanikolaou. Alloy informatics through ab initio charge density profiles: Case study of hydrogen effects in face-centred cubic crystals. *Acta Materialia*, 268:119773, 2024.

[36] Teddy Koker, Keegan Quigley, Eric Taw, Kevin Tibbetts, and Lin Li. Higher-order equivariant neural networks for charge density prediction in materials. *npj Computational Materials*, 10(1):161, 2024.

[37] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[39] Jimmy-Xuan Shen, Jason M. Munro, Matthew K. Horton, Patrick Huck, Shyam Dwaraknath, and Kristin A. Persson. A representation-independent electronic charge density database for crystalline materials, 2021.

[40] A. M. Ganose and A. Jain. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9, 2019.

A Appendix / supplemental material

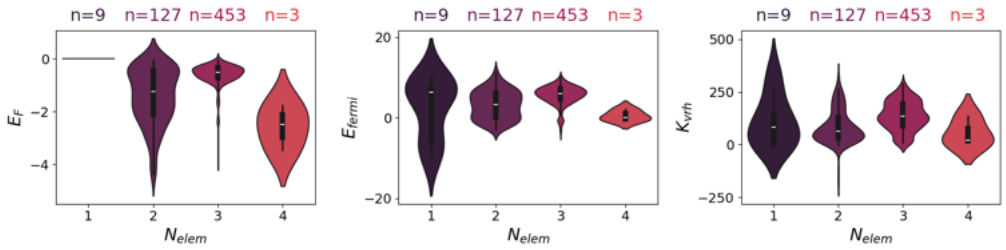
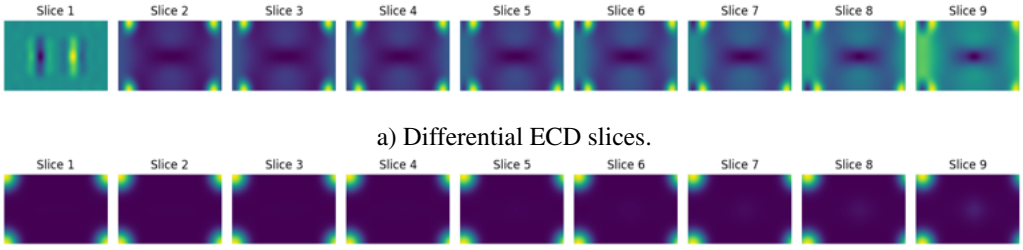


Figure 3: Visualization of the samples distributions according to the three target properties, the formation energy per atom (E_F in eV/atom), the Fermi energy (E_{fermi} in eV) and the bulk modulus (K_{vrh} in GPa), and to the number of constitutive atomic species (N_{elem}). For each violin plot, a number n on the upper edge of the frame indicates the corresponding number of samples. The plot clearly indicates the majority of samples being three-species compounds. A wider (narrower) horizontal section of a single distribution instead indicates more (fewer) data points are present at that value. The white mark indicates the median, while the black bar the interquartile (IQR) range. The distributions here shown correspond already to the cleaned version of the dataset, with 592 samples.



a) Differential ECD slices.

b) Total ECD slices.

Figure 4: An example of 2D slices along the z-axis of the ECD volumetric data for an MP sample (mp-72, Ti).

Table 2: Different trials (T#) for the hyperparameters used during the finetuning of the CLIP model and associated performance metrics. The images correspond to ECD images of the samples in viridis colorscale, while the target property for regression is always the formation energy.

	BS	LR	Patience	Patch	Ratio_T	$\langle \text{R}^2 \rangle_\text{f}$	$\langle \text{RMSE} \rangle_\text{f}$	R^2_T	RMSE_T
T#1	16	0.00015	17	32	0.15	0.6373	0.5873	0.8476	0.4127
T#2	32	0.00015	50	32	0.15	0.9064	0.2973	0.7792	0.4967
T#3	16	0.00010	50	32	0.15	0.8548	0.3670	0.8942	0.3438
T#4	32	0.00015	50	16	0.15	0.8298	0.4033	0.6508	0.6246
T#5	16	0.00010	50	16	0.15	0.6988	0.4990	0.7123	0.5670

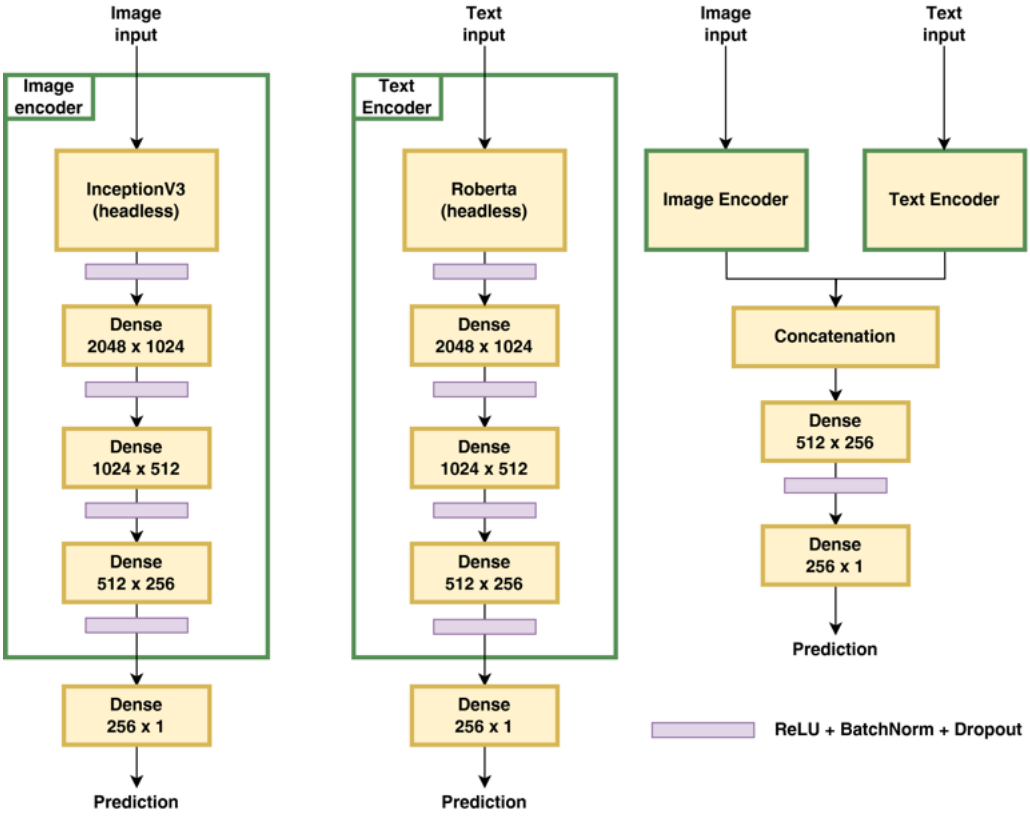


Figure 5: Schematic representation of the stacked custom multimodal model architecture. From the left: image-only model, text-only model, and a combined multimodal model that utilizes them both.

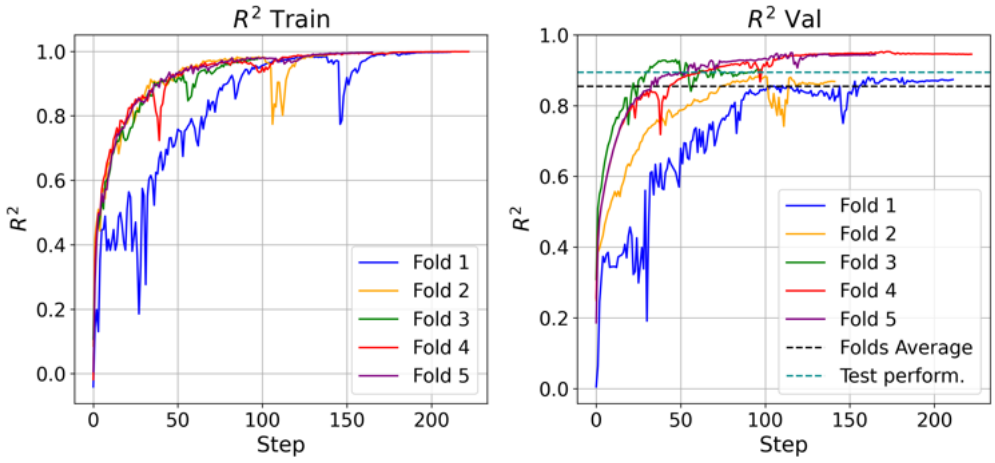


Figure 6: Performance of the selected hyperparameters set (T#3) on the ECDs dataset composed of images in viridis scale, in the regression of the formation energy property.

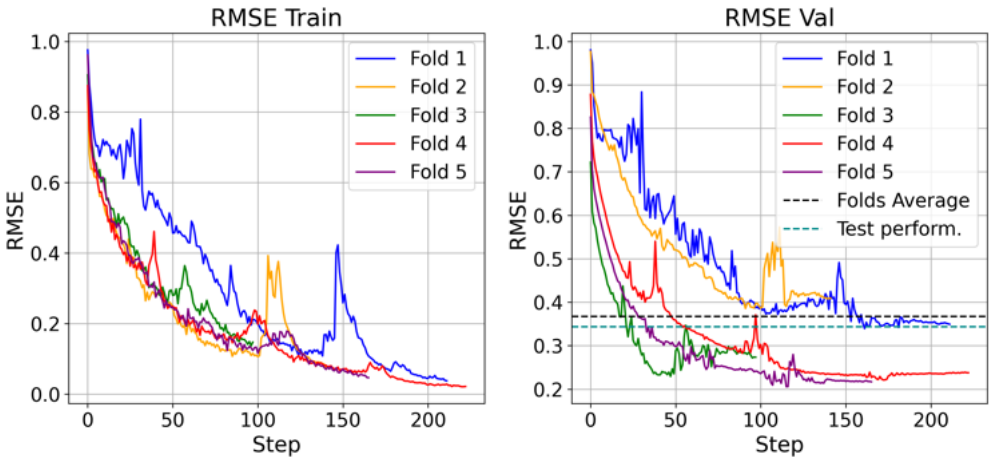


Figure 7: Performance of the selected hyperparameters set (T#3) on the ECDs dataset composed of images in viridis scale, in the regression of the formation energy property. Being the target property standardized, the RMSE is adimensional.

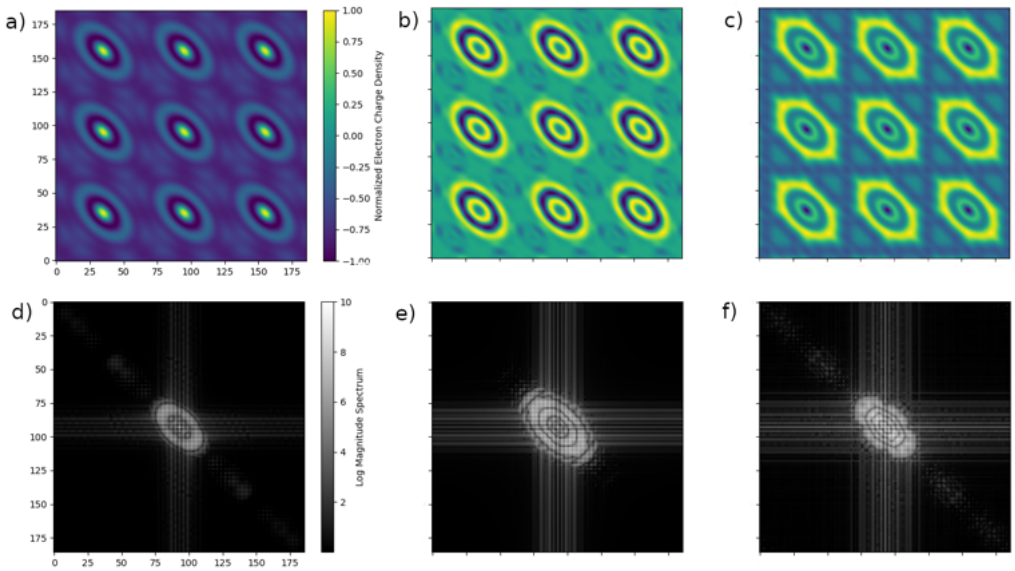


Figure 8: a-c) Visualization in viridis colorscale of three examples of periodic and normalized ECD patterns extracted from the center of the sliced planes. d-f) Fourier transforms of the upper ECD image data.

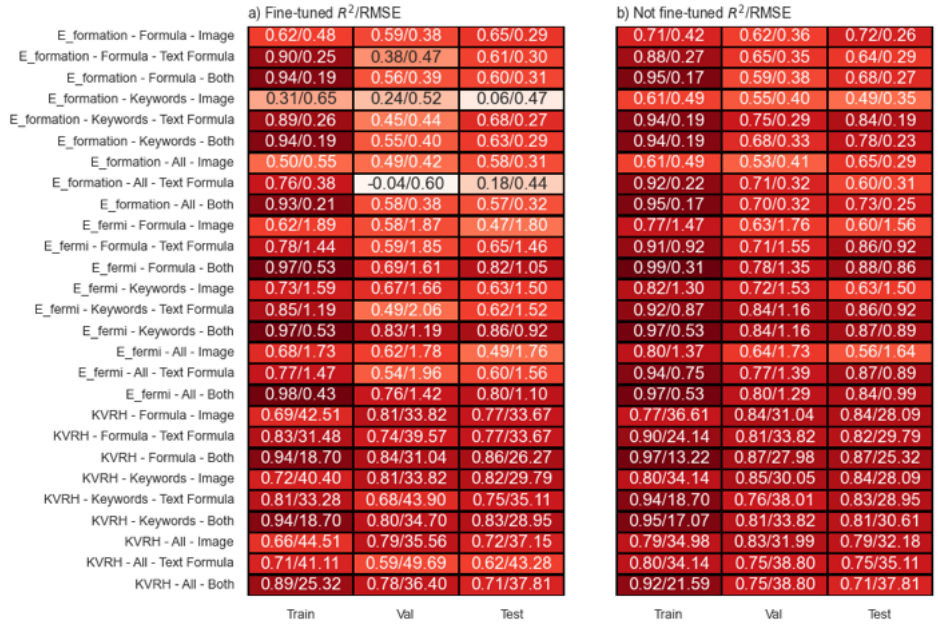


Figure 9: Performance of the models before and after fine-tuning.

Table 3: Different trials (T#) for the hyperparameters used during the finetuning of the CLIP model and associated performance metrics. The images correspond to the Fourier transform of the ECD images in grey scale, while the target property for regression is always the formation energy.

	BS	LR	Patience	Patch	Ratio_T	$\langle \text{R}^2 \rangle_\text{f}$	$\langle \text{RMSE} \rangle_\text{f}$	R^2_T	RMSE_T
T#1	16	0.00015	17	32	0.15	0.43	0.73	0.65	0.62
T#2	32	0.00015	50	32	0.15	0.84	0.38	0.72	0.56
T#3	16	0.00010	50	32	0.15	0.90	0.29	0.68	0.60
T#4	32	0.00015	50	16	0.15	0.81	0.41	0.72	0.56
T#5	16	0.00010	50	16	0.15	0.86	0.35	0.68	0.60

2.3 Manuscript III

"Substitutional alloying using crystal graph neural networks"

Dario Massa, Daniel Cieslinski, Amirhossein Naghdi and Stefanos Papanikolaou

Commentary

In the context of direct extraction of atomic descriptors, we here consider a particular open-source CGCNN model, MEGNet (MatErials Graph Network), in its pre-trained stage on the Materials Project (MP) database for the regression of a number of physical properties of bulk crystalline materials. The aim of the work is to propose a systematic procedure to explore and test the transferability of such models towards property predictions in *parent* systems of the training set, namely bulk crystals which have undergone minimal changes in their composition due to substitutional defecting. In particular, we propose two strategies of composition manipulation: (i) a specific single atom specie substitution in a large variety of host matrices, focusing on three quite different atoms for the substitution (H, Mn and Rb); and (ii) a specific host crystals set with a variety of single atom species substitutions, focusing on Al, Ni, Mo and Au as host matrices and almost all the periodic table of elements for the single atom substitution. We provide a small DFT validation set on the Al_X case, which shows promising results in the prediction of structural properties, considered the absence of any further training of the model on defected systems. However, we notice much worse predictions on the formation energies, leaving space for future work on larger validation sets and investigations for such deviations.

In this study, the PhD candidate contributed to: conceptualization of the work; DFT calculations; writing of the manuscript draft and all subsequent versions according to coauthors suggestions, plots.

RESEARCH ARTICLE | JANUARY 18 2024

Substitutional alloying using crystal graph neural networks

Dario Massa ; Daniel Cieśliński ; Amirhossein Naghdi ; Stefanos Papanikolaou

AIP Advances 14, 015023 (2024)
https://doi.org/10.1063/5.0163765

AIP Advances ARTICLE pubs.aip.org/aip/adv

Substitutional alloying using crystal graph neural networks

Cite as: AIP Advances 14, 015023 (2024); doi: 10.1063/5.0163765
Submitted: 11 December 2023 • Accepted: 12 December 2023 •
Published Online: 18 January 2024

Dario Massa, Daniel Cieśliński, Amirhossein Naghdi, and Stefanos Papanikolaou[a]

AFFILIATIONS

NOMATEN Centre of Excellence, National Centre for Nuclear Research, uł. Andreja Sołtana 7, Otwock, Poland

[a]Author to whom correspondence should be addressed: stefanos.papanikolaou@ncbj.gov.pl

ABSTRACT

Materials discovery, especially for applications that require extreme operating conditions, requires extensive testing that naturally limits the ability to inquire the wealth of possible compositions. Machine Learning (ML) has nowadays a well-established role in facilitating this effort in systematic ways. The increasing amount of available accurate Density Functional Theory (DFT) data represents a solid basis upon which new ML models can be trained and tested. While conventional models rely on static descriptors, generally suitable for a limited class of systems, the flexibility of Graph Neural Networks (GNNs) allows for direct learning representations on graphs, such as the ones formed by crystals. We utilize crystal graph neural networks (CGNNs) known to predict crystal properties with DFT level accuracy through graphs by encoding the atomic (node/vertex), bond (edge), and global state attributes. In this work, we aim at testing the ability of the CGNN MegNet framework in predicting a number of properties of systems previously unseen in the model, which are obtained by adding a substitutional defect to bulk crystals that are included in the training set. We perform DFT validation to assess the accuracy in the prediction of formation energies and structural features (such as elastic moduli). Using CGNNs, one may identify promising paths in alloy discovery.

I. INTRODUCTION

The use of machine learning (ML)[1,2] methods in materials science to accelerate materials discovery[3] is at the base of the so-called materials informatics (MI).[4–8] By training ML models on large databases, such as the OQMD (Open Quantum Materials Database) or the Materials Project high-throughput electronic structure calculation databases,[9–13] the goal is to achieve predictions of material properties with quantum accuracy.

As in statistical mechanics, with the need for identifying appropriate order parameters of novel phases and structures, the key challenge in ML algorithms is to identify effective system descriptors that can function as structure identifiers. A large variety of descriptors have been proposed, including fixed-length feature vectors of material, elemental, or electronic properties,[14–16] structural descriptors based on rotational and translational invariant transformation of atomic coordinates, such as the Coulomb matrix,[17] atom-centered symmetry functions (ACSFs),[18] social permutation invariant coordinates (SPRINTs),[19] smooth overlap of atomic positions (SOAP),[20] and global minimum of the root mean square distance.[21] However, these solutions are often system-specific and are not suitable for vast compositional and structural space exploration.

For this reason, a topic of fervent interest in the materials science community is the use of graph neural networks (GNNs),[22,23] which allow learning representations directly and in a flexible way, focused on molecular systems,[24–29] surfaces,[30–32] and periodic crystals.[25,33–39] GNNs can be regarded as the generalization of convolutional neural networks (CNNs) to graph-structured data, from which the internal materials representations can be learned and used for prediction of target properties;[40] even though larger amounts of data are required with respect to conventional ML models, GNNs take advantage of the unambiguous physics-guided real-space local associations between the system's degrees of freedom. Hence, they can be used for any type of atomic crystalline structure.[41] The common idea of GNN-based models is to represent atoms as nodes (V) and their chemical bonds as edges (E) in a graph $G(V, E)$, which can be fed to a trained neural network to create node-level embeddings (learned representation of each atom in its individual chemical environment) through convolutions with neighboring nodes and edges.[42] Therefore, in the context of a regression problem in a set

of target material properties y (output), given a graph-based data-structure G (input) encoding relevant features for the problem and a set of learnable weights W (model parameters) learnable from the data to make predictions on the input G, the GNN architecture reformulates the prediction task as the search for a mapping function f such that $f(G;\ W) \rightarrow y$.

A direct benefit of the crystal material GNN-converted graph encoding is the naturally derived vector characterization of the atoms and edges.[37] The work by Xie and Grossman[33] presented a pioneering example of a crystal graph convolutional neural network (CGCNN) architecture, which was later extended to the iCGCNN (*improved*-CGCNN) by Park and Wolverton[38] to include 3-body correlations on neighboring atoms, information on the Voronoi tessellated structure, and an optimized chemical representation of interatomic bonds in the crystal graphs.

For the discovery of new materials, one may take various exploring paths, involving high-throughput computational[3] and experimental[43] methods. However, the combined approach of machine-learning methods and compositional manipulation has very quickly acquired a well-established role in materials science, and it is applied in a wide range of property optimization searches, such as for zinc blende semiconductors,[44] perovskites,[45–50] and others.[51–58]

In this work, we utilize a particularly improved model of the originally proposed[33] CGCNN model, the MatErials Graph Network (MEGNet) model from Chen *et al.*,[35] which is introduced in Sec. II A 1 and which has the merit of being developed and tested both on molecules and crystals, with the possibility of defining global state attributes including temperature, pressure, and entropy. Indeed, the good performance of the MEGNet model has been assessed and compared in a multitude of works[36,59,60] on different databases. The Materials Project database, for example, contains multi-component systems, and the question of having a model being trained on them implies the possibility for it to be used on a new, or slightly different, alloy composition; in addition, to which errors its predictions would be consequently subject to remains, to our knowledge, unanswered.

We aim at assessing the capabilities of graph networks to predict the properties of single-atom substitutionally defected crystals, which do not belong to the dataset the model has been trained or tested on, questioning the predictive performances and therefore testing the transferability of the model learned knowledge of local atomic environments in systems where model-known atomic species are contained in unknown arrangements. We tackle these questions in the very low regime of defect contamination of substitutional alloying, with the aim of "isolating," or simplifying, the possible origins of differences from the known crystalline structures, shedding light over the model's knowledge. From the point of view of the selected samples, we propose two procedures of substitutional alloying manipulation: (i) a single atom substitution of a specific species in a variety of host matrices and (ii) a single atom substitution from a variety of species in a specific host matrix. From the point of view of the choice of the model, the simplification comes through the minimal input information used for MEGNet in this work: the atomic number for the nodes and the interatomic distances for the edges. After considering a pre-trained model on the Materials Project (MP) database (Sec. II A 3), we focus on the formation energies and bulk and shear modulus predictions, both comparing the results obtained in datasets of similarly defected structures (Sec. III A) and the effects of almost all the possible single-atom defects in the same matrices (Sec. III B). To validate the predictions, as described in Secs. II B and III C, we perform Density Functional Theory (DFT) calculations, and we find that CGNNs have both great potential and also limitations in predicting properties of defected bulk crystals, thus promoting materials discovery.

II. METHODS

A. Machine learning framework

1. MEGNet description

In the present work, we utilize the MEGNet model.[35] The reasons for this choice lie in the structure and performance of the model:

1. It is characterized by a low number of attributes, one for the atom (atomic number) and one for the bond (spatial distance), but MEGNet outperforms previous graph-based models,[35] such as the CGCNN[33] and MPNN,[24] with higher number of attributes, as well as SchNet,[25] with a similar low number.
2. The MEGNet framework includes a global state attribute, essential for state-property relationship predictions in materials.
3. The graph network of MEGNet has been developed and tested for both molecules and crystals. Here, we limit ourselves and present the main features of the model, but for a more exhaustive explanation, we recommend the reader to refer to the original work by Chen *et al.*[35] and the references therein. In particular, for a graph $G(E, V, \mathbf{u})$,

- V is the set of N_v atomic attribute vectors $\mathbf{v}_i$,
- $E = \{(\mathbf{e}_k, r_k, s_k)\}_{k=1\ldots N_e}$ is the set of N_e bond attribute vectors $\mathbf{e}_k$, with r_k and s_k being the indices of the atoms forming the k-th bond, and
- $\mathbf{u}$ is the global state attribute vector. The role of graph networks is to recursively update the input graph $G(E, V, \mathbf{u})$ to an output graph $G(E', V', \mathbf{u}')$ with progressive and inclusive information flow going from bonds to atoms, and finally to the global state. In particular, first the attributes of each bond are updated through a function ϕ_e. Applied on the concatenation of the self-attributes, the ones of the connecting $\mathbf{v}_{s_k}$ and $\mathbf{v}_{r_k}$ atoms, and of the global state $\mathbf{u}$, are

$$\mathbf{e}'_k = \phi_e(\mathbf{v}_{s_k} \oplus \mathbf{v}_{r_k} \oplus \mathbf{e}_k \oplus \mathbf{u}), \tag{1}$$

where $\oplus$ is the concatenation operator.

The update of atomic attributes involves the average over the i-th atom connecting bonds $\bar{\mathbf{v}}_i^e = \frac{1}{N^e}\sum_{k=1}^{N_i^e}\{\mathbf{e}'_k\}_{r_k=i}$, the i-th atom self-attributes $\mathbf{v}_i$, and the global state ones $\mathbf{u}$, as follows:

$$\mathbf{v}'_i = \phi_v\left(\bar{\mathbf{v}}_i^e \oplus \mathbf{v}_i \oplus \mathbf{u}\right). \tag{2}$$

Finally, the information flow from all three attribute groups involves updating the global state attributes, as follows:

$$\mathbf{u}' = \phi_u\left(\bar{\mathbf{u}}^e \oplus \bar{\mathbf{u}}^v \oplus \mathbf{u}\right), \tag{3}$$

where $\bar{\mathbf{u}}^e = \frac{1}{N^e}\sum_{k=1}^{N^e}\{\mathbf{e}'_k\}$ and $\bar{\mathbf{u}}^v = \frac{1}{N^v}\sum_{i=1}^{N^v}\{\mathbf{v}'_i\}$.

TABLE I. Parameters from the pre-trained MEGNet model.

Parameter	Value	Short description
nfeat_node	94	Number of atom features
nfeat_global	2	Number of state features
ngauss_centers	110	Number of Gaussians
converter_cutoff	4	Cutoff radius
megnet_blocks	3	Number of MEGNetLayer blocks
Optimizer	Adam	Optimizer of the model weights
lr	1×10^{-3}	Learning rate
n1	64	Number of hidden units in layer 1
n2	32	Number of hidden units in layer 2
n3	16	Number of hidden units in layer 3

As mentioned before, for our systems of interest, namely, periodic crystals, the atomic number is the only atomic attribute for each $\mathbf{v}_i$ in the set V. For bonds, the spatial distance is expanded in a Gaussian basis set, centered at a linearly spaced r_0 location between $r_0 = 0$ and $r_0 = r_{\text{cut}}$ and characterized by a given width σ; it therefore has a shape $\exp\left(-(r - r_0)^2/\sigma^2\right)$. Finally, the global state is simply a two zero placeholder for global information exchange.

2. Data collection

We consider crystal structures collected through the Python Materials Genomics interface (pymatgen)[61] for the Materials Application Programming Interface from Materials Project.[10] When creating the dataset, there were 126 301 structures in the database that had formation energy (E_{form}) property and 13 102 structures that had bulk modulus (K_{VRH}) and shear modulus (G_{VRH}) properties in the Voigt–Reuss–Hill (VRH) approximation.[62]

3. Pre-trained model

Our focus in this work is on the prediction capabilities of MEGNet for *minimally* defected systems that are clearly not in the training database, given the training of a dataset of undefected structures. In order to do so, we consider the substitution of a single atom in a supercell, hoping that CGNN training captures atomic similarities, based on combinations of atomic radii, valence electrons, and other atomic properties. Table I shows some of the parameters of the model, and a more complete list can be found at the default implementation of the class.[63]

TABLE II. MAEs of the model for the prediction of the bulk modulus (K_{VRH}), shear modulus (G_{VRH}), and formation energy (E_{form}).

Property	MAE
K_{VRH}	6.143 GPa
G_{VRH}	10.489 GPa
E_f	0.029 eV/atom

We report parity plots of Fig. 1 for all three properties of interest in this study: bulk modulus (K_{VRH}), shear modulus (G_{VRH}), and formation energy (E_{form}). To evaluate the model accuracy in predicting the properties of interest for the present study, the mean-absolute error (MAE) is used as the evaluation metric. Table II presents the MAE values for each predicted property over the dataset, which provides insights into the pre-trained model performance.

B. Validation with DFT

We verify the accuracy of the model's predictions for system properties such as bulk modulus K_{VRH}, shear modulus G_{VRH}, and formation energy E_{form} after single-atom substitution is implemented in $2 \times 2 \times 2$ supercells. We perform DFT calculations with `Quantum Espresso (QE)`[64–66] and its `THERMO_PW`[67] driver for the calculation of structural properties. Pseudo-potentials for all involved atomic species are ultra-soft and with the Perdew–Burke–Ernzerhof (PBE)[68] functional. Methfessel–Paxton smearing[69] has been introduced to correctly investigate metallic systems, and the calculations have been set as spin-polarized, for possible non-zero magnetization effects. Convergence is checked on the number of k-points and plane-wave cutoff energy. In addition, the energy smearing is spread (`degauss` parameter in QE) for

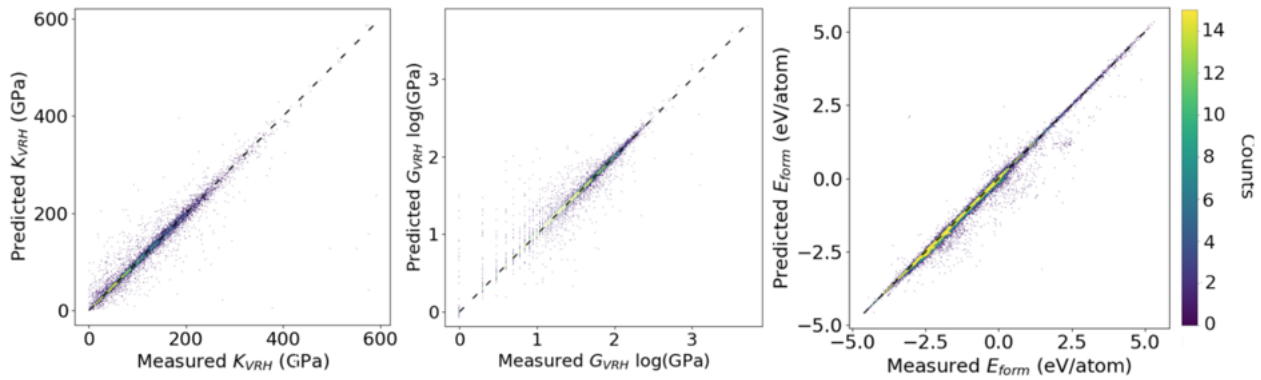


FIG. 1. Parity plots for the pre-trained model on the MP dataset. The plots involve the predictions on (a) the bulk modulus (K_{VRH}), (b) the shear modulus (G_{VRH}), and (c) the formation energy (E_{form}).

TABLE III. Differences in the atomic properties of the three elements considered for the single-atom substitutional process. As an example, we report here the atomic weight, radius, and electronic configuration.

	Weight (u)	Radius (pm)	El. configuration
H	1.008	53	$1s^1$
Mn	54.938	161	[Ar] $4s^2$ $3d^5$
Rb	85.468	265	[Kr] $5s^1$

each case after a preliminary variable-cell relaxation of pure crystals and further fixed-cell relaxation with optimal parameters for final equilibrium bulk structures. The common acceptance threshold in the variation in the total energy upon parameter change is set at 10^{-5} Ry. Forces and total energy convergence thresholds for ionic minimization are set to a common value of 10^{-5} and 10^{-6} a.u., respectively. After optimization of pure crystals, fixed-cell relaxation is performed on supercells with single-atom substitutions, and its structural properties are then extracted through the `THERMO_PW` driver.

The computation of formation energies for validation purposes is performed for the case of single-atom substitutional defects (D) applied on pure bulk crystals (M), as follows:

$$E_{\text{form}}(\mathrm{M_{1-x}D_x}) = E(\mathrm{M_{1-x}D_x}) - (1 - \mathrm{x})E(\mathrm{M}) - \mathrm{x}E(\mathrm{D}), \quad (4)$$

where x is the atomic fraction of substitutional defects, $E(\mathrm{M_{1-x}D_x})$ is the total energy per atom of the compound, and $E(\mathrm{M})$ and $E(\mathrm{D})$ are the total energies per atom of the precursor species that compose it; the latter energies are obtained by relaxing the ground-state lattices of these species using the same aforementioned QE parameters of the compound as necessary to ensure computational consistency in the evaluation of accurate values for the defect formation energies.

III. RESULTS: PROPERTY PREDICTIONS FOR SUBSTITUTIONAL ALLOYING WITH CGNNs

There is a large variety of ways to systematically evaluate the effect of substitutional alloying on the properties of crystals.[70] Here, we focus on two key questions:

1. Data science of defects: what is the effect of substituting the same defect in a large variety of systems?
2. Are there qualitative and quantitative effects from atom-substituting various elemental defects in the same host crystalline matrix?

While it has not yet been possible to perform an exhaustive search of the kind, in this work, for the first part, the basin of host systems is represented by the crystals dataset of Materials Project introduced in the previous sections,[10] while for the latter part, the host systems are pure metallic bulk crystalline supercells of Al, Ni, Mo, and Au.

A. An elemental defect seeing a wealth of different crystalline environments

First, we focus on the prediction of system properties, where the systematic substitutional defective process is applied using the same replacement atom on randomly selected sites of crystals that exist in the crystals dataset of Materials Project.[10] The aim is to explore how material properties change after single-atom substitution, evaluating deviations from original pure crystalline predictions, and how they depend on the specific substitution. For this, we consider three elemental cases, Rb, Mn, and H as the key replacement atoms that we will mutually compare. The reason for the choice lies in the drastic elemental differences among their unique characteristics and the assumption that, on the base of these, we might be able to gain a deeper understanding on how the model behaves in the prediction of previously unseen defect-induced changes in the system properties, based on its learned notion of the local environment. Table III

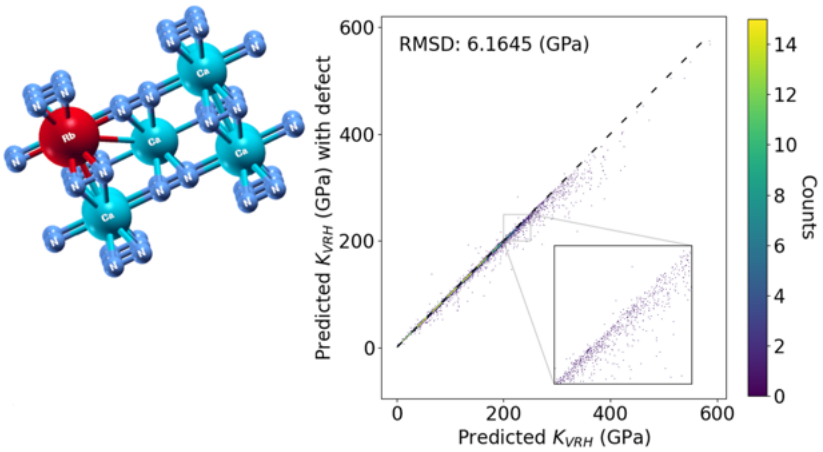


FIG. 2. Predicted bulk moduli in the Rb-defected MP crystals with respect to their prediction in non-defected ones (b). An example structure from the MP dataset is shown, CaN_2, in which one atom of Ca has been replaced with a Rb atom (a). Here, only the case of Rb-defected systems is shown due to its largest RMSD among the set of considered defects. Similar plots for the Mn- and H-defected systems are shown in the Appendix.

TABLE IV. RMSD (GPa) for the prediction of the bulk modulus (K_{VRH}) in Rb, Mn, and H single-atom substitutionally defected systems with respect to the non-defected ones.

Defect	RMSD (GPa)
Rb	6.1645
Mn	3.1274
H	1.2975

TABLE V. RMSD [log (GPa)] for the prediction of the shear modulus (G_{VRH}) in Rb, Mn, and H single-atom substitutionally defected systems with respect to the non-defected ones.

Defect	RMSD [log (GPa)]
Rb	0.0628
Mn	0.0350
H	0.0435

reports, as an example, the values of the atomic weight, radius, and electronic configuration of each considered replacement atom.

In Fig. 2, we display the predictions of MEGNet for the bulk moduli of Rb-defected systems with respect to the original, pure ones. As shown in Table IV, this kind of substitutional defect causes the largest root-mean-square (RMS) deviations among the three considered atomic species, with a value of ~6.2 GPa. In this section, we only report plots referring to the largest RMSD cases, but analogous plots can be found in Subsection 1 of the Appendix.

In addition, the shear modulus predictions show the largest RMSD for the case of a Rb-defect, with a value of 0.0628 log (GPa). As it can be visible from an analogous plot in the Appendix, without a log–log scale, the deviations cannot be correctly estimated. This is the reason why we decided to report the G_{VRH} prediction plots and RMSD values in these units. The results are shown in Fig. 3 and in Table V. According to the model, it seems that both K_{VRH} and G_{VRH} upon substitution of a Rb atom show a tendency of decrease in their value, implying an increase in their compressibility and decrease in hardness. It is worth noticing that even though the only physical atomic feature that the model exploits is the atomic number, the prediction of larger changes in the structural properties for involved defects with larger radii can be regarded as a reasonable one.

Furthermore, in Fig. 4, we show the predicted deviations in the formation energy of the MP-defected samples for the case of elemental H, and also for all three elements in Table VI. As it can be seen, the deviations are all in the order of 0.02 eV/atom and are progressively less prominent going from H to Rb and Mn.

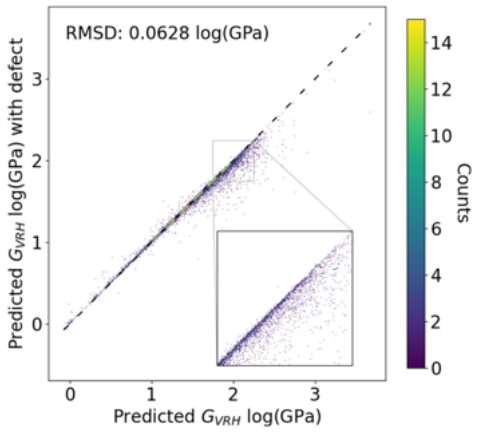


FIG. 3. Predicted shear modulus in the Rb-defected MP crystals with respect to the prediction in non-defected ones. Here, only the case of Rb-defected systems is shown due to its largest RMSD among the set of considered defects. Similar plots for the Mn- and H-defected systems are shown in the Appendix.

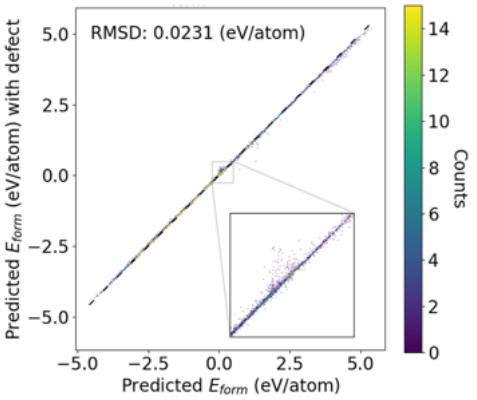


FIG. 4. Predicted formation energy in the H-defected MP crystals with respect to the prediction in non-defected ones. Here, only the case of H-defected systems is shown due to its largest RMSD among the set of considered defects. Similar plots for the Rb- and Mn-defected systems are shown in the Appendix.

B. A crystalline matrix seeing a wealth of different elemental defects

The case of systematically changing the single-atom substitutional defect species in the same crystalline matrix is complementary to the prior addressed one. We focus on pure metallic bulk host matrices, namely, Al, Ni, Mo, and Au, even though the test could be carried out on any crystalline matrix. In Fig. 5, we show the key aspects of the performed calculation, with the supercell shown on the left (a) and the elements shown on the right (b). The compositional space considered for the single-atom substitution covers almost the entire Periodic Table and also comprises the species highlighted as host matrices when they are not selected as such.

TABLE VI. RMSD (eV/atom) for the prediction of the formation energy (E_{form}) in Rb, Mn, and H single-atom substitutionally defected systems with respect to the non-defected ones.

Defect	RMSD (eV/atom)
Rb	0.0189
Mn	0.0150
H	0.0231

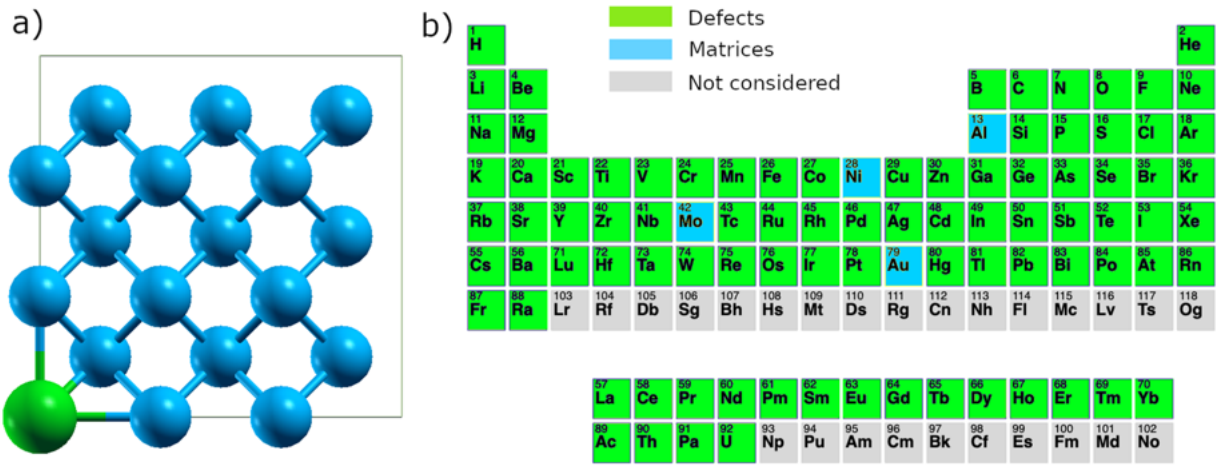


FIG. 5. Periodic Table plot (b) explaining the process of selecting a set of host matrices (in light blue) and substitutional defects (in green). To a given selected host matrix, the non-selected ones represent defects too. The host matrices are 3 × 3 × 3 supercells of the highlighted species (a).

Even though our focus is on simple systems, the process can be insightful on the capabilities of the model to learn and predict, with minimal input, physio-chemical trends throughout the Periodic Table.

We show Periodic Table plots for the properties of interest, K_{VRH}, G_{VRH}, and E_{form}, for the host matrix, which displays the largest deviations in the properties. This is the case of Mo, but analogous plots for the other matrices are included in the Appendix. With the help of the visualization style, we characterize the predictions of the effect on the chemical distance between the substitutional defects and the host matrix on the properties of interest and how it correlates with the well-known trends along the Periodic Table. In Fig. 6, we show the prediction of the host supercell bulk modulus variation when defected with one of the elements from the Periodic Table. In particular, in the plot, we refer to

$$K'_{\mathrm{VRH}} = K_{\mathrm{VRH}}^{\mathrm{Mo(X)}} - K_{\mathrm{VRH}}^{\mathrm{Mo}}, \tag{5}$$

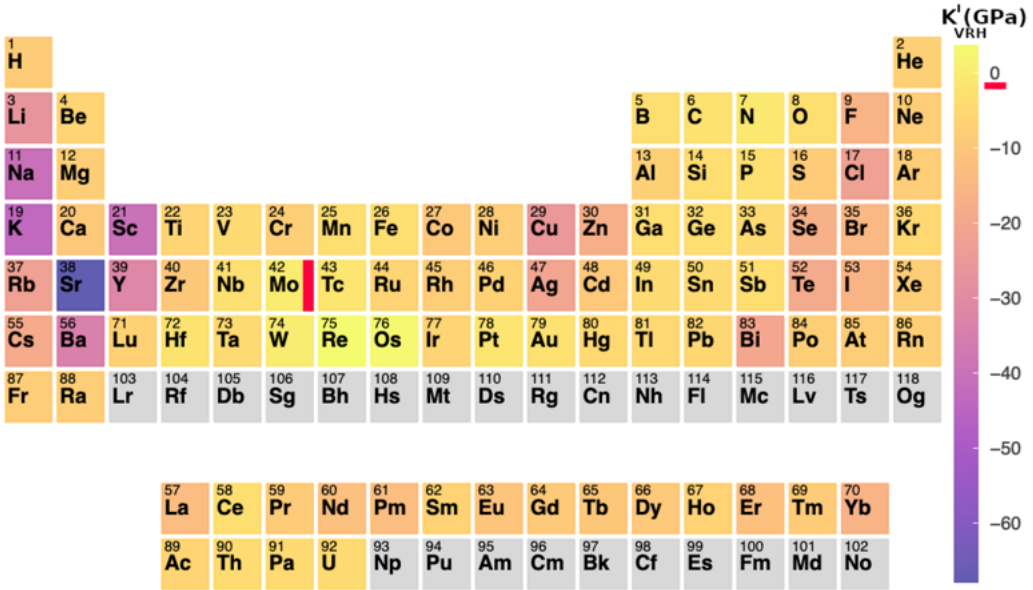


FIG. 6. Predicted bulk modulus variation (K'_{VRH}) for a single-atom substitutionally defected Mo supercell with respect to the undefected one, for each possible defect atomic species from the provided Periodic Table. Similar plots for Al, Au, and Ni supercells are provided in the Appendix. The red flag highlights the zero relative difference, meaning the pure Mo matrix selection.

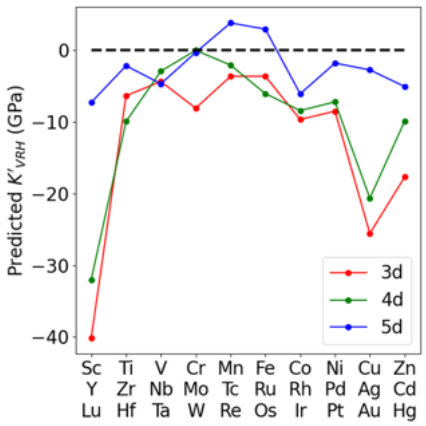


FIG. 7. Predicted bulk modulus variation (K'_{VRH}) for a single-atom substitutionally defected Mo supercell with respect to the undefected one along the 3d, 4d, and 5d series of the Periodic Table. The black dashed line highlights the pure Mo case.

where $K_{VRH}^{Mo(X)}$ is the bulk modulus of the Mo matrix defected by atom X and K_{VRH}^{Mo} is its value for pure Mo (labeled with a red flag in the figure).

Even though Sr represents an outstanding outlier in decreasing the bulk modulus of the Mo crystal, at this scale, it is still possible to appreciate how the variation happens along the periods: for the 3d, 4d, and 5d transition metals from third to 12th group, the defect-induced variations are mainly small, while a tendency to increase is observed in modulus, in the post-transition metals, and, remarkably, in the alkali and alkaline-earth metals. This behavior can be interpreted in terms of the well-known variations in the bulk modulus along the Periodic Table of elements, which seem correlated with respect to the defect-induced ones contained in this plot. The effect of substitutional alloying species, which under their bulk and standard temperature and pressure (STP) conditions show a lower bulk modulus, shows a tendency to lower the bulk modulus of their host system.

Due to the strong effect induced by the Sr defect, we further focus on transition metals to evaluate how defect-induced variations fluctuate in the compositional vicinity of the host matrix. Figure 7 shows the results along the list of 3d, 4d, and 5d transition elements, where the variations are harder to distinguish from the prior Periodic Table plot. We can recognize a decreasing trend that is supportive of our interpretation, but in the compositional vicinity of the host matrix, the fluctuations in the defect-induced variations are comparable with the MAE on bulk modulus prediction, just as the differences between the curves.

However, in view of a trend analysis of the variations along the Periodic Table, alkali (such as K), alkaline-earth (such as Sr and Ba), and post-transition metals (such as Se and Te) can support the given interpretation in terms of correlation with the bulk modulus of the impurity, given their associated fluctuations are much larger than the model prediction errors.

We may perform analogous investigations for the shear modulus of a pure host matrix, such as Mo, and analyze how it gets influenced by single-atom substitutional alloying, spanning over almost the entire Periodic Table, as shown in Fig. 8. We follow the same protocol and consider the variation $G'_{VRH} = G_{VRH}^{Mo(X)} - G_{VRH}^{Mo}$. In this case, the colormap reveals Si as an outlier toward hardening

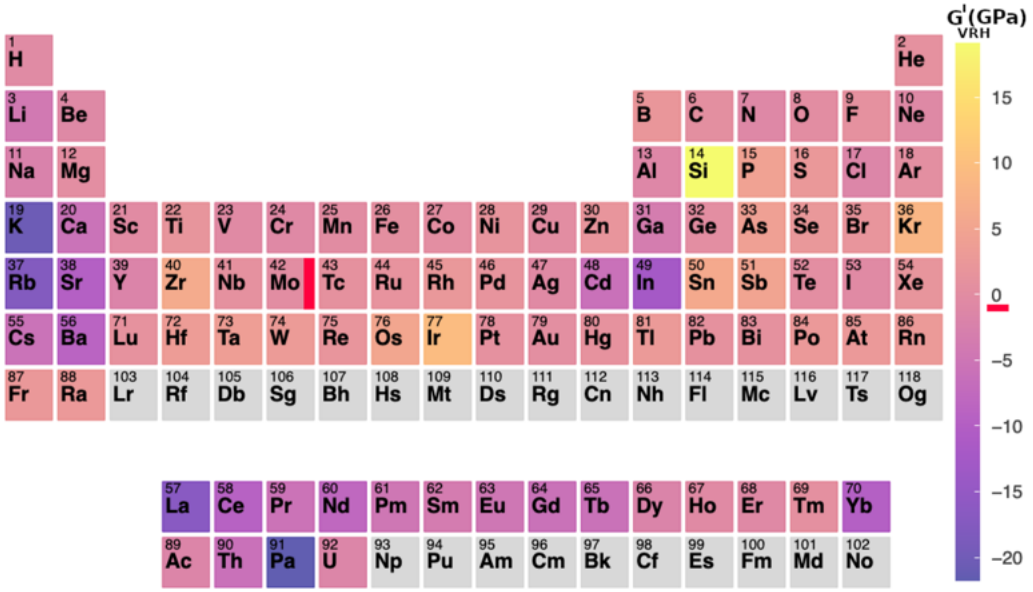


FIG. 8. Predicted shear modulus variation (G'_{VRH}) for a single-atom substitutionally defected Mo supercell with respect to the undefected one for each possible defect atomic species from the provided Periodic Table. Similar plots for Al, Au, and Ni supercells are provided in the Appendix. The red flag highlights the zero relative difference, meaning the pure Mo matrix selection.

of the Mo matrix. This is a feature that cannot be explained by a correlation in defect and defect-induced property trends since the predicted value is larger than the model prediction error on the shear modulus. We believe that the power of this investigation method is in paving the way to an efficient exploration of substitutional alloying, with the two-fold possibility of looking for comparable performances (discovery of alternatives) or outstanding ones (discovery of exceptionals). Similar to the investigation of the bulk modulus, the alkali and alkaline-earth metals such as K, Rb, Sr, and Ba are among the substitutional species providing the largest decreases in the shear modulus.

Overall, the effects and fluctuations caused by the substitutional defects on the defected host can always be highlighted, but it is not the aim of this work to find an exhaustive explanation for the existing predictions: The reasons for such trends may be due to any of the input parameters, such as the atomic number and bond lengths, or an abstract notion of the local environment, which is good enough to show reasonable correlations with existing alternative descriptors (i.e., atomic properties). The variations along the 3d, 4d, and 5d transition metals are, for most of the cases, below that of the MAE model for G_{VRH} predictions; therefore, no meaningful extrapolation is possible, but we show the plot in Fig. 23 of the Appendix.

For what concerns the formation energy, by definition of the latter, the results in Fig. 9 can be interpreted as the gain or loss in stability after the single-atom substitutional alloying has taken place. Fluorine represents a strong outlier, raising the formation energy by 0.14 eV/atom with respect to the pure Mo matrix. As shown in Fig. 10, it is evident that there is a trend that spans over the periods, and an overall interesting correlation could be found with the known trends for the atomic electronegativity along the Periodic Table, suggesting that the higher the latter for the substitutional defect in the Mo matrix is, the higher the resulting formation energy is.

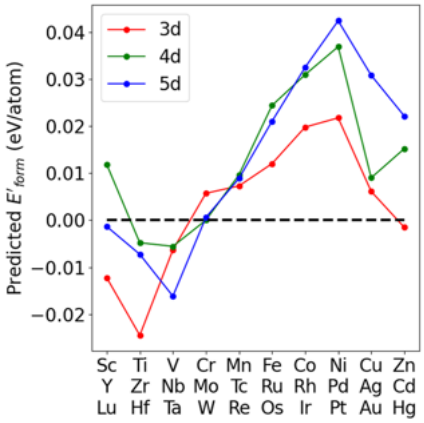


FIG. 10. Predicted formation energy variation (E'_{form}) for a single-atom substitutionally defected Mo supercell with respect to the undefected one along the 3d, 4d, and 5d series of the Periodic Table. The black dashed line highlights the pure Mo case.

C. Property prediction–DFT validation

The validation of artificial intelligence method predictions is a crucial step in order to quantify their quality. Even though the proposed graph network based method has already been validated for its predictions in bulk crystals, this work aims at testing its capabilities in the presence of single-atom substitutional defects. As explained in Sec. II B, we compare the model predictions of K_{VRH}, G_{VRH}, and E_{form} with the DFT based ones, obtained with the `THERMO_PW` driver of `QE`. Table VII shows the results for the validation on the properties

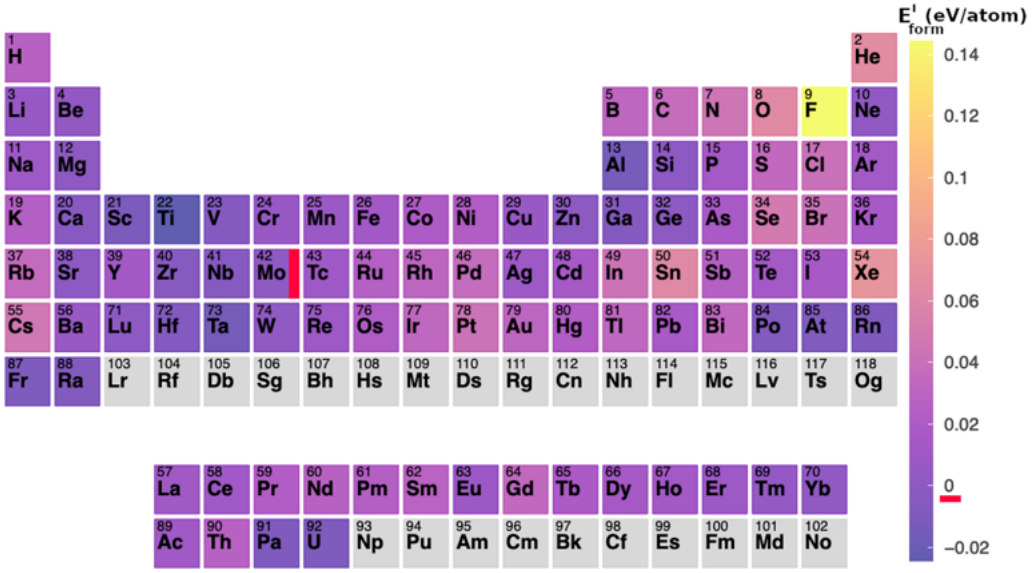


FIG. 9. Predicted formation energy variation (E'_{form}) for a single-atom substitutionally defected Mo supercell with respect to the undefected one for each possible defect atomic species from the provided Periodic Table. Similar plots for Al, Au, and Ni supercells are provided in the Appendix. The red flag highlights the zero relative difference, meaning the pure Mo matrix selection.

TABLE VII. Comparison of the DFT and MEGNet results for the three properties of interest in a small set of samples. Here, Al_B means a single B-atom substitution in an Al host 2 × 2 × 2 supercell matrix.

System	Method	K_{VRH} (GPa)	G_{VRH} (GPa)	E_{form} (eV/atom)
Al_B	MEGNet	84.456	60.405	0.050
	DFT	78.101	24.074	0.096
Al_C	MEGNet	87.165	34.762	0.113
	DFT	73.791	16.673	0.228
Al_I	MEGNet	72.697	29.796	0.056
	DFT	63.655	13.514	0.475
Al_{Ni}	MEGNet	84.683	30.437	0.009
	DFT	81.363	37.086	4.567
Al_{Zr}	MEGNet	87.881	29.158	−0.048
	DFT	79.672	30.068	1.262
MAE		8.060	15.653	1.290

in the case of an Al matrix and single-atom substitutional defects, including B, C, I, Ni, and Zr.

Comparing the MAE values of our DFT calculations with the ones of the model for non-defected systems in Table II, we find good performances of the model with respect to K_{VRH} and G_{VRH} predictions but large errors when it comes to the formation energies; the first may be regarded as a success, given that the model is predicting properties of a new class of systems and given the computational limitations; the second one is a negative signature even though the (i) defect dependent nature of the error order of magnitude opens up a deeper window of investigation on its reasons and (ii) validation set for defected systems is extremely small compared to the undefected MP dataset on which initial MAEs have been evaluated.

D. Size effects

In substitutional alloying, the defect atomic species is usually present in a dilute concentration, in the range of 0.1%–10%. For this reason, the study of how the property predictions vary with defect concentration is of interest, which we show in the saturation plots of Figs. 11–13. Our example system follows the choice of the Mo host matrix, with the H, Mn, and Rb substitutional defects present in the second and first part of the previously reported results. Interestingly, a hierarchy is conserved among the defect-induced variations for different host supercell sizes: the single Rb defect always causes the largest deviations of the studied properties from their asymptotic over-dilute level (<0.1%). Moreover, while the defect formation energy, as expected, shows a common descent to zero-level for all the defect species, the predicted structural properties seem to be sensitive on them, with the Rb-defected Mo crystal conserving a visible difference in the property value even at 0.1% concentration, both for K_{VRH} and G_{VRH}. Even though it is out of scope of the present work's

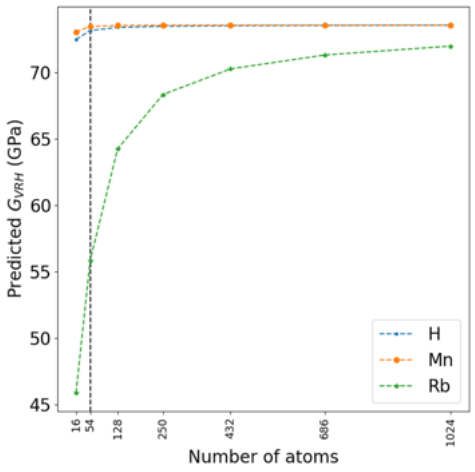


FIG. 12. Saturation plot of a Mo host matrix G_{VRH} when substitutionally defected with H, Mn, or Rb for different supercell sizes.

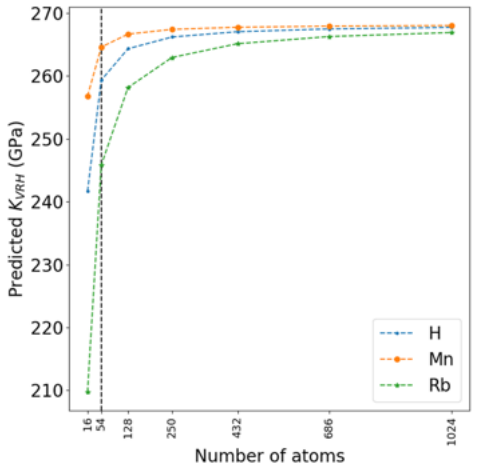


FIG. 11. Saturation plot of a Mo host matrix K_{VRH} when substitutionally defected with H, Mn, or Rb for different supercell sizes.

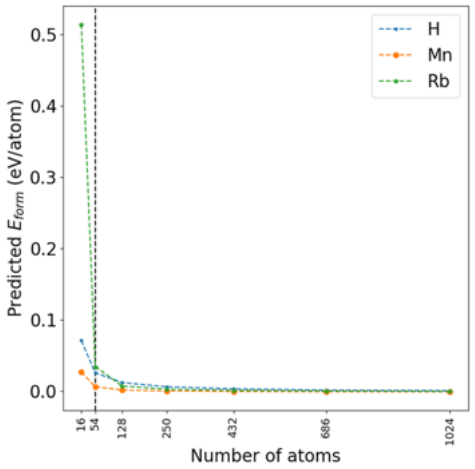


FIG. 13. Saturation plot of a Mo host matrix E_{form} when substitutionally defected with H, Mn, or Rb for different supercell sizes.

aims to validate the asymptotic-size behavior of the predictions with accurate but expensive DFT calculations, we believe these results to stand in favor of the positive model understanding of a crystalline defected system environment.

Let us continue along the previously paved path of analysis, which also looks into the effects of systematically changing the substitutional atomic species among the large variety contained in the Periodic Table, as previously shown in Fig. 6, with the only exception of selecting the smallest and largest supercell sizes from the saturation plots shown in Fig. 11, respectively: $2 \times 2 \times 2$ (1% concentration, 16 atoms cell) and $8 \times 8 \times 8$ (10% concentration, 1024 atoms cell). In Fig. 14, focusing on the first row of plots that deal with the K'_{VRH} variation, and comparing with the previously investigated supercell case of size $3 \times 3 \times 3$ (2% concentration) shown in Fig. 6,

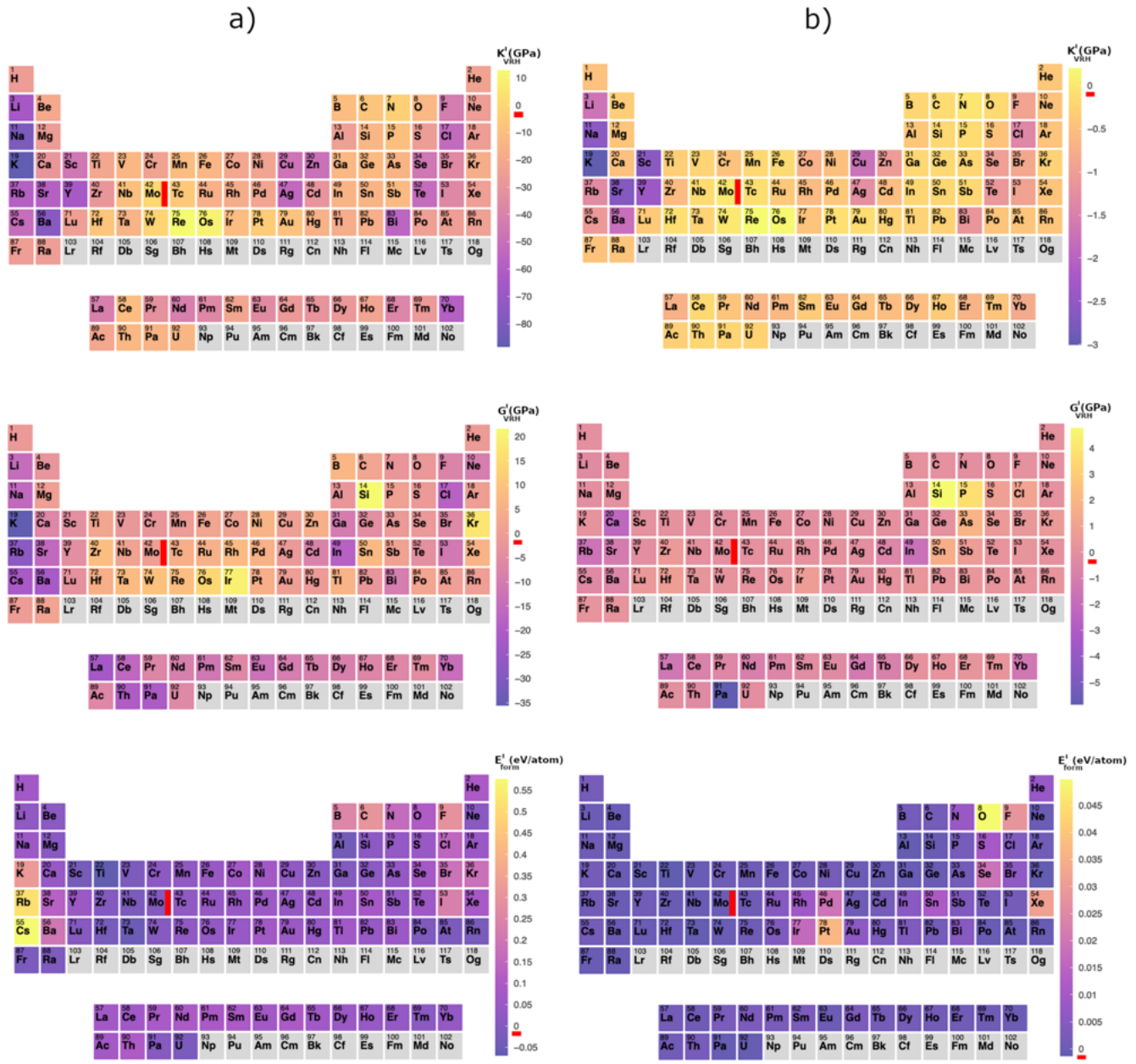


FIG. 14. Predicted variations K'_{VRH} [(a) and (d)], G'_{VRH} [(b) and (e)], and E'_{form} [(c) and (f)] for a single-atom substitutionally defected Mo supercell with respect to the undefected one for each possible defect atomic species from the provided Periodic Table. In plots (a)–(c), a $2 \times 2 \times 2$ supercell is considered, while in (d)–(f), an $8 \times 8 \times 8$ supercell is considered.

one can notice that the scale of property variations changes accordingly: smaller defect concentrations lead to smaller defect-induced effects, and vice versa, as expected. In particular, in the largest supercell case, the induced variation range is reduced to 3% of the $3 \times 3 \times 3$ system's one. However, the composition map outliers are left unchanged. The shear modulus variations, in the second row of tables, see the emergence of new outliers in the chemical neighborhoods of the previously obtained ones in $3 \times 3 \times 3$ supercells, while the formation energy variations undergo both a change in scale and a complete change in the map outliers. Following the brief assessment of prediction quality performed for each of the interesting properties in Sec. III C, we expect the formation energy variations to suffer by non-negligible fluctuations, leading to the possible need of further validation. However, the main aim of the present discussion is to underline the power of the overall approach in highlighting the path toward composition search in substitutional alloying, which can effectively drive toward specific desired emerging properties.

IV. CONCLUSIONS

In this work, we utilized a convolutional graph-neural network, based on the MEGNet architecture,[63] in order to attempt the design of novel alloys. Alloying involves substitutional and interstitial alloying at relatively low concentrations; thus, single-defect properties shall be informative on the overall designing capabilities and guidance. For this purpose, we utilized the MEGNet model, pre-trained on the MP database, and we focused on the prediction of the properties of single-atom substitutionally defected bulk crystals in the context of both (i) systematic substitution with a specific set of species in a wide variety of crystals from the entire Materials

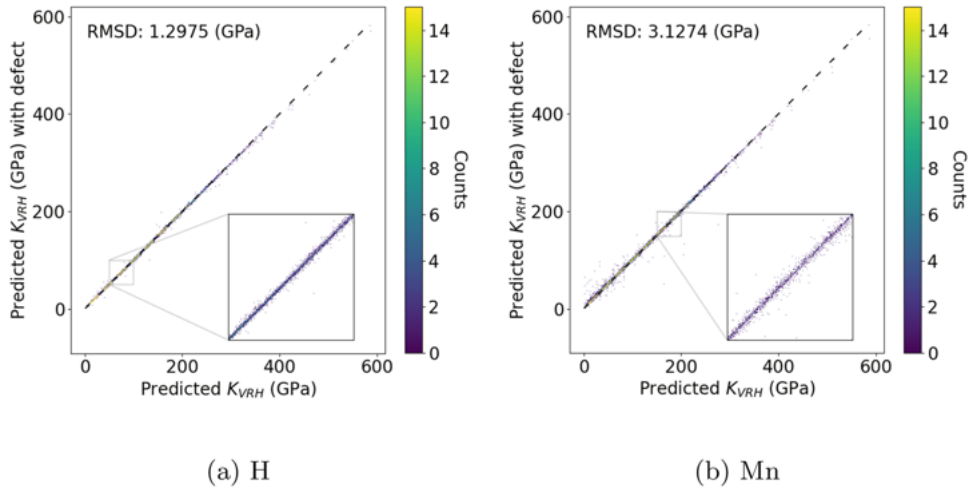


FIG. 15. Predicted bulk modulus in the (a) H- and (b) Mn-defected MP crystals with respect to its prediction in non-defected ones.

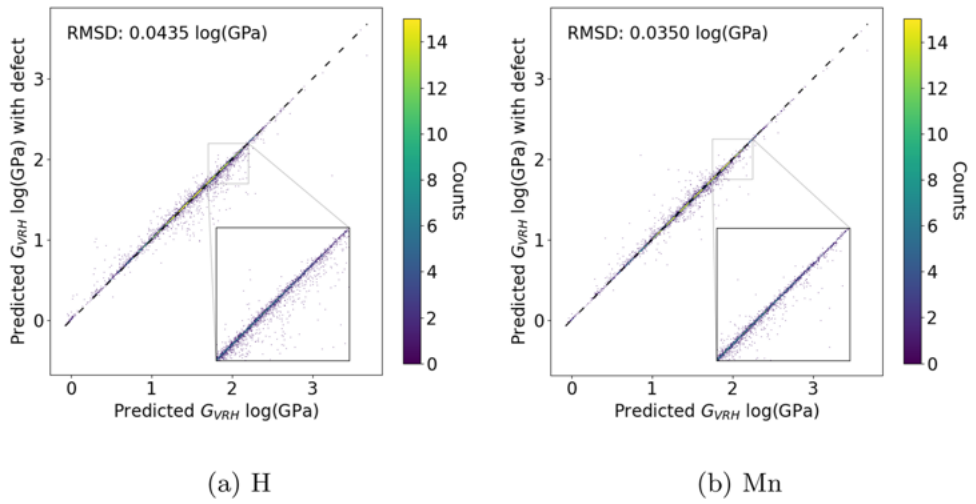


FIG. 16. Predicted shear modulus in the (a) H- and (b) Mn-defected MP crystals with respect to its prediction in non-defected ones.

Project dataset and (ii) systematic substitution with a variety of atomic species in a specific set of pure bulk crystals. We validated some of the results with our own DFT calculations. The resulting findings do not claim to be a detailed quantification of the quality and nature of atomic environment representations learned by the model as the novelty of the work lies in the testing procedure of their transferability and ultimately their usefulness in an example task related to materials discovery, a task for which such models have been created. We believe that the proposed approaches might provide novel insights into alloy design, especially if predictions include extended lattice defects such as dislocations and/or grain boundaries, as well as into the importance of defect inclusion into database design, on which present or future state-of-the-art GNN architectures can be trained on.

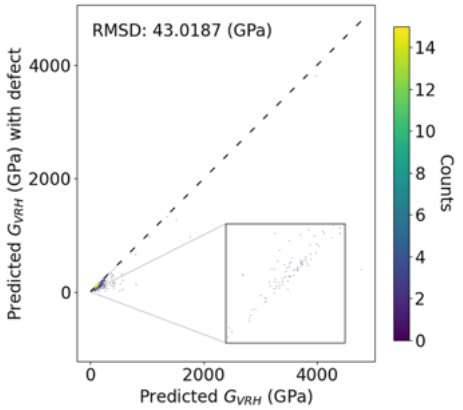


FIG. 18. Predicted formation energy in the Rb-defected MP crystals with respect to its prediction in non-defected ones. The plot has the aim to display the need of a log–log scale to appreciate the deviations.

ACKNOWLEDGMENTS

We would like to thank J. Llorca for insights and fruitful discussions and suggestions. This research was funded in part by the European Union Horizon 2020 Research and Innovation Program, under Grant Agreement No. 857470, and by the European Regional Development Fund via the Foundation for Polish Science International Research Agenda PLUS Program, Grant No. MAB PLUS/2018/8.

AUTHOR DECLARATIONS

Conflict of Interest

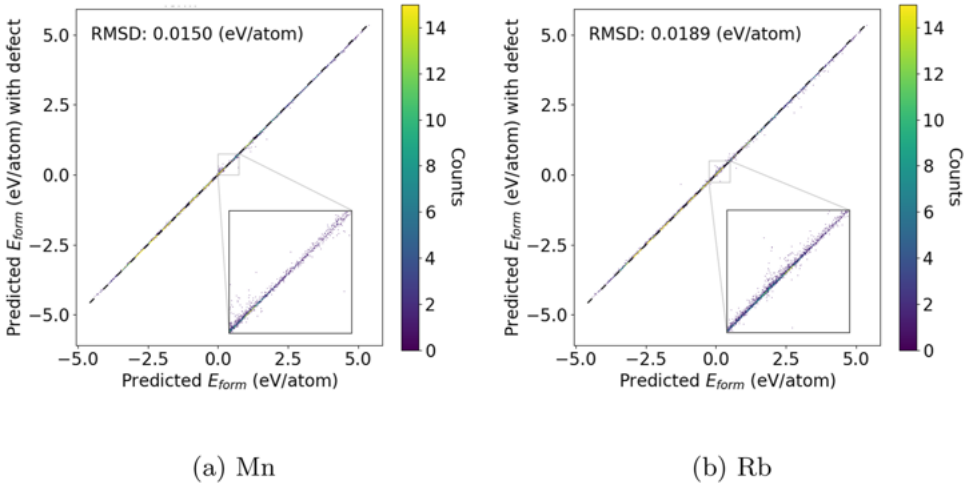
The authors have no conflicts to disclose.

Author Contributions

Dario Massa: Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal); Visualization (equal); Writing – original draft (lead); Writing – review & editing (equal). **Daniel Cieśliński**: Data curation (equal); Formal analysis (equal); Methodology (equal); Software (equal). **Amirhossein Naghdi**: Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Writing – review & editing (equal). **Stefanos Papanikolaou**: Conceptualization (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable



(a) Mn (b) Rb

FIG. 17. Predicted formation energy in the (a) Mn- and (b) Rb-defected MP crystals with respect to its prediction in non-defected ones.

request. The ML model and database are openly available in GitHub, and the `suballoy` repository is available at https://github.com/danielcieslinski/suballoy.[71]

APPENDIX: ADDITIONAL RESULTS

1. Further results for property prediction–a defect in a wide range of matrices

As mentioned in Sec. III B, we first focus on the prediction of defected system properties when the systematic substitutional defective process is applied with the same atom on the Materials Project crystal dataset. There, we show the RMSD for the predictions of bulk modulus, shear modulus, and defect formation energy, respectively, in Fig. 2 and Table IV, Fig. 3 and Table V, and Fig. 4 and Table VI. However, the figures report only the defects leading to the corresponding largest RMSD values. Here, for completeness, we report the missing ones in Figs. 15–17.

For the shear modulus predictions, the reader might have noticed the choice of a different scale for the plots. Apart from noticing a common use in the literature of the log–log scale for this quantity and adopting the same for the sake of comparison, we also decided to understand the reason behind this choice by plotting in a normal scale, as shown in Fig. 18; the wide scale over which a few predictions are spreading does not allow appreciation of the distribution of the data.

2. Further results for property prediction–a wide range of defects in the same matrix

Following a similar selection criterion for the results to which the main body of the article is dedicated, for the second part of our investigation, in which we consider the systematic change in the single-atom substitutional defect species in the same crystal structure, we decided to show the results for the molybdenum host matrix, which displays, overall, the largest variations in the interesting quantities. In Figs. 19 and 20, we report the collection of defect-induced variations in the bulk modulus, shear modulus, and formation energy for the all host-matrices investigated here, namely,

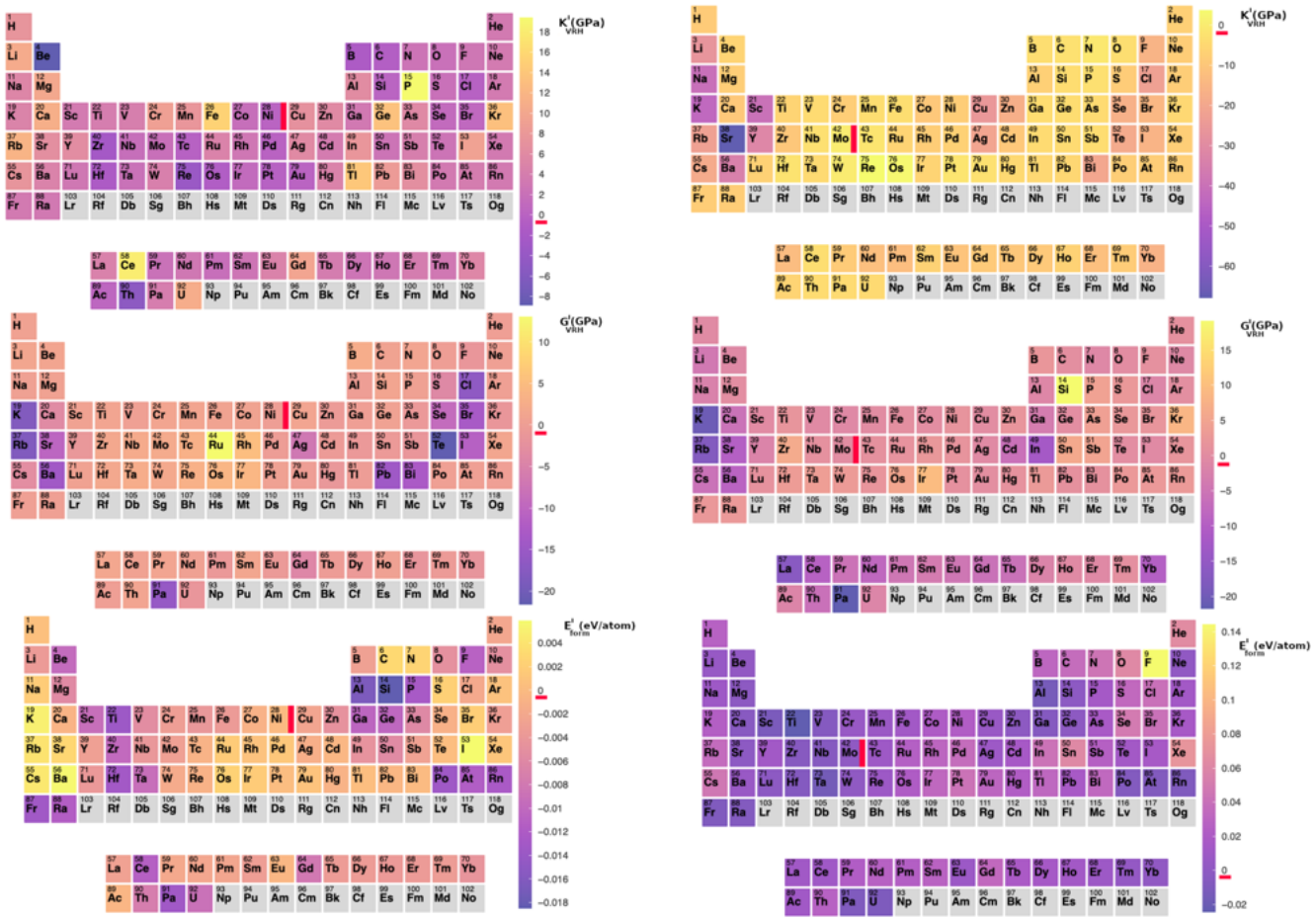


FIG. 19. Comparison between the Periodic Table plots of the predicted bulk modulus, shear modulus, and formation energy for a single-atom substitutionally defected Ni [(a)–(c)] and Mo [(d)–(f)] supercell with respect to the undefected one for each possible atomic species from the provided Periodic Table.

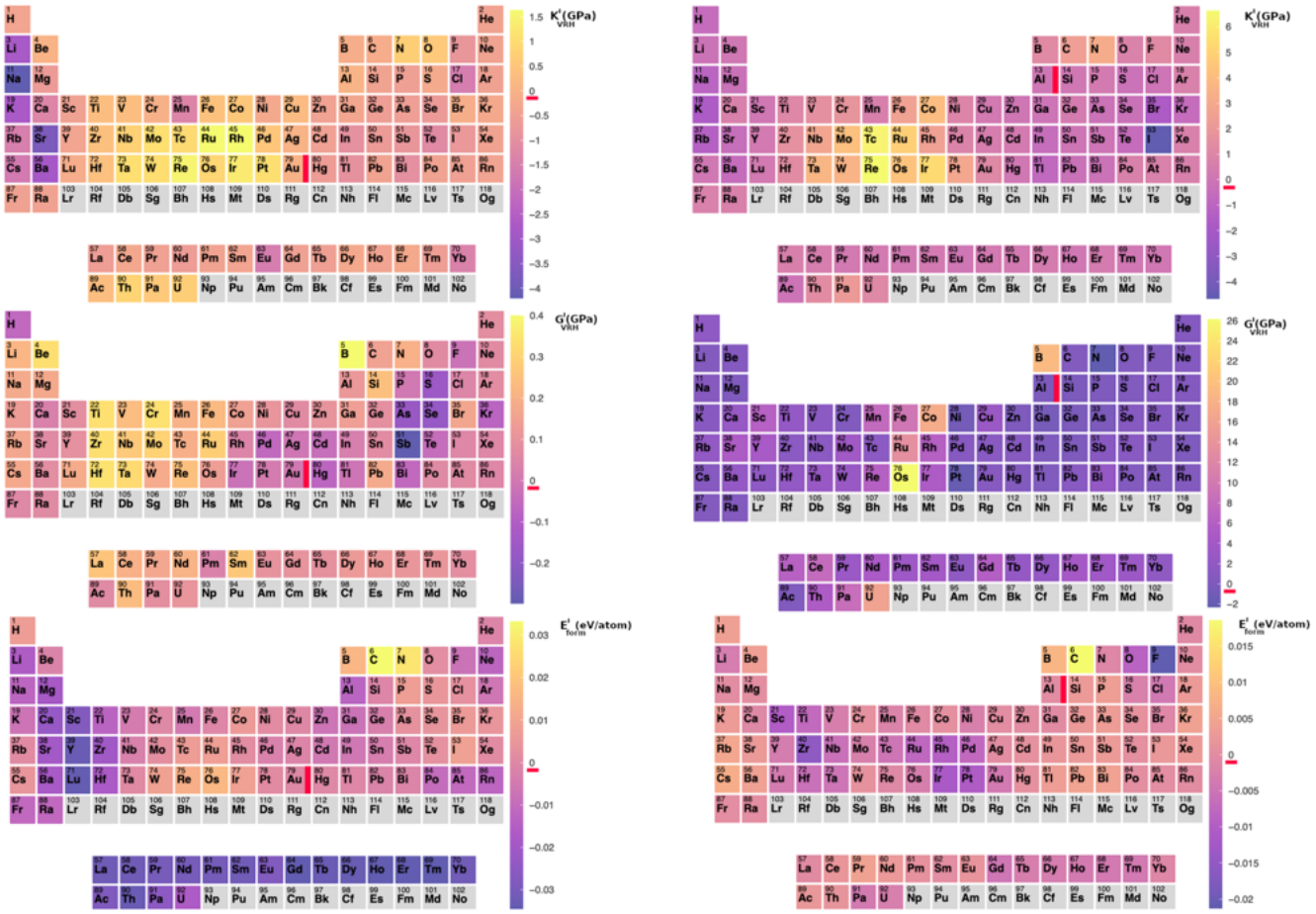
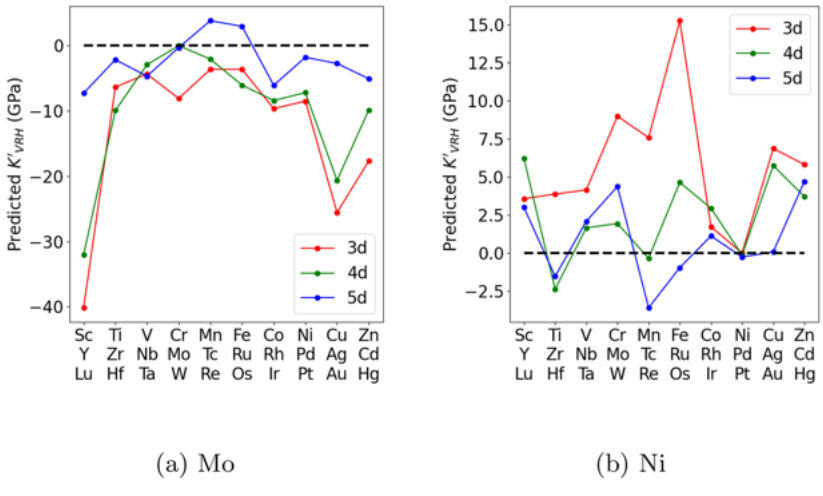


FIG. 20. Comparison between the Periodic Table plots of the predicted bulk modulus, shear modulus, and formation energy for a single-atom substitutionally defected Au [(a)–(c)] and Al [(d)–(f)] supercell with respect to the undefected one for each possible atomic species from the provided Periodic Table.



(a) Mo (b) Ni

FIG. 21. Predicted bulk modulus variation (K'_{VRH}) for a single-atom substitutionally defected Mo (a) and Ni (b) supercell with respect to the undefected one along the 3d, 4d, and 5d series of the Periodic Table. The black dashed line highlights the pure host matrix case.

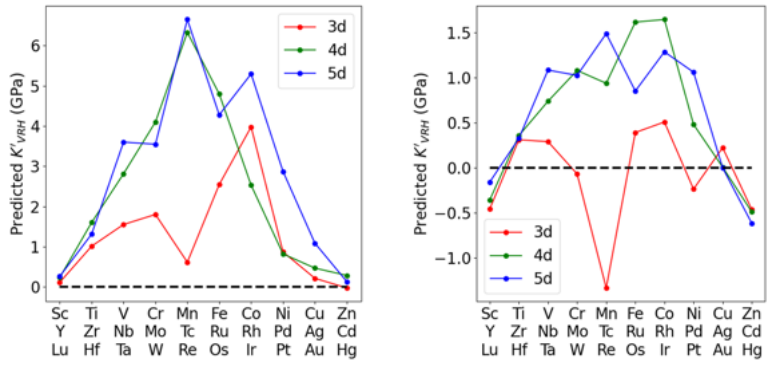


FIG. 22. Predicted bulk modulus variation (K'_{VRH}) for a single-atom substitutionally defected Al (a) and Au (b) supercell with respect to the undefected one along the 3d, 4d, and 5d series of the Periodic Table. The black dashed line highlights the pure host matrix case.

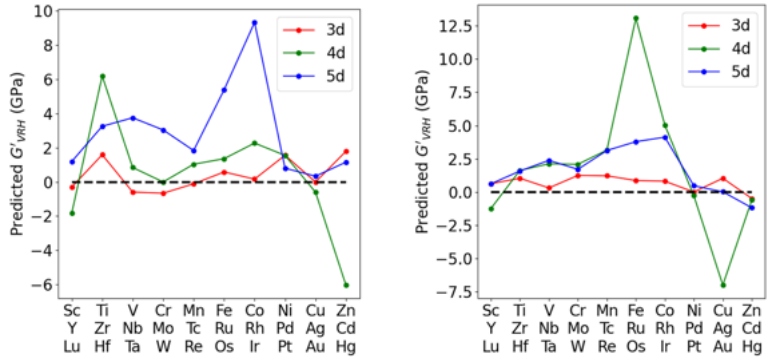


FIG. 23. Predicted shear modulus variation (G'_{VRH}) for a single-atom substitutionally defected Mo (a) and Ni (b) supercell with respect to the undefected one along the 3d, 4d, and 5d series of the Periodic Table. The black dashed line highlights the pure host matrix case.

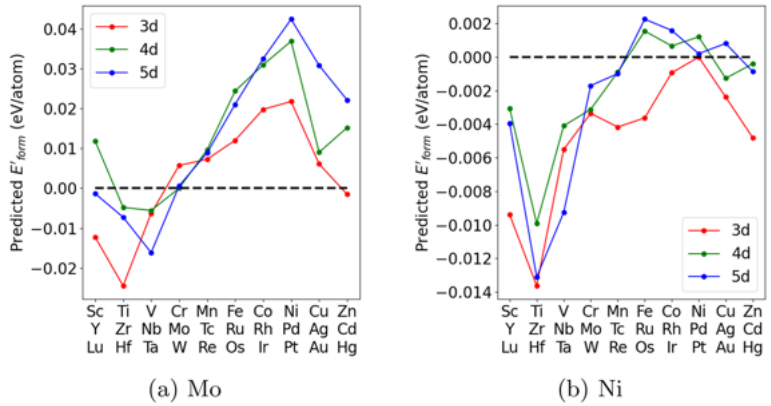


FIG. 24. Predicted formation energy variation (E'_{form}) for a single-atom substitutionally defected Mo (a) and Ni (b) supercell with respect to the undefected one along the 3d, 4d, and 5d series of the Periodic Table. The black dashed line highlights the pure host matrix case.

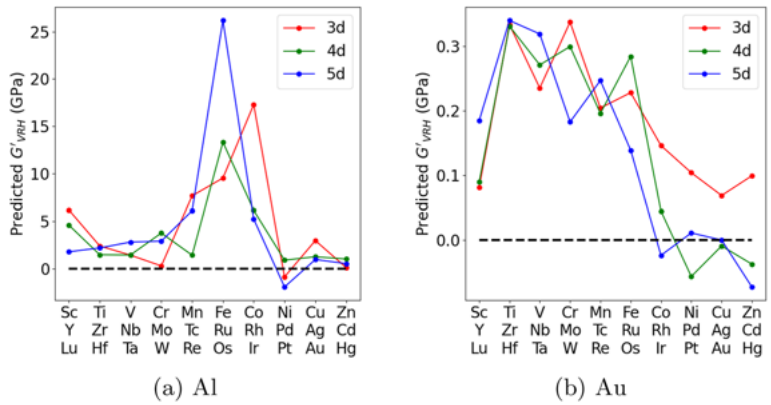


FIG. 25. Predicted shear modulus variation (G'_{VRH}) for a single-atom substitutionally defected Al (a) and Au (b) supercell with respect to the undefected one along the 3d, 4d, and 5d series of the Periodic Table. The black dashed line highlights the pure host matrix case.

Ni and Mo, and Au and Al. A composite visualization of all the variational periodic tables allows for an overall comparison and helps in appreciating that interesting effects not shown in the previous matrices are not missing:

- the shear modulus variation scale of the Ni matrix upon single-atom substitution is comparable to the one of Mo, and they also share alkaline and alkaline earth metals in the lower bound variations;
- similar to the previous point, Au and Al share a similar variation scale and outlier map for the bulk modulus and, in particular, for the formation energy.

Even though the variations in the interesting properties upon substitutional-defecting of the host matrices are dominant for the Mo matrix, here, we want to report, as previously done, all the variations and draw a brief comparison between them in a combined visualization. While the trends in the bulk modulus variations of Mo and Ni matrices are noticeably different, see Figs. 21(a) and 21(b), in Figs. 22(a) and 22(b), Al and Au show pretty similar trends and hierarchies on different scales: the central 4d and 5d elements tend to maximize the variations, with a minimization happening at extrema for both, as well as for a Mn substitution. In particular, it is interesting to notice that in the Al matrix, nearly all the Periodic Table elements considered lead to a positive bulk modulus variation. A similarity between variations in Mo and Ni holds for the shear modulus variations, see Figs. 23(a) and 23(b), which share comparable scales and a similar exchanged role of 5d (in Mo) and 4d (in Ni) substitutions, and for the formation energy variations, see Figs. 24(a)

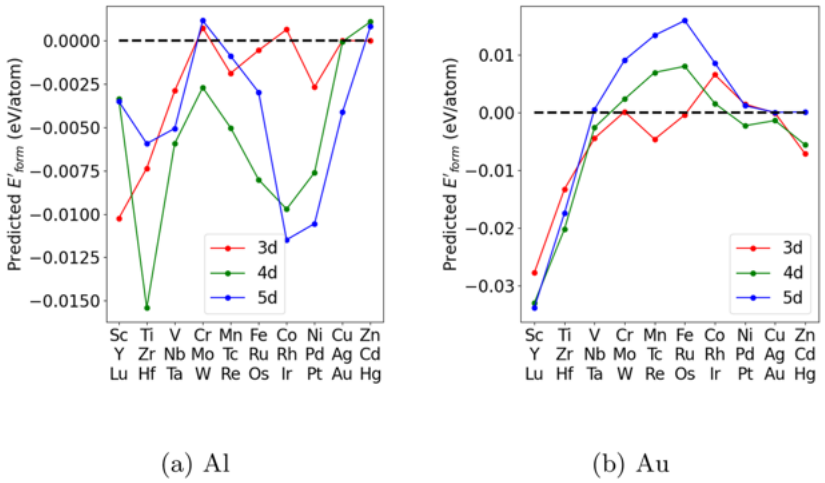


FIG. 26. Predicted formation energy variation (E'_{form}) for a single-atom substitutionally defected Al (a) and Au (b) supercell with respect to the undefected one along the 3d, 4d, and 5d series of the Periodic Table. The black dashed line highlights the pure host matrix case.

and 24(b), showing remarkably similar trends over scales differing by an order of magnitude. The shear modulus and formation energy variations in Al and Au matrices instead, shown in Figs. 25(a), 25(b), 26(a), and 26(b), do not share the trends.

REFERENCES

[1] R. S. Michalski *et al.*, *Machine Learning an Artificial Intelligence Approach* (Springer Science and Business Media, 2013).
[2] Y. LeCun *et al.*, "Deep learning," Nature **521**, 436 (2015).
[3] S. Curtarolo *et al.*, "The high-throughput highway to computational materials design," Nat. Mater. **12**, 191 (2013).
[4] S. Ramakrishna *et al.*, "Materials informatics," J. Intell. Manuf. **30**, 2307–2326 (2019).
[5] R. Ramprasad *et al.*, "Machine learning in materials informatics: Recent applications and prospects," npj Comput. Mater. **3**, 54 (2017).
[6] K. Takahashi and Y. Tanaka, "Materials informatics: A journey towards material design and synthesis," Dalton Trans. **45**, 10497 (2016).
[7] L. Ward and C. Wolverton, "Atomistic calculations and materials informatics: A review," Curr. Opin. Solid State Mater. Sci. **21**(3), 167 (2017).
[8] K. Rajan, "Materials informatics," Mater. Today **8**(10), 38–45 (2005).
[9] E. Saal *et al.*, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD)," JOM **65**, 1501 (2013).
[10] A. Jain *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).
[11] S. Curtarolo *et al.*, "AFLOWLIB.ORG: A distributed materials properties repository from high-throughput *ab initio* calculations," Comput. Mater. Sci. **58**, 227 (2012).
[12] J. Hachmann *et al.*, "The Harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid," J. Phys. Chem. Lett. **2**, 2241 (2011).
[13] NOMAD, https://nomad-coe.eu.
[14] A. Seko *et al.*, "Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization," Phys. Rev. Lett. **115**, 205901 (2015).
[15] D. Xue *et al.*, "Accelerated search for materials with targeted properties by adaptive design," Nat. Commun. **7**, 11241 (2016).
[16] O. Isayev *et al.*, "Universal fragment descriptors for predicting properties of inorganic crystals," Nat. Commun. **8**, 15679 (2017).
[17] M. Rupp *et al.*, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 058301 (2012).
[18] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys. **134**, 074106 (2011).
[19] F. Pietrucci and W. Andreoni, "Graph theory meets *ab initio* molecular dynamics: Atomic structures and transformations at the nanoscale," Phys. Rev. Lett. **107**, 085504 (2011).
[20] S. De *et al.*, "Comparing molecules and solids across structural and alchemical space," Phys. Chem. Chem. Phys. **18**, 13754–13769 (2016).
[21] A. Sadeghi *et al.*, "Metrics for measuring distances in configuration spaces," J. Chem. Phys. **139**, 184118 (2013).
[22] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," AI Open **1**, 57–81 (2020).
[23] Z. Wu *et al.*, "A comprehensive survey on graph neural networks," IEEE Trans. Neural Networks Learn. Syst. **32**, 4–24 (2021).
[24] P. B. Jørgensen *et al.*, "Neural message passing with edge updates for predicting properties of molecules and materials," in *32nd Conference on Neural Information Processing Systems*, Montreal, Canada (NIPS, 2018).
[25] K. Schütt, P. J. Kindermans, H. E. Sauceda Felix *et al.*, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions," Advances in Neural Information Processing Systems (NeurIPS Proceedings 30, (2017).
[26] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre *et al.*, "Convolutional networks on graphs for learning molecular fingerprints," Advances in Neural Information Processing Systems (NeurIPS Proceedings 28, 2015).
[27] Z. Wu, B. Ramsundar, E. N. Feinberg *et al.*, "MoleculeNet: A benchmark for molecular machine learning," Chem. Sci. **9**(2), 513–530 (2018).
[28] S. Kearnes *et al.*, "Molecular graph convolutions: Moving beyond fingerprints," J. Comput.-Aided Mol. Des. **30**, 595–608 (2016).
[29] C. W. Coley *et al.*, "Convolutional embedding of attributed molecular graphs for physical property prediction," J. Chem. Inf. Model. **57**, 1757–1772 (2017).
[30] S. Back *et al.*, "Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts," J. Phys. Chem. Lett. **10**, 4401–4408 (2019).
[31] A. Palizhati *et al.*, "Toward predicting intermetallics surface properties with high-throughput DFT and convolutional neural networks," J. Chem. Inf. Model. **59**, 4742–4749 (2019).
[32] G. H. Gu *et al.*, "Practical deep-learning representation for fast heterogeneous catalyst screening," J. Phys. Chem. Lett. **11**, 3185–3191 (2020).
[33] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," Phys. Rev. Lett. **120**, 145301 (2018).
[34] T. Xie and J. C. Grossman, "Hierarchical visualization of materials space with graph convolutional neural networks," J. Chem. Phys. **149**, 174111 (2018).
[35] C. Chen *et al.*, "Graph networks as a universal machine learning framework for molecules and crystals," Chem. Mater. **31**, 3564–3572 (2019).
[36] A. Dunn *et al.*, "Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm," npj Comput. Mater. **6**(1), 138 (2020).
[37] S. Y. Louis *et al.*, "Graph convolutional neural networks with global attention for improved materials property prediction," Phys. Chem. Chem. Phys. **22**, 18141–18148 (2020).
[38] C. W. Park and C. Wolverton, "Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery," Phys. Rev. Mater. **4**, 063801 (2020).
[39] M. Karamad *et al.*, "Orbital graph convolutional neural network for material property prediction," Phys. Rev. Mater. **4**, 093801 (2020).
[40] P. Reiser *et al.*, "Graph neural networks for materials science and chemistry," Commun. Mater. **3**, 93 (2022).
[41] W. Gong and Q. Yan, "Graph-based deep learning frameworks for molecules and solid-state materials," Comput. Mater. Sci. **195**, 110332 (2021).
[42] V. Fung *et al.*, "Benchmarking graph neural networks for materials chemistry," npj Comput. Mater. **7**, 84 (2021).
[43] Y. Liu *et al.*, "High-throughput experiments facilitate materials innovation: A review," Sci. China: Technol. Sci. **62**, 521–545 (2019).
[44] A. Mannodi-Kanakkithodi *et al.*, "Universal machine learning framework for defect predictions in zinc blende semiconductors," Patterns **3**, 100450 (2022).
[45] X. Zhai *et al.*, "Predicting the formation of fractionally doped perovskite oxides by a function-confined machine learning method," Commun. Mater. **3**, 42 (2022).
[46] P. V. Balachandran *et al.*, "Predictions of new ABO_3 perovskite compounds by combining machine learning and density functional theory," Phys. Rev. Mater. **2**(4), 043802 (2018).
[47] W. Ye *et al.*, "Deep neural networks for accurate predictions of crystal stability," Nat. Commun. **9**, 3800 (2018).
[48] V. Sharma *et al.*, "Machine learning substitutional defect formation energies in ABO_3 perovskites," J. Appl. Phys. **128**, 034902 (2020).
[49] M. Klug *et al.*, "Tailoring metal halide perovskites through metal substitution: Influence on photovoltaic and material properties," Energy Environ. Sci. **10**, 236 (2017).
[50] M. Sampson *et al.*, "Transition metal-substituted lead halide perovskite absorbers," J. Mater. Chem. A **5**, 3578 (2017).
[51] Z. Guan *et al.*, "Compositional engineering of multinary Cu–In–Zn-based semiconductor nanocrystals for efficient and solution-processed red-emitting quantum-dot light-emitting diodes," Org. Electron. **74**, 46 (2019).
[52] C. Z. Ning *et al.*, "Bandgap engineering in semiconductor alloy nanomaterials with widely tunable compositions," Nat. Rev. Mater. **2**, 17070 (2017).

[53]F. Oba and Y. Kumagai, "Design and exploration of semiconductors from first principles: A review of recent advances," Appl. Phys. Express **11**, 060101 (2018).
[54]A. Deml *et al.*, "Intrinsic material properties dictating oxygen vacancy formation energetics in metal oxides," J. Phys. Chem. Lett. **6**, 1948 (2015).
[55]A. Deml *et al.*, "Oxide enthalpy of formation and band gap energy as accurate descriptors of oxygen vacancy formation energetics," Energy Environ. Sci. **7**, 1996 (2014).
[56]Z. Wan *et al.*, "Data-driven machine learning model for the prediction of oxygen vacancy formation energy of metal oxide materials," Phys. Chem. Chem. Phys. **23**, 15675 (2021).
[57]J. Varley *et al.*, "Descriptor-based approach for the prediction of cation vacancy formation energies and transition levels," J. Phys. Chem. Lett. **8**, 5059 (2017).
[58]A. Mannodi-Kanakkithodi *et al.*, "Machine-learned impurity level prediction for semiconductors: The example of Cd-based chalcogenides," npj Comput. Mater. **6**(1), 39 (2020).
[59]K. Choudhary and B. DeCost, "Atomistic line graph neural network for improved materials property predictions," npj Comput. Mater. **7**(1), 185 (2021).
[60]S. S. Omee, S. Y. Louis, N. Fu *et al.*, "Scalable deeper graph neural networks for high-performance materials property prediction," Patterns **3**, 100491 (2022).
[61]S. P. Ong *et al.*, "Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis," Comput. Mater. Sci. **68**, 314 (2013).
[62]D. Chung and W. Buessem, "The Voigt-Reuss-Hill approximation and elastic moduli of polycrystalline MgO, CaF_2, β-ZnS, ZnSe, and CdTe," J. Appl. Phys. **38**(6), 2535–2540 (1967).
[63]C. Chen *et al.* (2022). "The megnet model on GitHub," GitHub. https://github.com/materialsvirtuallab/megnet/blob/master/megnet/models/megnet.py
[64]P. Giannozzi *et al.*, J. Phys.: Condens. Matter **21**, 395502 (2009).
[65]P. Giannozzi *et al.*, J. Phys.: Condens. Matter **29**, 465901 (2017).
[66]P. Giannozzi *et al.*, J. Chem. Phys. **152**, 154105 (2020).
[67]A. Dal Corso (2023). "Welcome to Thermo_pw," GitHub. https://dalcorso.github.io/thermo_pw/
[68]J. P. Perdew *et al.*, Phys. Rev. Lett. **77**, 3865 (1996).
[69]M. Methfessel and A. T. Paxton, Phys. Rev. B **40**, 3616 (1989).
[70]E. Kaxiras, *Atomic and Electronic Structure of Solids* (Cambridge University Press, 2003).
[71]D. Cieslinski (2022), "'Suballoy' repository containing the ML model and python-based codes for the reproducibility of the proposed results," GitHub. https://github.com/danielcieslinski/suballoy

2.4 Manuscript IV

"Hydrogen Diffusion in Magnesium Using Machine Learning Potentials: a comparative study"

Andrea Angeletti, Luca Leoni, Dario Massa, Luca Pasquini, Stefanos Papanikolaou, and Cesare Franchini

Commentary

The work tackles the computational modeling complexities of hydrogen diffusion in bulk crystals, for the case of magnesium, by using ML-accelerated molecular dynamics (ML-MD) simulations. We present a comparison among multiple models for MLIPs development, each characterized by different training approaches — VASP-MLFF (*on-the-fly* or *active*-learning), MACE and CHGNet (universal pre-trained and fine-tuned models). The active-learning procedure, sampling configurations from the diffusive MD trajectory at various temperatures, tailors datasets that include critical representative states for the system and its dynamics. The methodology employs fine-tuning of the UIPs on the VASP-MLFF on-the-fly database, adapting them specifically for systems with diffusing defects. The results show how the fine-tuned UIPs achieve DFT-level prediction of energies and forces on a validation set of AIMD configurations, and replicate with great accuracy the experimental temperature dependence of hydrogen diffusion coefficients, as well as the associated transition energy barrier. The pre-trained counterparts instead fail in capturing the correct order of magnitude of the dynamical properties of interest, underscoring the limitations of general-purpose UIPs on defected materials. This is due to a lack of representative configurations in the original training databases, such as high-temperature or transition-state data, which reduces the prediction accuracy. Such systematic approach for MLIPs development in diffusive dynamics establishes a framework to extend their applicability without compromising ab-initio precision.

In this study, the PhD candidate contributed to: conceptualization of the work; preliminary ab-initio MD simulations in VASP; MD simulations at the different presented temperatures and ensembles with the CHGNet potential (pre-trained and finetuned) for multiple replicas; CHGNet finetuning on the DFT database; extraction of the dynamical properties of interest; NEB calculations with CHGNet; writing and figures.

Hydrogen Diffusion in Magnesium Using Machine Learning Potentials: a comparative study

Andrea Angeletti[1,+], Luca Leoni[2,*,+], Dario Massa[3], Luca Pasquini[2], Stefanos Papanikolaou[3], and Cesare Franchini[1,2]

[1]University of Vienna, Vienna Doctoral School in Physics, Boltzmanngasse 5, 1090, Vienna, Austria, Europe
[2]Department of Physics and Astronomy 'Augusto Righi', Alma Mater Studiorum - Università di Bologna, Viale Berti Pichat 6/2, 40127, Bologna, Italy, Europe
[3]NOMATEN Centre of Excellence, National Center for Nuclear Research, 05-400, Otwock, Poland, Europe
[*]Corresponding author: luca.leoni12@unibo.it
[+]These authors contributed equally to this work

ABSTRACT

Understanding and accurately predicting hydrogen diffusion in materials is challenging due to the complex interactions between hydrogen defects and the crystal lattice. These interactions span large length and time scales, making them difficult to address with standard ab initio techniques. This work addresses this challenge by employing accelerated machine learning (ML) molecular dynamics simulations through active learning. We conduct a comparative study of different ML-based interatomic potential schemes, including VASP, MACE, and CHGNet, utilizing various training strategies such as on-the-fly learning, pre-trained universal models, and fine-tuning. By considering different temperatures and concentration regimes, we obtain hydrogen diffusion coefficients and activation energy values which align remarkably well with experimental results, underlining the efficacy and accuracy of ML-assisted methodologies in the context of diffusive dynamics. Particularly, our procedure significantly reduces the computational effort associated with traditional transition state calculations or ad-hoc designed interatomic potentials. The results highlight the limitations of pre-trained universal solutions for defective materials and how they can be improved by fine-tuning. Specifically, fine-tuning the models on a database produced during on-the-fly training of VASP ML force-field allows the retrieving of DFT-level accuracy at a fraction of the computational cost.

1 Introduction

The global imperative for sustainable and green energy solutions has intensified the search for efficient hydrogen storage materials. With the highest energy density among fuels[1], hydrogen represents a promising alternative to fossil sources that can be produced with zero CO_2 emissions powered by surplus renewable energy[2], through methods such as electrolysis[3–5]. The main barrier preventing a future economy based on hydrogen energy is the cost of production, and the absence of a green, safe and efficient way to store and transport it. Solid state hydrogen storage technologies are the most studied in this regard, being the safest, offering higher volumetric densities[6] than cryogenic or high-pressure gaseous alternatives[7–9]. Despite these advantages, the technology remains in its early stages and the search for materials allowing large-scale applications, like in the automotive industry[10], remains open[11,12].

Metal hydrides currently represent promising efficient and economical solutions[13]. Particularly, Magnesium stands out for its excellent hydrogen storage capacity[14], environmental friendliness, and natural abundance, displaying theoretical storage capacities as high as 7.6% wt[15]. However, the slow kinetics of hydrogen in magnesium-based compounds still pose a limit to possible applications. Therefore, understanding and optimizing hydrogen diffusion pathways through theoretical modeling[16–18] and experimental studies[19–21] is crucial in order to improve performances of future Mg-based hydrogen storage materials. Despite numerous efforts over the past decade, modeling hydrogen dynamics in solid-state compounds remains challenging[22–25]. The low hydrogen diffusivity in magnesium requires prolonged simulation times, in the order of nanoseconds, for accurate studies using ab-initio Molecular Dynamics (MD). Consequently, ab-initio transition state calculations, such as the nudged elastic band (NEB) method, have emerged as the most effective approaches to reproduce and interpret the experimental data documented in the literature to date[17]. However, this technique is cumbersome and often impracticable for systems with high defect concentrations, or complex potential energy landscapes: manually describing all possible paths required by NEB may be very challenging and virtually impossible[17].

Recently, Machine Learning accelerated MD (MLMD) has revolutionized the world of MD by making accurate simulations of large systems accessible over long time scales[26]. The application of such approach to study hydrogen defective systems is of high interest[27], since the prediction of dynamical properties would highly expand the limited landscape offered by today's

transition state computations. MLMD does allow the study of multi-component systems[28] and can efficiently account for interaction between defects[29]. However, developing accurate interatomic potentials, especially for hydrogen-defective materials, is notoriously challenging[30,31]. Still, the field is rapidly growing and several new approaches were proposed to explore new and complex phase spaces. On one side, various pre-trained universal solutions[32,33] start to be available, aiming to offer a convenient and versatile way of tackling the problem. However, as discussed in this work, while their training datasets are rich in chemical compositional space, the limited configurational sampling can significantly compromise their accuracy on previously unseen defected, metastable and transition states, leading to ungranted generalization capabilities. On the other side, active-learning approaches based on Bayesian force fields are showing a large versatility thanks to the construction of on-the-fly databases[34,35]. The error-oriented sampling of configurations allows these models to easily collect high-quality data widely spanning the configurational space, thus making them highly accurate despite their architecture, constrained compared to neural networks. The current study aims at illustrating a systematic procedure for ML-potentials applications in diffusive dynamics conditions, which can be used to enhance the study of different embedded defects, without departing from ab-initio accuracy. This procedure specifically consists in improving ML-based pre-trained models' performance via actively learned configurations, generated by on-the-fly training of the Vienna Ab Initio Simulation Package Bayesian ML Force-Field (VASP-MLFF)[34,35]. We consider four different concentrations ($MgH_{0.03125}$, $MgH_{0.046875}$, $MgH_{0.0625}$ and $MgH_{0.078125}$) and compute the hydrogen diffusion coefficient at three different temperatures (300 K, 480 K, and 673 K), employing a proper methodology which ensures accurate analysis of unbiased dynamical properties. Two different Universal Interatomic Potentials (UIPs), CHGNet[32] and MACE[33], were considered. Dynamical properties were computed for VASP-MLFF alongside the pre-trained and fine-tuned versions of the UIPs. The comparison between the different results and experimental data showed excellent agreement, both with VASP-MLFF and fine-tuned potentials, while the pre-trained versions fail to reach a satisfactory accuracy. Interestingly, the fine-tuned potentials outperform VASP-MLFF by correctly predicting the temperature dependence of the diffusion coefficient.

2 Material and Methods

2.1 MLFF-MD

The Density Functional Theory (DFT), MLFF-MD and climbing-image-NEB[36] (ciNEB) calculations were performed using VASP version 6.4.2 and version 6.4.3[34,35,37,38], respectively. We utilized a $4 \times 4 \times 4$ supercell, shown in Fig. 1, comprising of 128 Mg atoms in hcp crystal symmetry, with 8 H atoms ($MgH_{0.0625}$) randomly distributed in the lattice. We performed non-spin-polarized calculations, at the Perdew-Burke-Ernzerhof (PBE) functional level of theory[39], using an energy cutoff of 600 eV and a $4 \times 4 \times 4$ k-points grid, with convergence threshold of 0.1 meV, employing a Gaussian smearing with a sigma value of 0.05 eV. To ensure consistency and avoid any discrepancies with our database, we adopted the same convergence parameters as those used in the Materials Project Database. For VASP-MLFF we used the default parameters. The temperature was incrementally raised from 0 to 700 K over a 0.2 ns interval with VASP on-the-fly MLFF-MD. In this setup, derived configurations—including structure energy, forces acting on each atom, atomic coordinates, stress tensor, and lattice parameters—were used to train the interatomic potential. Whenever the Bayesian error surpassed the set fixed threshold, VASP reverted to DFT to generate a new

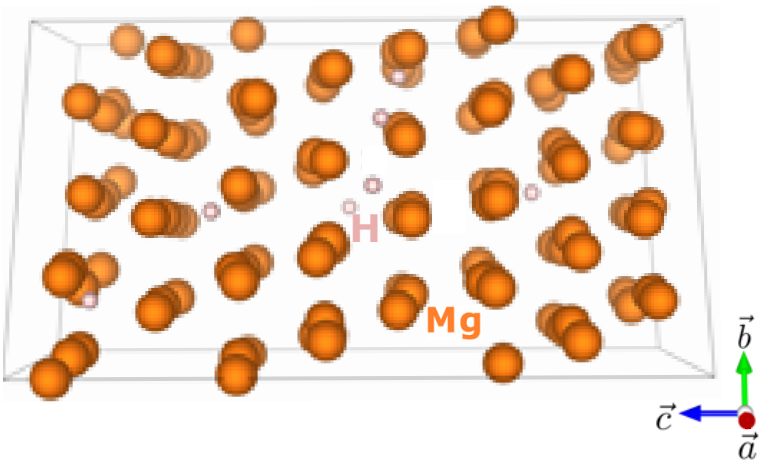


Figure 1. Perspective view of a MLFF-MD frame, at 673 K, of the $4 \times 4 \times 4$ $MgH_{0.0625}$ supercell employed in the calculations. Mg atoms are displayed in orange and H atoms in pink.

configuration in the database. Such on-the-fly procedure allows the model to use the accumulated ab-initio configurations to gradually improve the predictions on subsequent steps. The threshold value of 5 meV/Å has been set to ensure a collection of diverse and well-spread representation of structures in the database from the explored configurational phase space. During the thermalization phase, we opted for an NpT ensemble while constraining the cell shape, employing the Langevin thermostat with a friction coefficient for the lattice and atomic degrees-of-freedom equal to 10 ps^{-1}. We applied zero external pressure and a time step of 1 fs. At key temperatures of 300 K, 480 K and 673 K, we conducted further 100 ps long NpT simulations in training mode to accumulate additional configurations. A total of over 3,700 ab-initio configurations have been stored, fewer than 1,000 being recorded at each constant temperature, and the remaining configurations captured during the ramping phase. Subsequently, we switched to MLFF-MD in run mode and determined the average lattice volume over a 100 ps period at fixed temperatures of 300 K, 480 K, and 670 K. The average volume configurations were then utilized to conduct NVT simulations. Following the 100 ps NVT simulation, we extracted three structures with an energy close to the average value of the NVT run. To avoid any correlations, we ensured that the time interval between each of them was at least 10 ps. Subsequently, starting from those, we performed three distinct NVE simulations, where we computed the mean squared displacements (MSD) of hydrogen atoms as the ensamble average of

$$\mathrm{MSD}(t) = \frac{1}{T-t}\int_0^{T-t} [\mathbf{r}(t+\Delta) - \mathbf{r}(\Delta)]^2 d\Delta. \tag{1}$$

Where T is the total simulation time, and $\mathbf{r}$ is the trajectory of the atoms under analysis. During the final NVE run a reduced time step of 0.5 fs was employed to reduce energy fluctuations. All the MSD were constructed over NVE trajectories of 1 ns, and used to extract the diffusion coefficient D by fitting the linear part of the function with the Einstein relation

$$\mathrm{MSD}(t) = 6Dt \tag{2}$$

by fitting the initial 0.2 ns linear region of the simulation. The value of D at each temperature was computed as the average value of the three NVE replicas. The associated uncertainty was computed as the standard error of the mean SEM = SD/$\sqrt{3}$, where SD is the standard deviation, and 3 the number of our collected predictions. An example of MSD fitting procedures is included in the Supplementary Figure SF1.
The ciNEB calculations were performed using a 2x2x2 supercell with one Hydrogen positioned along the main paths between interstitial sites identified in the literature[17]. A visualization of those sites is reported in SF3. Also, every computation was performed using ten images and a 0.01 eV/Å forces convergence criterion.

2.2 Universal inter-atomic potentials UIPs

We consider two state-of-the-art best performing[40] pre-trained UIPs: MACE[41], based on an equivariant message passing neural network, and CHGNet[32], a graph-based neural network. The models are imported in their pre-trained versions on the Materials Project (MP) relaxation trajectories database[42], comprising 1.6 millions crystal structures with the associated energies, forces, stresses and magnetic moments. In order to tackle the predicted dynamical properties from such UIPs, we followed the same NEB and MD procedure (excluding the on-the-fly training phase) as explained in the previous section. Subsequently, CHGNet and MACE were fine-tuned on the VASP-DFT generated data obtained during the on-the-fly MLFF-MD. MACE was also trained from scratch on the dataset. Both traning and fine-tuning of all the models were performed using the standard procedure and parameters suggested by the developers in the respective github repositories. During CHGNet's MD simulations, we observed the need of a 0.5 fs time step to stabilize the dynamic also in the NPT and NVT runs, avoiding drifts in temperature for NVE. In this regard, we show examples of temperature oscillations during NVE simulations from 0.5 fs and 1.0 fs cases in the SF4 and SF5. To perform the MD simulations we employed LAMMPS[43] and ASE[44], for MACE and CHGNet respectively. A schematic view of the employed workflow is represented in Fig. (2). Furthermore, using the best performing model, (MACE_FT), we followed the same protocol to estimate D at three additional hydrogen concentrations, containing 10 H ($MgH_{0.03125}$), 6 H ($MgH_{0.046875}$) and 4 H ($MgH_{0.046875}$), respectively. To assess the performance of every model, we evaluated the Root Mean Square Errors (RMSE) for the prediction of energy and forces over a test dataset of 1200 configurations. These were randomly sampled, with proper equal spacing in time, from the NVE replicas of the MACE_FT model, with 400 samples for each temperature. This validation process was independently repeated for each different hydrogen content to test the models' accuracy beyond the training concentration. At last, NEB calculation on the same path studied using VASP were performed with every model using the ciNEB implementation of ASE employing the same number of images and convergence criteria.

3 Results and Discussions

The results reported in Fig. (3) show how the predictions of the VASP-MLFF achieve an accuracy below 0.3 meV/atom for energies and below 10 meV/Å for forces, respectively, as reasonably expected compared to other studies involving MLFF-

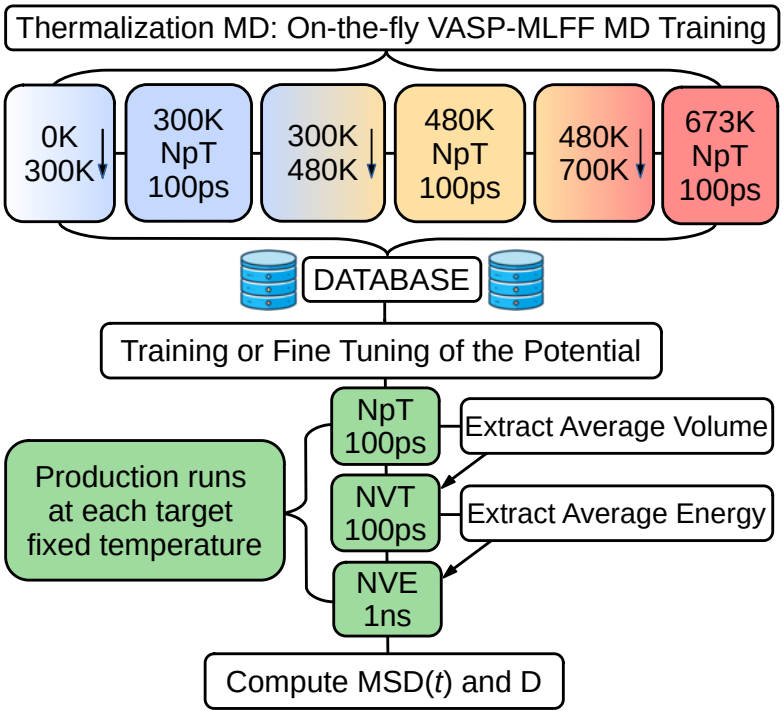


Figure 2. Methodological protocol employed to improve the performance of interatomic potentials and obtain dynamical properties. The database of configurations is built both during NpT-MD thermalizations of the system from 0 K to 700 K via active learning of the VASP-MLFF, and at target temperatures (300 K, 480 K and 673 K). Subsequentially, the machine learned potentials are fine tuned or trained, and after system-equilibration the MSD and the diffusion coefficient D at fixed T are computed.

MD[45–48]. On the other hand, the UIPs pre-trained on the MP-database, respectively named on CHGNet_MP and MACE_MP , failed to reach such level of accuracy by more than one order of magnitude. The discrepancy was significantly reduced after fine-tuning the two models on the VASP-generated database: the error obtained shows that the CHGNet_FT performance reaches a level comparable with the VASP-MLFF, and MACE_FT even outperforms it. The performance differences between CHGNet_FT and MACE_FT may stem from the equivariant architecture employed by the latter. MACE turned out to be highly data-efficient, leading to better fine-tuning results over small datasets, compared to the more data-hungry architectures like CHGNet. The MACE model was also trained from scratch on the VASP-DFT configurations (MACE_TR), achieving slightly smaller errors on forces with respect to MACE_FT. Please note that all energy values at concentrations different from the training concentration exhibit a systematic shift dependent on the hydrogen content. However, the underlying physics of hydrogen dynamics is unaffected, as this shift remains consistent throughout simulations with a constant number of particles and does not influence the forces. Consequently, the models are still capable of accurately describing the hydrogen dynamics. For completeness, we also present the pristine validation and the corresponding R^2 values, as shown in SF2.

The ciNEB calculations conducted with the ML-potentials, allowed to estimate the transition energy barrier E_b between various octahedral and tetragonal hydrogen interstitial sites, and to further validate the accuracy of the ML potentials. In fact, we compared the obtained E_b values through these methods with the VASP-DFT benchmark results. Our findings, shown in Fig. 3-b), reveal that the universal potentials perform significantly less accurately in predicting the energy barriers. However, the performance of the trained and fine-tuned counterpart dramatically improve, producing results that are much closer to the VASP-DFT reference values. The VASP-MLFF method also gave excellent agreement with the DFT benchmark. As expected, the model with better accuracy on the validation set, especially on the forces, obtained results closer to the DFT values.

The diffusion coefficient D predicted from every model during the NVE runs is reported in Fig. 4-a), and compared with experimental[19] and NEB[17] results, at each investigated temperature. The experimental data from Nishimura et. al. were collected in the temperature range of 474-493 K and then extrapolated at higher and lower temperatures using an Arrhenius relation. The results clearly show that our procedure provides multiple solutions with excellent experimental agreement, outperforming NEB computations[16] that until now has represented the standard for such applications. This holds true for the MACE_FT potential in particular, which not only predicts the correct order of magnitude across all temperatures, but

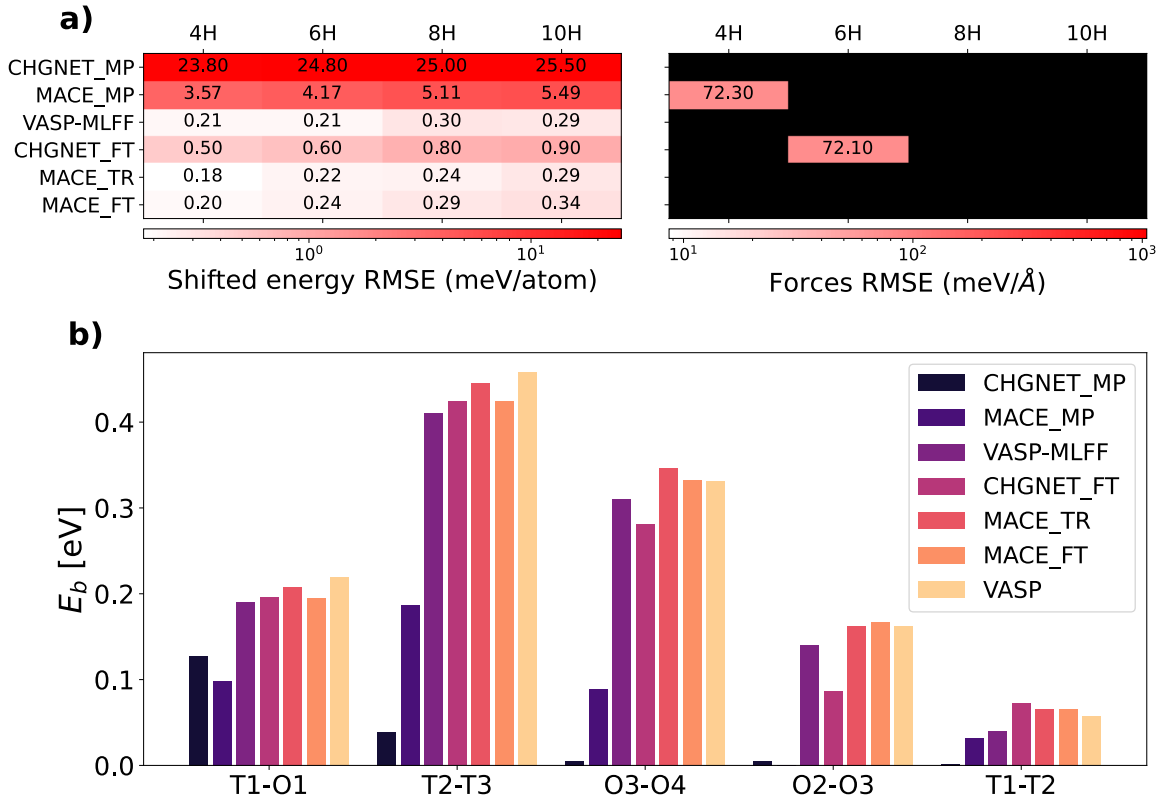


Figure 3. a) RMSE values for the predictions of energies and forces on the validation set from every model considered in the study. These results are given at different concentrations, obtained using a supercell with a variable number of H atoms, removing the corresponding constant energy shift by setting the mean of all the configuration at same concentration to zero. The subscript 'MP' refers to the pre-trained model on the Material Project database, 'TR' to the one trained from scratch and 'FT' to the fine-tuned version. b) Energy barrier values associated to the hydrogen transition between different octahedral (O_i) and tetragonal (T_i) sites, obtained through ciNEB calculations, using different ML-potentials and DFT reference (in yellow).

significantly agrees at 480 K. The VASP-MLFF instead, shows a good agreement at 480 K and 673 K, while missing the room temperature by one order of magnitude. Concerning CHGNet_FT, it provides remarkable agreement with the experimental value at 480 K, but at the lowest temperatures underestimates the result by one order of magnitude. As expected from the error analysis on energies and forces, both the pre-trained versions of the UIPs result in much lower agreement, with deviations of at least one order of magnitude from experimental values at most temperatures. A closer inspection of the temperature dependence of the diffusion coefficients, as shown in Fig. 4-b), reveals that CHGNet_FT still outperforms the VASP-MLFF by better reproducing the Arrhenius curve observed in the experiments, while the MACE_FT solutions outperform both. This shows how with smaller errors, the deep networks are capable to better represent the shape of the energy landscape, with respect to the Bayesian alternative. In this regard, further quantitative analysis can be provided by comparing the predicted value of the activation energy E_a, obtained from the linear fits in the Arrhenius plots. In particular, we achieved values of 0.4 eV for VASP-MLFF, 0.19 eV for CHGNet_FT and 0.28 eV for MACE_FT, where the experiment value corresponds to 0.25 eV[19]. MACE_FT clearly excels at representing the energy landscape of the system.

Furthermore, we computed D at various hydrogen concentrations using MACE_FT, along with the associated E_a, as shown in Fig. 6-a). The results reveal a decrease in hydrogen mobility with increasing concentration, leading to a corresponding reduction in the diffusion coefficient. Simultaneously, the activation energy increases with higher hydrogen densities. The evaluation of D across different concentrations enabled a comparison with experimental values, demonstrating that the results for lower concentrations, specifically $MgH_{0.03}$ and $MgH_{0.045}$, show the best alignment. It should be noted that Nishimura et al.[19] did

not report the hydrogen concentration within their samples. These findings suggest that the target hydrogen concentration in Nishimura's experiments may be lower than 0.0625, assumed in previous DFT studies[17].
Further analysis have been performed on the dynamics of the system by evaluating the radial distribution function (RDF) in all of the NVE run performed. We report in Fig. 4-c) the predicted behaviour by the best performing models in the system at 673 K, while other temperature cases can be found in Fig. SF6. A very good agreement between MACE_FT and VASP-MLFF is found, while the RDF of CHGNet_FT departs from the others at larger distances, by smoothing out peaks. From such curves it is possible to retain information about the behaviour of hydrogen during the simulation. Proceeding in order with increasing distance radius, the first peak appears just before 2 Å in the Mg-H curve, in correspondence of the average distance of one Magnesium atom from the center of the nearest octahedral sites on which hydrogen tend to sit[16]. The second peak, belonging to the H-H pair, is very pronounced at around 2.6 Å, reflecting a correlation between hydrogen atoms at this distance. This behaviour finds agreement in the literature for molecular dynamics with magnesium hydride nanoclusters[30]. In fact, the distance of 2.6 Å corresponds to the one between two octahedral sites along the c-direction[16], as also reported in literature[30], implying how hydrogen tend to sit on near sites. To highlight such behaviour we also report in Fig. (5) the extensive diffusion path of a representative hydrogen atom within the magnesium matrix during a 100 ps simulation, at 673 K. The trajectory

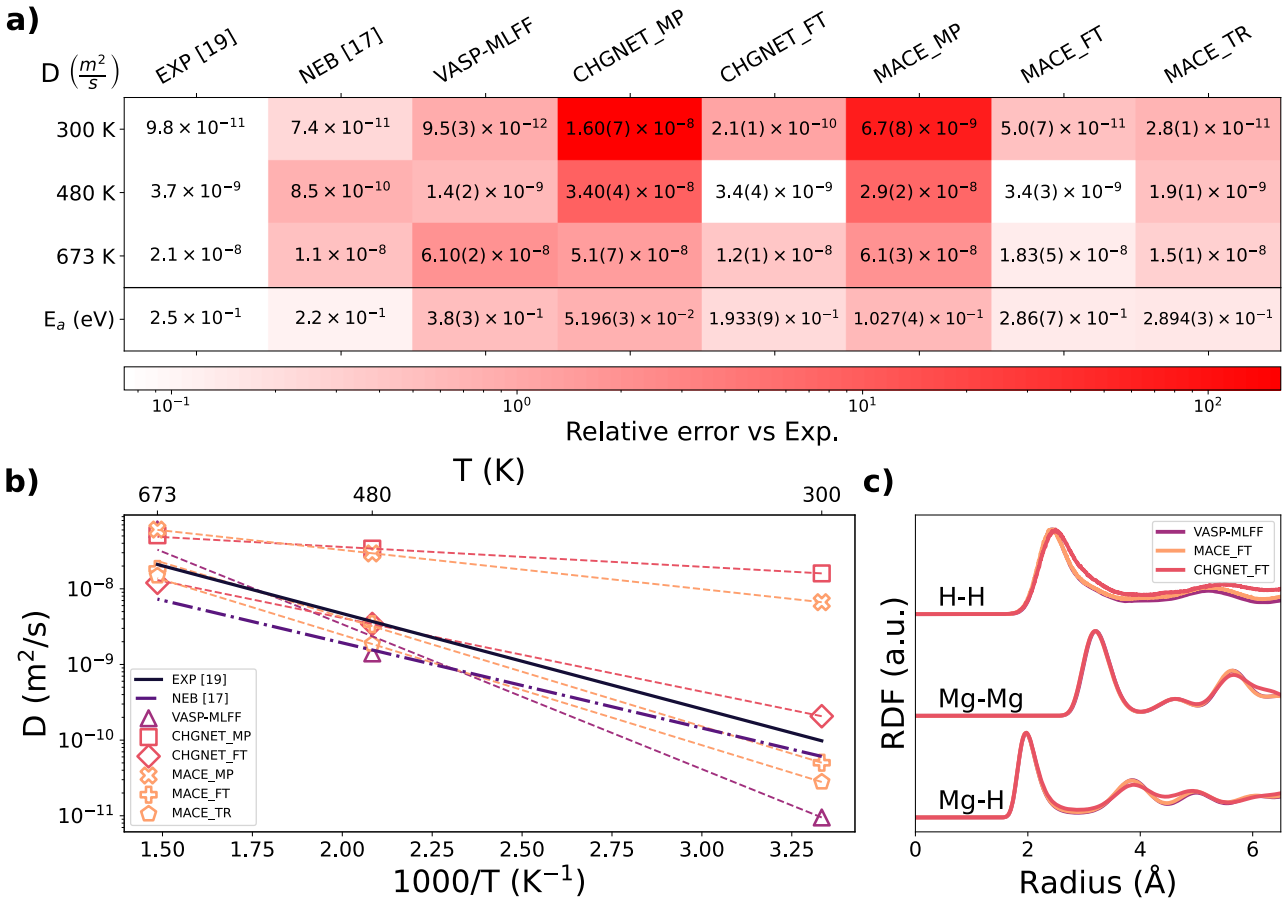


Figure 4. a) Diffusion coefficient values of hydrogen in $MgH_{0.0625}$ at three different constant average temperatures of 380 K, 480 K and 673 K, and the corresponding activation energy, comparing results achieved via different potentials in our investigations, and previous studies[17,19]. The uncertainty is reported in bracket for the last digit. The colorbar highlights in logscale the relative deviation with respect to the experimental values. b) Comparison between the dependence of the diffusion coefficient with temperature, for ML-models, the NEB[17] and experimental results[19]. The experimental curve is extrapolated from data obtained at 474-493 K. c) Radial distribution function of atoms pairs at 673 K during 1 ns long NVE simulations. VASP-MLFF and MACE show a remarkable agreement over the whole domain for every pair, while CHGNet starts to differ at larger distances.

is unwrapped in the replicas of the periodic images to enhance visibility and interpretation. The color gradient serves as a temporal marker, with blue indicating the initial position of the hydrogen atom at the beginning of the simulation (t = 0 ps) and red indicating its position at the end of the studied interval (t = 100 ps). Intermediate colors (cyan, green, yellow, and orange) represent the progression of time between these two extremes, providing a visual cue for the temporal evolution of the atom's diffusion path. The black circles highlight interstitial regions where the hydrogen atom tends to oscillate around the magnesium sites, indicating temporary trapping sites within the lattice structure, before continuing its diffusion trajectory. The third significant peak in the RDF is observed around 3 Å in the Mg-Mg curve, consistently with the typical magnesium distances in hcp structures. Multiple smaller peaks indicate further neighbor interactions in the crystal lattice. Analogous results were found in the RDF at 300 K and 480 K, where the peaks are sharper due to the reduced effect of thermal motion, see SF6. In particular, higher pronounced RDF at lower temperature indicates that hydrogen tend to spend more time in the vicinity of Magnesium, and diffuse less in the crystal structure.

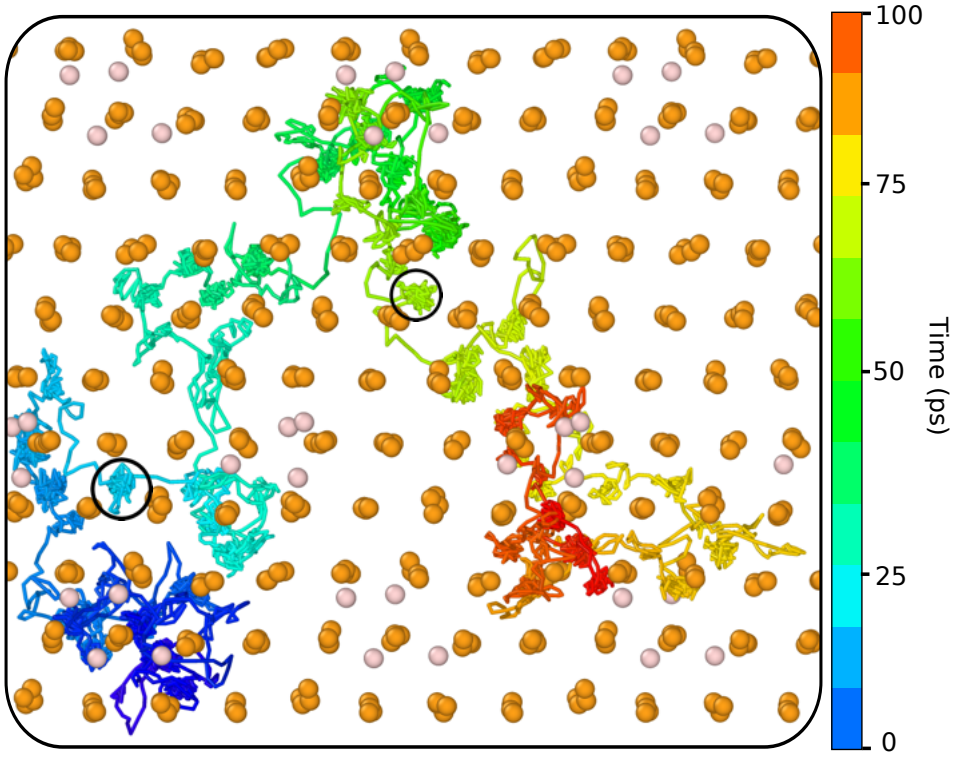


Figure 5. Diffusion path of a representative hydrogen atom in $MgH_{0.0625}$ during 100 ps at 673K, depicted using a color-gradient line to represent the progression of time, from blue to red. The black circles highlight the interstitial regions where H atoms tend to oscillate around Mg lattice site before continuing their diffusing trajectory.

The analysis of hydrogen pair (n_p) distributions presented in Fig. 6-b) provides a clear insight into the dependence of pair formation on temperature and concentration. At elevated temperatures, such as 480 K and 673 K, the simulated distributions align closely with the theoretical Poisson distribution, calculated as:

$$f(n_p;\lambda) = \frac{e^{-\lambda}\lambda^{n_p}}{n_p!},$$

where λ represents the average number of pairs. This agreement demonstrates the random nature of hydrogen pairing, indicating a lack of correlation and independence from the initial atomic configuration. Furthermore, the convergence of λ values for different concentrations as temperature increases underscores this intrinsic randomness. Conversely, at 300 K, a deviation from the Poissonian behavior emerges, with λ showing a stronger dependence on hydrogen concentration. This behavior can be attributed to the reduced mobility of hydrogen atoms at low temperatures, where the thermal energy is insufficient to overcome the energy barriers for interstitial site transitions. Consequently, the hydrogen atoms remain in proximity to their initial positions, leading to non-equilibrated pair distributions within the simulation timescale. These results suggest that, while the system approaches a random distribution at higher temperatures, longer simulations would be required to achieve equilibration and Poissonian behavior at room temperature for low-mobility regimes.

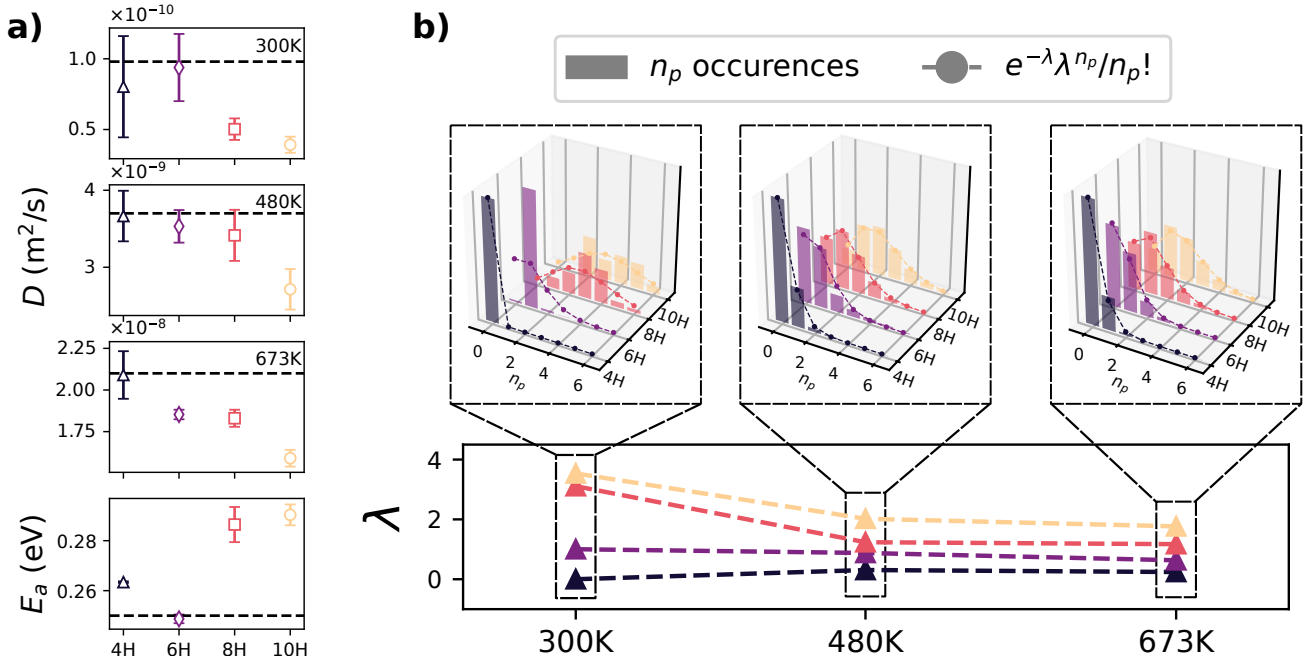


Figure 6. Results obtained from runs at different hydrogen content using MACE_FT. a) Diffusion coefficients at different temperature and H concentration, on top, and obtained activation energy at different concentration, on the bottom. The target experimental value from[19] is reported as the dashed line in every plot. b) Comparison of the simulated hydrogen pair distributions (n_p occurrences, bars) with theoretical Poisson distributions (points and dashed lines) at three temperatures (300 K, 480 K, and 673 K) for varying hydrogen concentrations. The lower panel shows the λ parameter (average number of pairs) as a function of temperature and concentration.

4 Conclusion

Our investigation enabled to thoroughly characterize the kinetic properties and mobility of hydrogen within a structured environment, such as pure magnesium, across various temperatures through a rigorous and efficient methodology. We performed a systematic and comparative study of ML-based interatomic potential MD schemes, particularly focusing on Bayesian versus equivariant and graph neural networks (MACE and CHGNet), under different training modes (universal and fine-tuned). The results were validated by estimates of the diffusion coefficient and the activation energy values, which showed excellent agreement with experimental data. The obtained results proved that the VASP-DFT configurations, collected during the MLFF on-the-fly training, represent a complete set for the accurate modeling of inter-atomic interactions, between hydrogen and magnesium atoms. This strategy could be consistently applied to generate a comprehensive dataset for the proper training of existing or forthcoming potentials. In fact, we identified the limitations of pretrained UIPs in studying materials with diffusing defects, due to the absence of representative high temperatures, defective and metastable states in their datasets. Particularly, we demonstrated the ability of state-of-the-art machine learning models to achieve DFT-level accuracy, after fine-tuning on actively learned DFT-configurations, highlighting the importance of focusing on efficient dataset-building methods and their quality. Specifically, the importance of including defected and transition-state configurations, as well as the role of transfer-learning, allowing pre-trained solutions to adapt to new systems. The obtained data offer valuable new insights into the collective dynamical properties at varying hydrogen concentrations, accurately predicting a decrease in hydrogen mobility as concentration increases. Additionally, the statistical analysis of hydrogen pair distributions indicates the system approaches a random distribution at higher temperatures. However, longer simulations would be necessary to achieve equilibration and Poissonian behavior at room temperature in low-mobility regimes.

These progresses not only improve our understanding of hydrogen interactions in magnesium, but also pave the way for future research into a broader range of multi-component systems and defected compositions. This is especially significant for systems with complex potential energy surfaces, where traditional ab-initio methods become impractical. Successfully modeling the hydrogen diffusion mechanism in magnesium via machine learning-accelerated molecular dynamics could facilitate the study and discovery of new, more efficient materials for hydrogen storage and beyond it, contributing to the transition towards a greener and more sustainable energy future.

Acknowledgments

C. F. and A. A. acknowledges the ”Doctoral College Advanced Functional Materials – Hierarchical Design of Hybrid Systems DOC 85 doc.funds” funded by the Austrian Science Fund (FWF) and by the Vienna Doctoral School in Physics (VDSP). For Open Access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission. D. M. and S. P. were supported by the European Union Horizon 2020 research and innovation program under Grant Agreement No. 857470, from the European Regional Development Fund under the program of the Foundation for Polish Science International Research Agenda PLUS, grant No. MAB PLUS/2018/8, and the initiative of the Ministry of Science and Higher Education ’Support for the activities of Centers of Excellence established in Poland under the Horizon 2020 program’ under agreement No. MEiN/2023/DIR/3795. L.P. and C.F. acknowledges the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Project NEST (Network 4 Energy Sustainable Transition) of Ministero dell’Università e della Ricerca (MUR), funded by the European Union – NextGenerationEU. L.L. and C.F. acknowledges the NRRP, CN-HPC grant no. (CUP) J33C22001170001, SPOKE 7, of MUR, funded by the European Union – NextGenerationEU. The computational results were obtained using the Vienna Scientific Cluster (VSC) and the LEONARDO cluster. We acknowledge access to LEONARDO at CINECA, Italy, via an AURELEO (Austrian Users at LEONARDO supercomputer) project

Supplementary material

Supplementary information and figures are available at website...
The training dataset can be openly accessed at https://figshare.com/articles/dataset/_b_ML_AB_b_/27922470?file=50852211.

Author Contribution

Conceptualisation: C.F., A. A., D. M. and L. L. Methodology and calculations: A. A., L.L. and D. M. Supervision: C.F. Writing: A.A , D.M. , L. L. and C. F. Discussion and reviewing: All authors

Competing Interests

The authors declare no competing interests.

References

1. Yang, J., Sudik, A., Wolverton, C. & Siegel, D. J. High capacity hydrogen storage materials: Attributes for automotive applications and techniques for materials discovery. *Chem. Soc. Rev.* **39**, 656 – 675 (2010).
2. Bhimineni, S. H. *et al.* Machine-learning-assisted investigation of the diffusion of hydrogen in brine by performing molecular dynamics simulation. *Ind. & Eng. Chem. Res.* **62**, 21385–21396, DOI: 10.1021/acs.iecr.3c01957 (2023). https://doi.org/10.1021/acs.iecr.3c01957.
3. Tarasov, B. P. *et al.* Metal hydride hydrogen storage and compression systems for energy storage technologies. *Int. J. Hydrog. Energy* **46**, 13647–13657 (2021).
4. Turner, J. A. A realizable renewable energy future. *Science* **285**, 687 – 689 (1999).
5. Hydrogen production from renewables. *Renew. Energy Focus.* **9**, 34–37 (2009).
6. Aceves, S. M., Berry, G. D., Martinez-Frias, J. & Espinosa-Loza, F. Vehicular storage of hydrogen in insulated pressure vessels. *Int. J. Hydrog. Energy* **31**, 2274–2283 (2006).
7. Sun, Y. *et al.* Tailoring magnesium based materials for hydrogen storage through synthesis: Current state of the art. *Energy Storage Mater.* **10**, 168–198, DOI: https://doi.org/10.1016/j.ensm.2017.01.010 (2018).
8. Felderhoff, M., Weidenthaler, C., von Helmolt, R. & Eberle, U. Hydrogen storage: the remaining scientific and technological challenges. *Phys. Chem. Chem. Phys.* **9**, 2643–2653 (2007).
9. Hua, T. *et al.* Technical assessment of compressed hydrogen storage tank systems for automotive applications. *Int. J. Hydrog. Energy* **36**, 3037–3049 (2011).
10. US Department of Energy. Doe technical targets for onboard hydrogen storage for light-duty vehicles.
11. Yartys, V. *et al.* Magnesium based materials for hydrogen based energy storage: Past, present and future. *Int. J. Hydrog. Energy* **44**, 7809–7859, DOI: https://doi.org/10.1016/j.ijhydene.2018.12.212 (2019). A special issue on hydrogen-based Energy storage.
12. Hirscher, M. *et al.* Materials for hydrogen-based energy storage – past, recent progress and future outlook. *J. Alloy. Compd.* **827**, 153548, DOI: https://doi.org/10.1016/j.jallcom.2019.153548 (2020).
13. Klopčič, N., Grimmer, I., Winkler, F., Sartory, M. & Trattner, A. A review on metal hydride materials for hydrogen storage. *J. Energy Storage* **72**, 108456, DOI: https://doi.org/10.1016/j.est.2023.108456 (2023).
14. Allendorf, M. D. *et al.* Challenges to developing materials for the transport and storage of hydrogen. *Nat. Chem.* **14**, 1214–1223, DOI: 10.1038/s41557-022-01056-2 (2022).
15. Gupta, A. & Faisal, M. Magnesium based multi-metallic hybrids with soot for hydrogen storage. *Int. J. Hydrog. Energy* **53**, 93–104, DOI: https://doi.org/10.1016/j.ijhydene.2023.12.020 (2024).
16. Schimmel, H., Kearley, G., Huot, J. & Mulder, F. Hydrogen diffusion in magnesium metal (α phase) studied by ab initio computer simulations. *J. Alloy. Compd.* **404-406**, 235–237, DOI: https://doi.org/10.1016/j.jallcom.2004.11.100 (2005). Proceedings of the 9th International Symposium on Metal-Hydrogen Systems, Fundamentals and Applications (MH2004).
17. Klyukin, K., Shelyapina, M. G. & Fruchart, D. Dft calculations of hydrogen diffusion and phase transformations in magnesium. *J. Alloy. Compd.* **644**, 371–377, DOI: https://doi.org/10.1016/j.jallcom.2015.05.039 (2015).
18. Shelyapina, M. G. Hydrogen diffusion on, into and in magnesium probed by dft: A review. *Hydrogen* **3**, 285–302, DOI: 10.3390/hydrogen3030017 (2022).
19. Nishimura, C., Komaki, M. & Amano, M. Hydrogen permeation through magnesium. *J. Alloy. Compd.* **293-295**, 329–333, DOI: https://doi.org/10.1016/S0925-8388(99)00373-4 (1999).
20. Uchida, H. *et al.* Absorption kinetics and hydride formation in magnesium films: Effect of driving force revisited. *Acta Materialia* **85**, 279–289, DOI: https://doi.org/10.1016/j.actamat.2014.11.031 (2015).
21. Renner, J. & Grabke, H. J. Bestimmung von diffusionskoeffizienten bei der hydrierung von legierungen. *Int. J. Mater. Res.* **69**, 639–642, DOI: doi:10.1515/ijmr-1978-691005 (1978).
22. Kwon, H., Shiga, M., Kimizuka, H. & Oda, T. Accurate description of hydrogen diffusivity in bcc metals using machine-learning moment tensor potentials and path-integral methods. *Acta Materialia* **247**, 118739, DOI: https://doi.org/10.1016/j.actamat.2023.118739 (2023).
23. Tang, H. *et al.* Reinforcement learning-guided long-timescale simulation of hydrogen transport in metals. *Adv. Sci.* **11**, 2304122, DOI: https://doi.org/10.1002/advs.202304122 (2024). https://onlinelibrary.wiley.com/doi/pdf/10.1002/advs.202304122.

24. Kimizuka, H., Thomsen, B. & Shiga, M. Artificial neural network-based path integral simulations of hydrogen isotope diffusion in palladium. *J. Physics: Energy* **4**, 034004, DOI: 10.1088/2515-7655/ac7e6b (2022).

25. Lu, G. M., Witman, M., Agarwal, S., Stavila, V. & Trinkle, D. R. Explainable machine learning for hydrogen diffusion in metals and random binary alloys. *Phys. Rev. Mater.* **7**, 105402, DOI: 10.1103/PhysRevMaterials.7.105402 (2023).

26. Friederich, P., Häse, F., Proppe, J. & Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. *Nat. Mater.* **20**, 750–761, DOI: 10.1038/s41563-020-0777-6 (2021).

27. Yu, F., Xiang, X., Zu, X. & Hu, S. Hydrogen diffusion in zirconium hydrides from on-the-fly machine learning molecular dynamics. *Int. J. Hydrog. Energy* **56**, 1057–1066, DOI: https://doi.org/10.1016/j.ijhydene.2023.12.241 (2024).

28. Vandenhaute, S., Cools-Ceuppens, M., DeKeyser, S., Verstraelen, T. & Van Speybroeck, V. Machine learning potentials for metal-organic frameworks using an incremental learning approach. *npj Comput. Mater.* **9**, 1–8 (2023).

29. Kývala, L., Angeletti, A., Franchini, C. & Dellago, C. Diffusion and coalescence of phosphorene monovacancies studied using high-dimensional neural network potentials. *The J. Phys. Chem. C* **127**, 23743–23751, DOI: 10.1021/acs.jpcc.3c05713 (2023). https://doi.org/10.1021/acs.jpcc.3c05713.

30. Wang, N. & Huang, S. Molecular dynamics study on magnesium hydride nanoclusters with machine-learning interatomic potential. *Phys. Rev. B* **102**, 094111, DOI: 10.1103/PhysRevB.102.094111 (2020).

31. Lu, G. M., Witman, M., Agarwal, S., Stavila, V. & Trinkle, D. R. Explainable machine learning for hydrogen diffusion in metals and random binary alloys. *Phys. Rev. Mater.* **7**, 105402, DOI: 10.1103/PhysRevMaterials.7.105402 (2023).

32. Deng, B., Zhong, P., Jun, K. *et al.* Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5**, 1031–1041, DOI: 10.1038/s42256-023-00716-3 (2023).

33. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csanyi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. In Koyejo, S. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 35, 11423–11436 (Curran Associates, Inc., 2022).

34. Jinnouchi, R., Karsai, F., Verdi, C., Asahi, R. & Kresse, G. Descriptors representing two- and three-body atomic distributions and their effects on the accuracy of machine-learned inter-atomic potentials. *The J. Chem. Phys.* **152**, 234102, DOI: 10.1063/5.0009491 (2020). https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/5.0009491/15575269/234102_1_online.pdf.

35. Jinnouchi, R., Karsai, F. & Kresse, G. On-the-fly machine learning force field generation: Application to melting points. *Phys. Rev. B* **100**, 014105, DOI: 10.1103/PhysRevB.100.014105 (2019).

36. Henkelman, G., Uberuaga, B. P. & Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The J. Chem. Phys.* **113**, 9901–9904, DOI: 10.1063/1.1329672 (2000). https://pubs.aip.org/aip/jcp/article-pdf/113/22/9901/19259681/9901_1_online.pdf.

37. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186, DOI: 10.1103/PhysRevB.54.11169 (1996).

38. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775, DOI: 10.1103/PhysRevB.59.1758 (1999).

39. Perdew, J. P., Burke, K. & Wang, Y. Generalized gradient approximation for the exchange-correlation hole of a many-electron system. *Phys. Rev. B* **54**, 16533–16539, DOI: 10.1103/PhysRevB.54.16533 (1996).

40. Riebesell, J. *et al.* Matbench discovery–an evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920* (2023).

41. Batatia, I. *et al.* A foundation model for atomistic materials chemistry (2023). 2401.00096.

42. Jain, A. *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials* **1** (2013).

43. Thompson, A. P. *et al.* LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.* **271**, 108171, DOI: 10.1016/j.cpc.2021.108171 (2022).

44. Larsen, A. H. *et al.* The atomic simulation environment—a python library for working with atoms. *J. Physics: Condens. Matter* **29**, 273002, DOI: 10.1088/1361-648X/aa680e (2017).

45. Verdi, C., Ranalli, L., Franchini, C. & Kresse, G. Quantum paraelectricity and structural phase transitions in strontium titanate beyond density functional theory. *Phys. Rev. Mater.* **7**, L030801, DOI: 10.1103/PhysRevMaterials.7.L030801 (2023).

46. Ranalli, L. *et al.* Temperature-dependent anharmonic phonons in quantum paraelectric ktao3 by first principles and machine-learned force fields. *Adv. Quantum Technol.* **6**, 2200131, DOI: https://doi.org/10.1002/qute.202200131 (2023). https://onlinelibrary.wiley.com/doi/pdf/10.1002/qute.202200131.

47. Verdi, C., Karsai, F., Liu, P., Jinnouchi, R. & Kresse, G. Thermal transport and phase transitions of zirconia by on-the-fly machine-learned interatomic potentials. *npj Comput. Mater.* **7**, 156 (2021).

48. Liu, P., Verdi, C., Karsai, F. & Kresse, G. Phase transitions of zirconia: Machine-learned force fields beyond density functional theory. *Phys. Rev. B* **105**, L060102, DOI: 10.1103/PhysRevB.105.L060102 (2022).

Supplementary Materials for

"Hydrogen Diffusion in Magnesium Using Machine Learning Potentials: a comparative study"

Andrea Angeletti,[1] Luca Leoni,[2] Dario Massa,[3] Luca Pasquini,[2] Stefanos Papanikolaou,[3] and Cesare Franchini[2,4]

[1] *University of Vienna, Vienna Doctoral School in Physics, Boltzmanngasse 5, 1090 Vienna, Austria*
[2] *Department of Physics and Astronomy 'Augusto Righi', Alma Mater Studiorum - Università di Bologna, Bologna, 40127 Italy*
[3] *NOMATEN Centre of Excellence, National Center for Nuclear Research, 05-400 Otwock, Poland*
[4] *University of Vienna, Faculty of Physics and Center for Computational Materials Science, Kolingasse 14-16, 1090 Vienna, Austria*

In this section we show supplementary figures to the main work:

Fig. SF1 shows the mean square displacement of one the VASP-MLFF replicas at the highest investigated temperature, and the component decomposition along three directions. Overall across the various simulations we have not observed a consistent difference in the behavior between the different components.

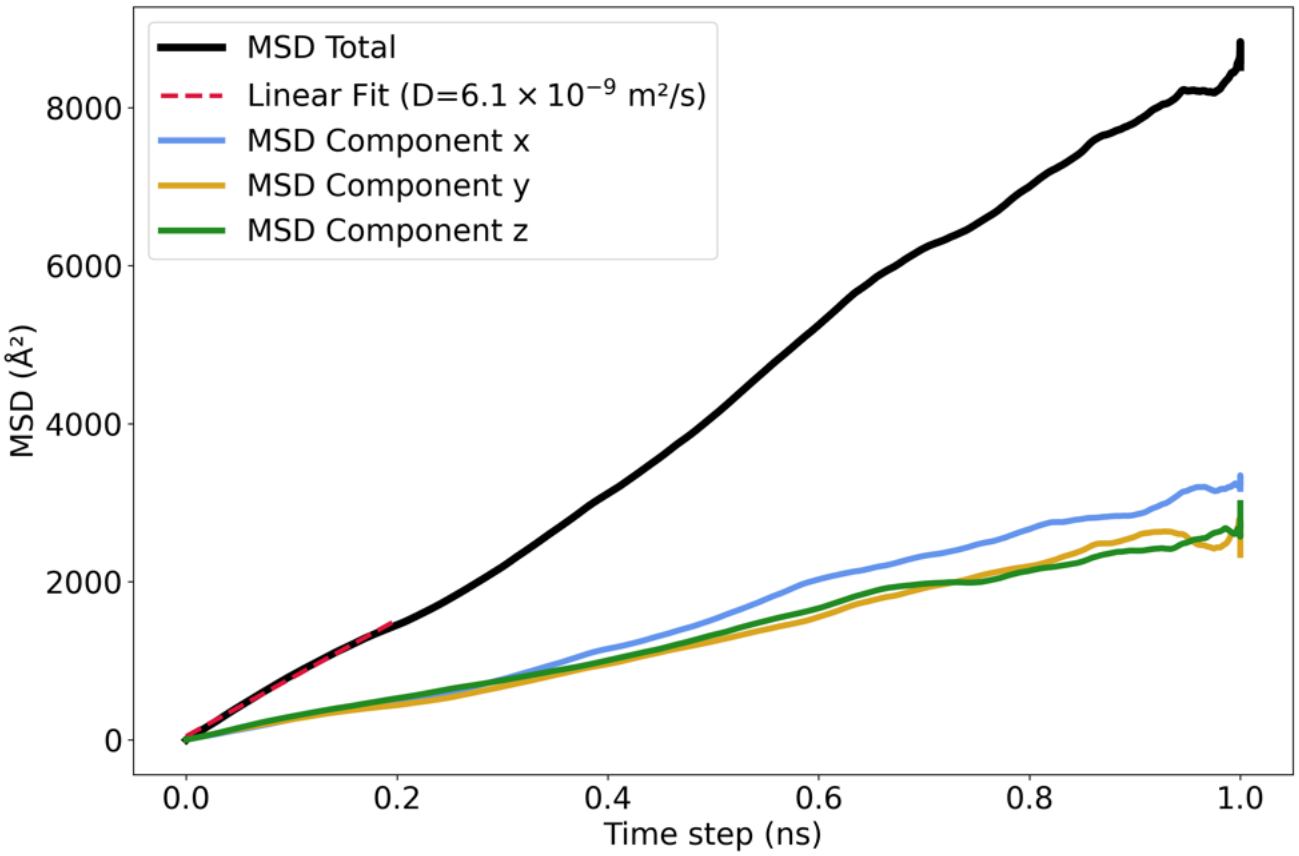
Fig. SF2 shows the un-shifted energy prediction errors of the models during their validation and the corresponding R^2 values.

Fig. SF3 shows the octahedral O_i and tetragonal T_i interstitial hydrogen sites in the magnesium matrix, between which we computed the associated energy barrier via ciNEB, with the ML-models and VASP-DFT.

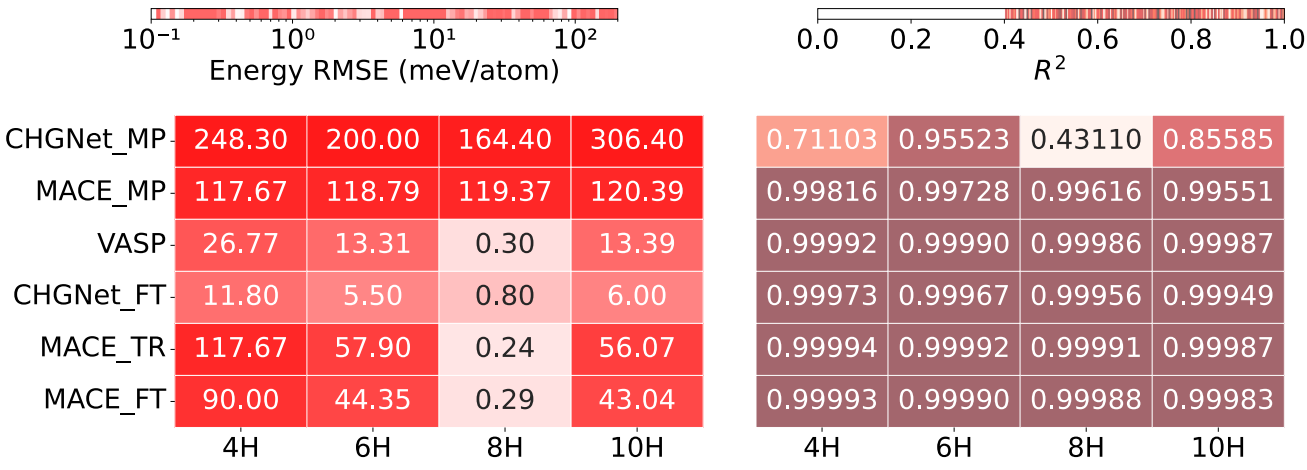
Fig. SF4 shows the drifting in energy for the CHGNet_MP model using a time step of 1 fs. By reducing it to 0.5 fs the drifting disappears as can be see in Fig. SF5. Moreover, we can also observe the temperature fluctuations during the molecular dynamics simulations of CHGNet_MP model, which turned to out to be most unstable model overall.

Fig. SF6 shows the Radial Distribution Function of all atom pairs for three ML-model at 480 K and 300 K.

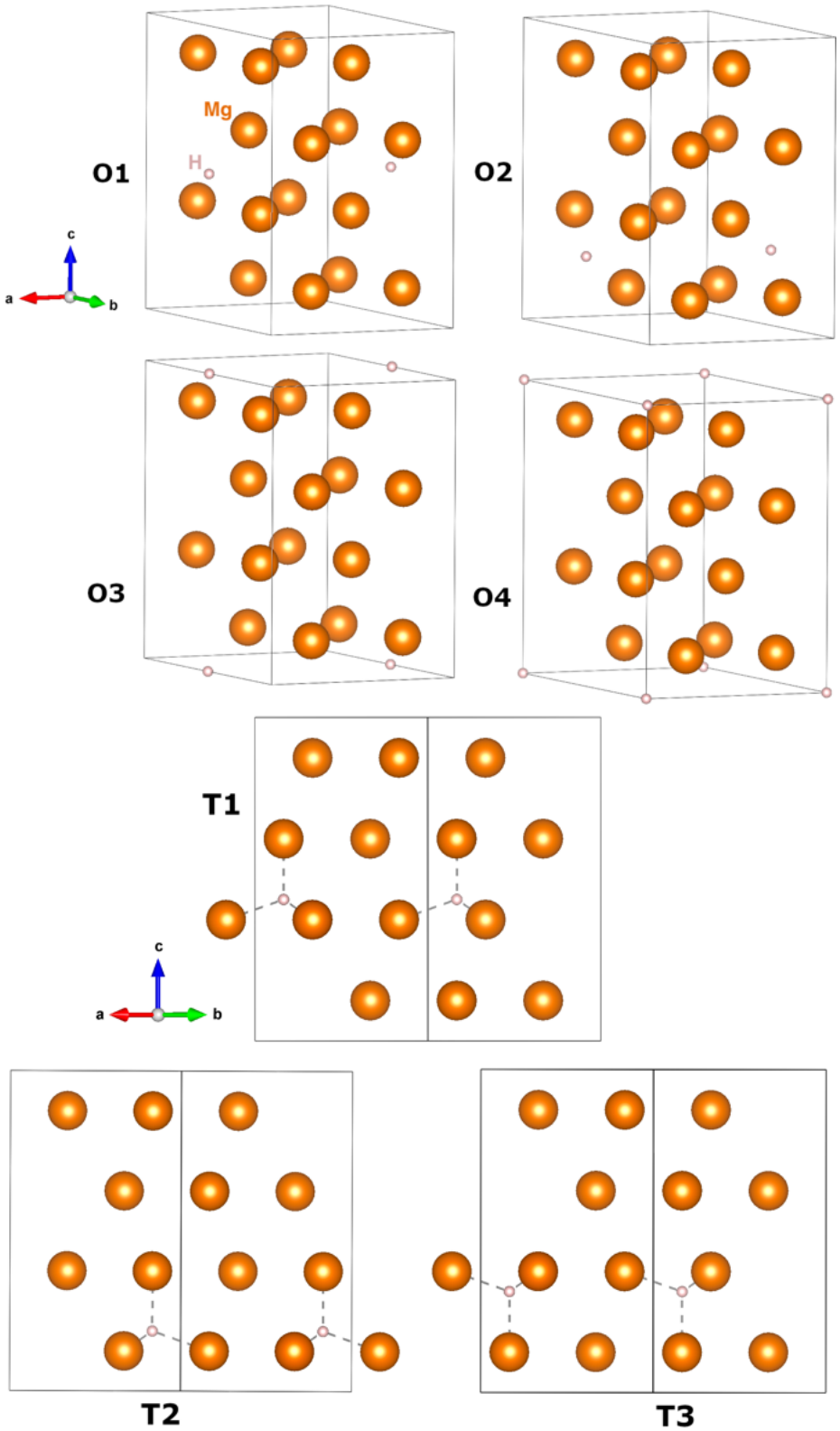
Fig. SF7 shows estimates of the MSD at different temperature for the MACE_FT model using different time lags (parameter δ of MSD formula in Ref. [1]). In particular, the results are calculated for the standard $\sigma = 1$ and the fully uncorrelated $\sigma = \tau$ cases, estimating also the diffusion coefficient to show a possible dependence on the parameter. The two estimates showed nearly the same linear trend with differences mainly related to the lower statistic available to the $\delta = \tau$ case that makes it more noisy. This is well reflected by the estimates of D fitted in the range described previously for the $\delta = 1$ case and in a smaller range of 0 and 0.1 ns for the uncorrelated case, where the smaller range is needed due to the much higher noise in longer times. The estimates deviates of at most 2.5 % between the two cases still remaining perfectly compatible within the estimated error, leading to variations in the estimates of the energy barrier of less than 1 %. This shows how no major correlation is present inside the MSD of the case under study, making the values obtained with $\delta = 1$, corresponding to the formula reported in the main text, more reliable due to their much larger statistic that allows for a converged estimates in the simulation time under investigation.



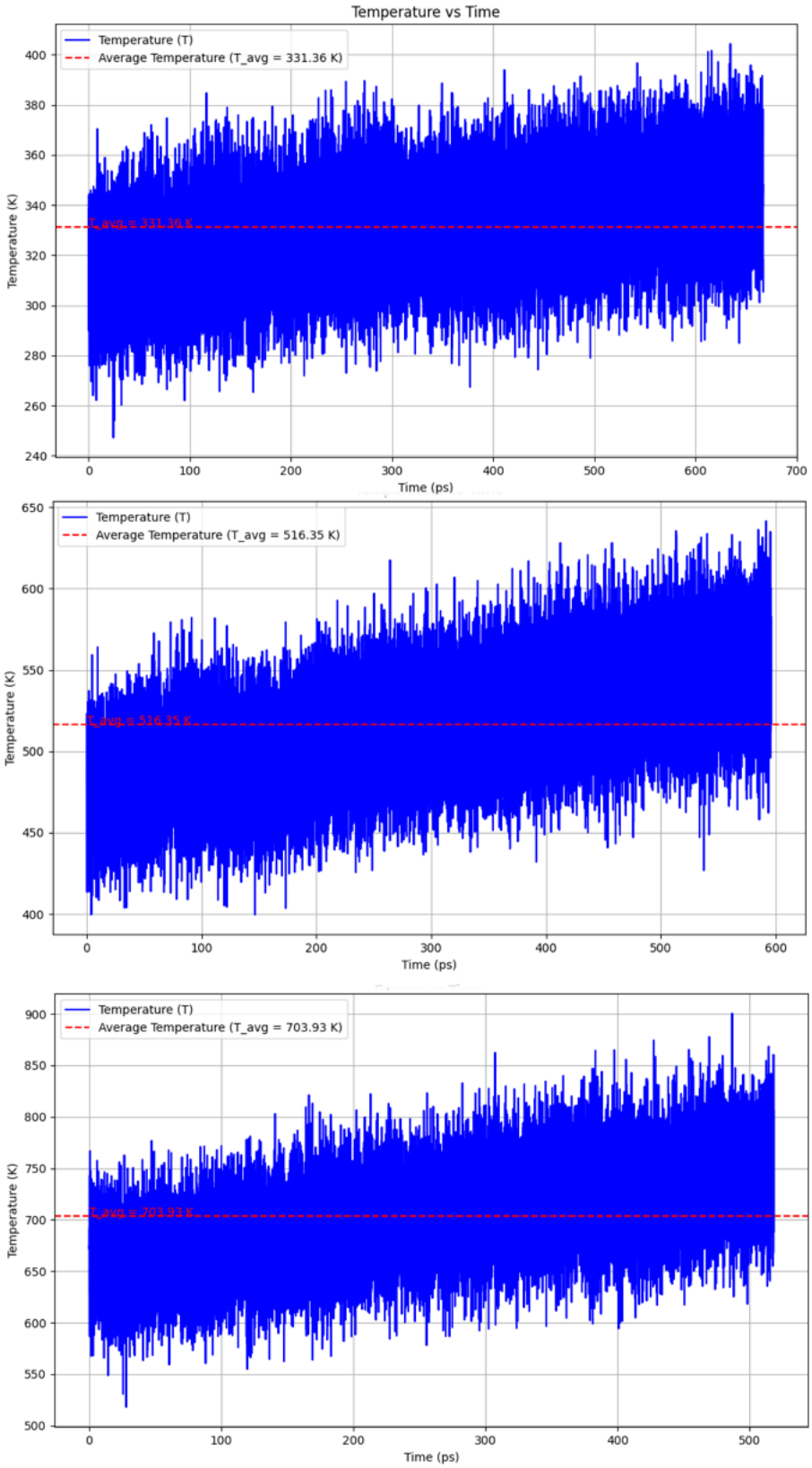
Supplementary Figure 1: Example of MSD fitting procedure for the case of VASP MLFF model at 673K and the values along different directions.



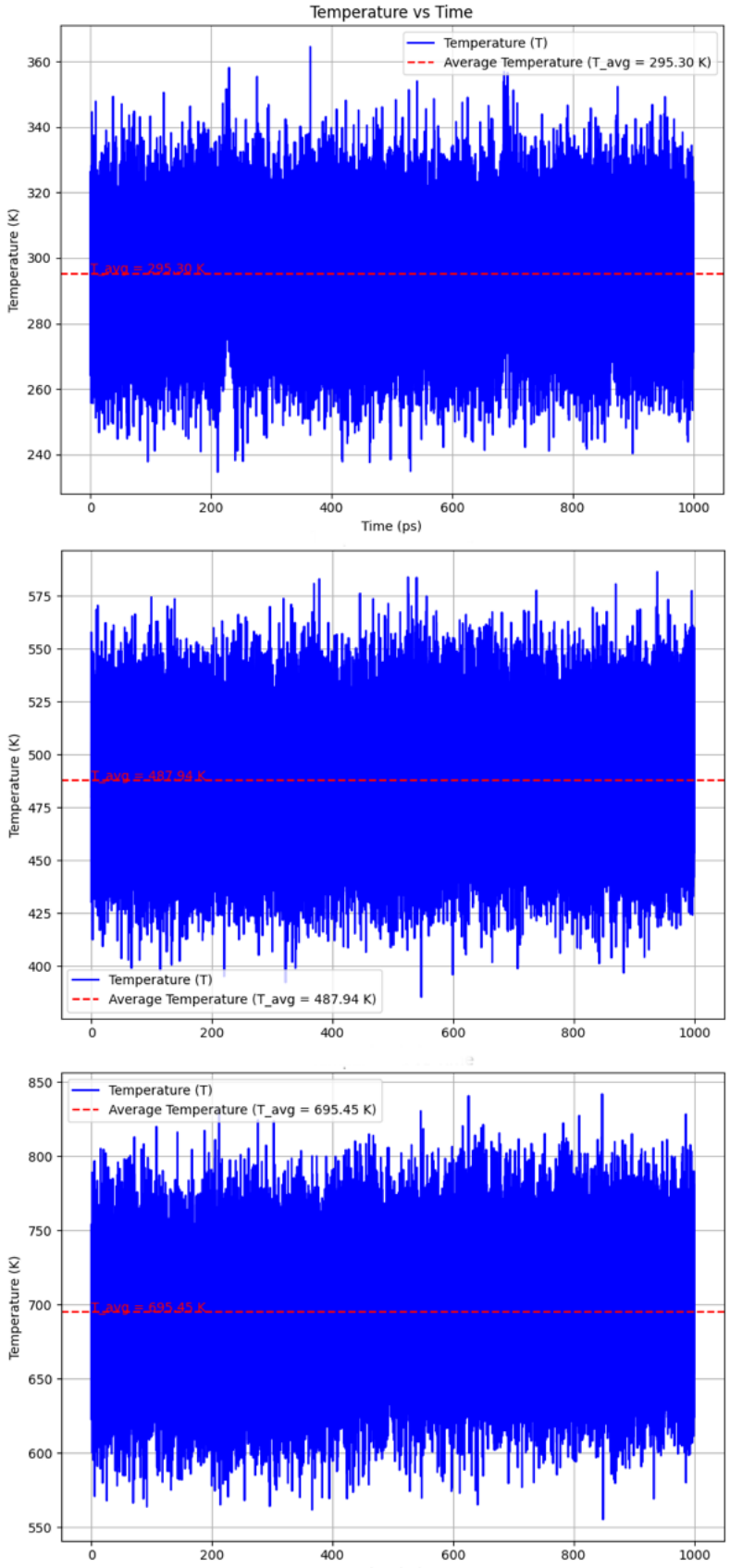
Supplementary Figure 2: The left table shows the RSME for energies and forces for all models sand concentrations, without energy shifting. The corresponding R^2 values for the predictions are on the right table.



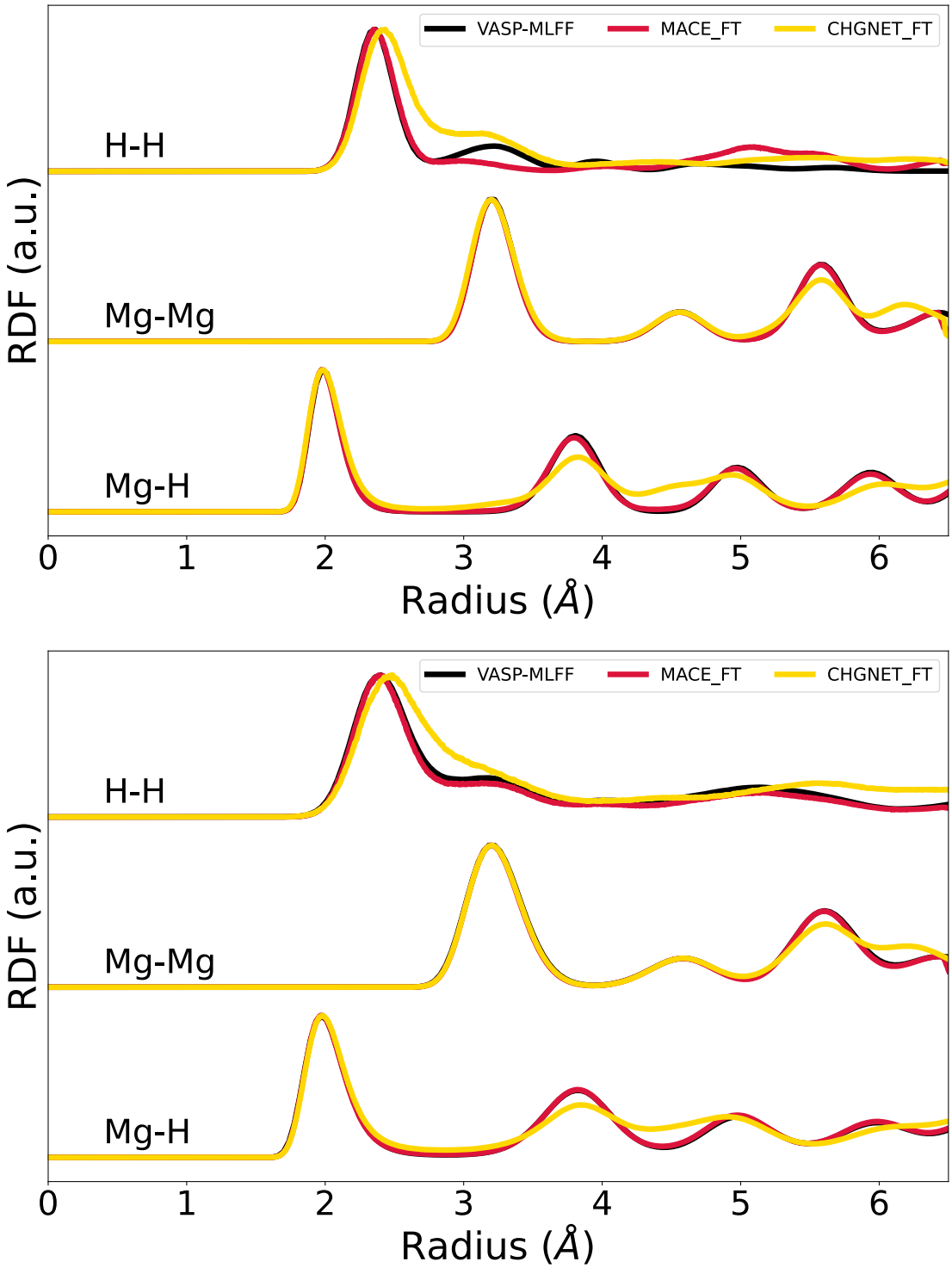
Supplementary Figure 3: Interstitial states considered for NEB calculations with all models.



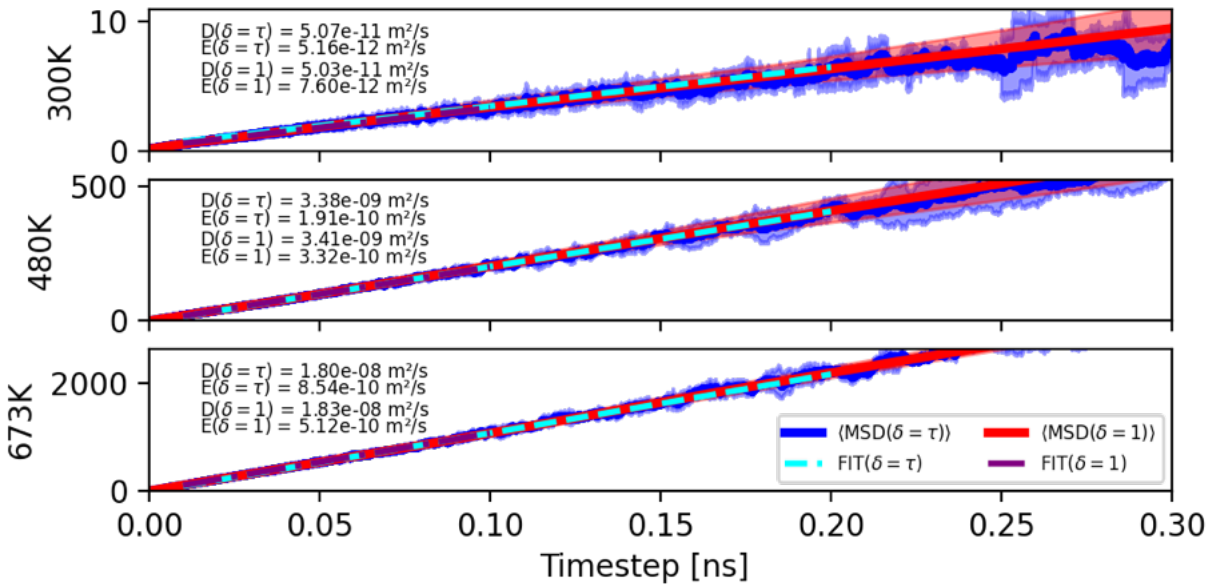
Supplementary Figure 4: Example of drifting in the NVE temperatures for the case of 1 fs timestep simulations with the pretrained CHGNet_MP model.



Supplementary Figure 5: Example of stable temperature oscillations in the NVE simulations for the case of 0.5 fs timestep with the CHGNet_MP model.



Supplementary Figure 6: Radial distribution function at different temperatures for the best performing models, at 300 K on top and 480K at the bottom.



Supplementary Figure 7: Values of the MSD at different temperatures, with following linear fit, for the MACE_FT model estimated for different time lags, δ, following the formula in Ref. [1].

[1] G. Pranami and M. H. Lamm, Journal of Chemical Theory and Computation **11**, 4586 (2015), https://doi.org/10.1021/acs.jctc.5b00574.

Chapter 3

Conclusions

The Thesis objectives were: (i) proposing statistical methods to treat electron charge density fields as reliable, bias-free descriptors for predicting and classifying material properties and defect interactions within crystalline structures, (ii) testing the predictive ability of well established atomic descriptors and ML models (e.g. graph-based) towards structure-property relationships and dynamical properties in defected materials. Regarding the first aim, performed research proves the suitability of the proposed ECD based methods for different purposes (classification, regression), target materials (defected, stable) and descriptor building techniques (post-processing of the perturbed fields, automatic extraction with DL). Regarding the second aim, the research results suggest that even models trained on very large materials databases, commonly presented as *Universal* or general-purpose, need to be adapted (fine-tuned) on relevant representative configurations for the target class of systems and their conditions in order to meet high-accuracy standards; we show this both from the point of view of static structure-property prediction tasks, as well as in defect room- and high-temperature dynamics.

The most relevant achievements and results of the research work include:

- Development of an unsupervised learning based technique for *Alloy Informatics* [4]. Impurity-induced ECD perturbations are proposed as novel descriptors of local atomic chemical environments, proving their validity in the case of hydrogen interstitial defects in a variety of FCC bulk crystals. An in-depth correlation analysis between the charge responses and host crystals lays the foundation for the method;

- Classification, within the Alloy Informatics approach [4], of the charge responses of the host crystals. Finding proof for a connection between the static density-based composition maps and the defect mobility in the matrices, hinting at the possibility of new multi-scale predictive approaches. The method can be applied to other real-case studies in which charge-density perturbations play a role (e.g. functionalization, optimization of materials properties);

- Development of a supervised learning approach on ECD fields of a variety of bulk crystals compositions for the direct regression of their properties, extracting representations directly with DL [5]. In particular, proposing low-effort strategies of transfer-learning in highly-specialized frameworks of multi-modal models (images and text). The observations underline easier predictability and interpretability of performance results in the regression of structural properties, e.g. the bulk moduli, as well as the power of multi-modalities in the presence of more challenging energetics-related properties. The approach could be applied in larger portions of open-source databases and provide powerful prediction tools, also inspiring inverse design solutions;

- Definition and application of a protocol to zero-shot test the predictions of a GNN foundation models in substitutional alloying for materials discovery [2]. During a small DFT validation, the observations underline again lower prediction errors for the structural properties, but failure in proper generalization for the prediction of energetics (e.g. formation energies);

- Development of a database-creation strategy through on-the-fly building of MLIPs for defect dynamics, in the context of hydrogen diffusion in magnesium [3]. Such database is used to fine-tune Universal (foundation) IPs;

- Providing a proof of the importance of training configurations for accurate predictions in defect dynamics [3]: the universal models, fine-tuned on the proposed database, reproduce energy and forces of the system with DFT accuracy, and deliver multiple predictions in striking accordance with experimental results, both for diffusion coefficients and transition barriers. The approach could be potentially applied to a variety of similar real-case scenarios.

Chapter 4

Other inherent research activity

Manuscript A

Amirhossein Naghdi, Franco Pellegrini, Emine Küçükbenli, Dario Massa, F. J. Dominguez–Gutierrez, Efthimios Kaxiras and Stefanos Papanikolaou. "Neural network interatomic potentials for open surface nano-mechanics applications" in Acta Materialia 277 (2024): 120200.
DOI:10.1016/j.actamat.2024.120200

Abstract: Material characterization in nano-mechanical tests may provide information on the potential heterogeneity of mechanical properties. Here, we develop a robust neural-network interatomic potential (NNIP), and we provide a test for the example of molecular dynamics (MD) nanoindentation, and the case of body-centered cubic crystalline molybdenum (Mo). We employ a similarity measurement protocol, using standard local environment descriptors, to select ab initio configurations for the training dataset that capture the behavior of the indented sample. We find that it is critical to include generalized stacking fault (GSF) configurations, featuring a dumbbell self-interstitial on the surface, to capture dislocation cores, and also high-temperature configurations with frozen atom layers for the indenter tip contact. We develop a NNIP with distinct dislocation nucleation mechanisms, realistic generalized stacking fault energy (GSFE) curves, and an informative energy landscape for the atoms on the sample surface during nanoindentation. We compare our NNIP results with nanoindentation simulations, performed with three existing potentials – an embedded atom method (EAM) potential, a gaussian approximation potential (GAP), and a tabulated GAP (tabGAP) potential – that predict different dislocation nucleation mechanisms, and display the absence of essential information on the shear stress at the sample surface in the elastic region. Finally, we compared our NNIP nanoindentation results with experiments, showing reliable predictions for reduced Young's modulus and observable slip traces.

Manuscript B

Cameron J. Owen and Amirhossein D. Naghdi and Anders Johansson and Dario Massa and Stefanos Papanikolaou and Boris Kozinsky "Unbiased Atomistic Predictions of Crystal Dislocation Dynamics using Bayesian Force Fields" in arXiv preprint (2024).

DOI: arxiv.org/abs/2401.04359

Abstract: Crystal dislocation dynamics, especially at high temperatures, represents a subject where experimental phenomenological input is commonly required, and parameter-free predictions, starting from quantum methods, have been beyond reach. This is especially true for phenomena like stacking faults and dislocation cross-slip, which are computationally intractable with methods like density functional theory, as $10^5 \div 10^6$ atoms are required to reliably simulate such systems. Hence, this work extends quantum-mechanical accuracy to mesoscopic molecular dynamics simulations and opens unprecedented possibilities in material design for extreme mechanical conditions with direct atomistic insight at the deformation mesoscale. To accomplish this, we construct a Bayesian machine-learned force field (MLFF) from ab initio quantum training data, enabling direct observations of high-temperature and high-stress dislocation dynamics in single-crystalline Cu with atomistic resolution. In doing so, a generalizable training protocol is developed for construction of MLFFs for dislocation kinetics, with wide-ranging applicability to other single element systems and alloys. The resulting FLARE MLFF provides excellent predictions of static bulk elastic properties, stacking fault widths and energies, dynamic evolutions and mobilities of edge and screw dislocations, as well as cross-slip energy barriers for screw dislocations, all of which are compared to available experimental measurements. This work marks the first reliable quantitative determination of dislocation mobilities and cross-slip barriers, demonstrating a substantial advancement over previous empirical and machine learned force field models.

Manuscript C

Grzegorz Kaszuba and Amirhossein D. Naghdi and Dario Massa and Stefanos Papanikolaou and Andrzej Jaszkiewicz and Piotr Sankowski "In-Context Learning of Physical Properties: Few-Shot Adaptation to Out-of-Distribution Molecular Graphs" in arXiv preprint (2024).

DOI:arxiv.org/abs/2406.01808

Abstract: Large language models manifest the ability of few-shot adaptation to a sequence of provided examples. This behavior, known as in-context learning, allows for performing nontrivial machine learning tasks during inference only. In this work, we address the question: can we leverage in-context learning to predict out-of-distribution materials properties? However, this would not be possible for structure property prediction tasks unless an effective method is found

to pass atomic-level geometric features to the transformer model. To address this problem, we employ a compound model in which GPT-2 acts on the output of geometry-aware graph neural networks to adapt in-context information. To demonstrate our model's capabilities, we partition the QM9 dataset into sequences of molecules that share a common substructure and use them for in-context learning. This approach significantly improves the performance of the model on out-of-distribution examples, surpassing the one of general graph neural network models.

Book Chapter

Amirhossein Naghdi, Dario Massa, Kamran Karimi and Stefanos Papanikolaou, "High Entropy Alloy Composition Design for Mechanical Properties", Chapter of Book "High Entropy Alloys - Composition and Microstructure Design", In: IntechOpen (2024).
DOI:10.5772/intechopen.1004868

Abstract: Multi-component high-entropy alloys (HEAs) are a novel class of materials exhibiting outstanding material properties that often surpassing their traditional counterparts. Despite their ubiquity, the underlying microstructure-property relationships in HEAs remain elusive. This chapter addresses this gap by exploring the application of cutting-edge machine learning tools to establish robust connections between HEAs' chemical composition, microstructure, and mechanical response. The survey begins by discussing the current state of micro-structural characterization techniques in HEAs, giving insights into their complex underlying microstructure. The development of ML force fields for HEAs based on ab initio datasets is then highlighted, addressing challenges posed by the expansive composition space associated with HEAs. The chapter further outlines machine learning-assisted composition search strategies for HEAs with specific functional properties, offering a systematic and efficient approach to explore material properties. Overall, the present overview demonstrates the potential of machine learning in unraveling the intricate nature of HEAs and accelerating their tailored design for diverse applications.

Chapter 5

Appendix

5.1 Density Functional Theory

In this section, Density Functional Theory will be introduced, following the literature [118–120]. The Hamiltonian of a N-electrons system, in the Born-Oppenheimer and non-relativistic approximation, in atomic units,

$$H\left(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\right) = -\frac{1}{2}\sum_{i=1}^{N}\nabla_{\mathbf{r}_i}^2 + \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{i=1}^{N} v_{\mathrm{ne}}\left(\mathbf{r}_i\right) \tag{5.1}$$

with the last term being the nuclei-electrons interaction potential $v_{\mathrm{ne}}\left(\mathbf{r}_i\right) = -\sum_\alpha Z_\alpha / \left|\mathbf{r}_i - \mathbf{R}_\alpha\right|$ The time-independent Schroedinger equation can be than written as

$$H\left(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\right)\Psi\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right) = E\Psi\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right) \tag{5.2}$$

being E the energy associated with a N-electrons antisymmetric wave function $\Psi\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right)$, where $\mathbf{x}_i = (\mathbf{r}_i, \sigma_i)$, with $\sigma_i \in \{\uparrow\downarrow\}$. In this framework, the variational principle allows to express the ground-state electronic energy

$$E_0 = \min_{\Psi}\langle\Psi|\hat{H}|\Psi\rangle \ , \ \langle\Psi|\Psi\rangle = 1 \tag{5.3}$$

where the Dirac notation has been used, in which $\hat{H} = \hat{T} + \hat{W}_{ee} + \hat{V}_{ne}$. In DFT, the variational principle is reformulated in terms of a one-electron density

$$n(\mathbf{r}) = N\int\ldots\int\left|\Psi\left(\mathbf{x}, \mathbf{x}_2, \ldots, \mathbf{x}_N\right)\right|^2 \ \mathrm{d}\sigma \ \mathrm{d}\mathbf{x}_2 \ldots \ \mathrm{d}\mathbf{x}_N \ , \ \int n(\mathbf{r})\, d\mathbf{r} = N \tag{5.4}$$

If in place of $v_{ne}(\mathbf{r})$ one assumes an arbitrary external potential $v(\mathbf{r})$, while in principle solving the Schroedinger equation would allow to obtain the ground state density $n(\mathbf{r})$ and so to define

a mapping from the potential to the density, the Hohenberg-Kohn theorem proves the opposite mapping to hold as well, up to an arbitrary constant. In other words, the potential $v(\mathbf{r})$ is a functional of the ground state density, as well as all other quantities. Along this line, the universal density functional $F[n]$ and total electronic energy functional $E[n]$ can be defined

$$F[n] = \langle \Psi[n] | \hat{T} + \hat{W}_{\mathrm{ee}} | \Psi[n] \rangle \ , \ E[n] = F[n] + \int v_{\mathrm{ne}}(\mathbf{r}) n(\mathbf{r}) \mathrm{d}\mathbf{r} \tag{5.5}$$

as well as a variational principle expressing the ground state energy

$$E_0 = \min_n \left\{ F[n] + \int v_{\mathrm{ne}}(\mathbf{r}) n(\mathbf{r}) \mathrm{d}\mathbf{r} \right\} \tag{5.6}$$

with n being N-electron densities. The ground state energy is found for a ground state density $n_0(\mathbf{r})$ associated with $v_{ne}(\mathbf{r})$. Defining a non-interacting kinetic-energy functional

$$T_{\mathrm{s}}[n] = \min_{\Phi \to n} \langle \Phi | \hat{T} | \Phi \rangle = \langle \Phi[n] | \hat{T} | \Phi[n] \rangle \tag{5.7}$$

where the minimization regards normalized single-determinant wave functions[1] Φ yielding a fixed density n, Kohn and Sham proposed the following decomposition

$$F[n] = T_s[n] + E_{Hxc}[n] \tag{5.8}$$

where $E_{Hxc}[n]$ is the Hartree exchange-correlation functional, therefore leading to

$$\begin{aligned} E_0 &= \min_n \left\{ F[n] + \int v_{\mathrm{ne}}(\mathbf{r}) n(\mathbf{r}) \mathrm{d}\mathbf{r} \right\} \\ &= \min_n \left\{ \min_{\Phi \to n} \langle \Phi | \hat{T} | \Phi \rangle + E_{\mathrm{Hxc}}[n] + \int v_{\mathrm{ne}}(\mathbf{r}) n(\mathbf{r}) \mathrm{d}\mathbf{r} \right\} \\ &= \min_n \min_{\Phi \to n} \left\{ \langle \Phi | \hat{T} + \hat{V}_{\mathrm{ne}} | \Phi \rangle + E_{\mathrm{Hxc}} \left[n_\Phi \right] \right\} \\ &= \min_\Phi \left\{ \langle \Phi | \hat{T} + \hat{V}_{\mathrm{ne}} | \Phi \rangle + E_{\mathrm{Hxc}} \left[n_\Phi \right] \right\} \end{aligned} \tag{5.9}$$

Focusing now on $E_{Hxc}[n]$, it can be decomposed as the sum of two terms

$$E_{Hxc}[n] = E_H[n] + E_{xc}[n] \tag{5.10}$$

the first term being the Hartree energy functional, associated with the electrostatic repulsion

$$E_{\mathrm{H}}[n] = \frac{1}{2} \iint \frac{n\left(\mathbf{r}_1\right) n\left(\mathbf{r}_2\right)}{\left|\mathbf{r}_1 - \mathbf{r}_2\right|} \mathrm{d}\mathbf{r}_1 \, \mathrm{d}\mathbf{r}_2 \tag{5.11}$$

[1] $\Phi[n]$ takes the name of KS wave function.

and the second term being the exchange-correlation energy functional, as it is composed of and exchange $E_x[n]$ and correlation $E_c[n]$ terms,

$$E_x[n] = \langle \Phi[n] | \hat{W}_{ee} | \Phi[n] \rangle - E_H[n] \tag{5.12}$$

$$E_{\mathrm{c}}[n] = \langle \Psi[n] | \hat{T} + \hat{W}_{\mathrm{ee}} | \Psi[n] \rangle - \langle \Phi[n] | \hat{T} + \hat{W}_{\mathrm{ee}} | \Phi[n] \rangle = T_{\mathrm{c}}[n] + U_{\mathrm{c}}[n] \tag{5.13}$$

Coming back now to Eqn.5.9, being Φ composed by a set of N orthonormal occupied spin-orbitals $\phi_i(\mathbf{r})\delta_{\sigma_i,\sigma}$ such that $n(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2$, the total energy can be rewritten as

$$E\left[\{\phi_i\}\right] = \sum_{i=1}^{N} \int \phi_i^*(\mathbf{r}) \left(-\frac{1}{2}\nabla^2 + v_{\mathrm{ne}}(\mathbf{r}) \right) \phi_i(\mathbf{r}) \mathrm{d}\mathbf{r} + E_{\mathrm{Hxc}}[n] \tag{5.14}$$

and, for its minimization over the set of orthonormal orbitals, a Lagrangian is introduced

$$\mathcal{L}\left[\{\phi_i\}\right] = E\left[\{\phi_i\}\right] - \sum_{i=1}^{N} \varepsilon_i \left(\int \phi_i^*(\mathbf{r}) \phi_i(\mathbf{r}) \mathrm{d}\mathbf{r} - 1 \right) \tag{5.15}$$

with ϵ_i the multiplier for the orthonormalization constraint. Requiring the stationary condition one obtains

$$0 = \frac{\delta \mathcal{L}}{\delta \phi_i^*(\mathbf{r})} = \left(-\frac{1}{2}\nabla^2 + v_{\mathrm{ne}}(\mathbf{r}) \right) \phi_i(\mathbf{r}) + \frac{\delta E_{\mathrm{Hxc}}[n]}{\delta \phi_i^*(\mathbf{r})} - \varepsilon_i \phi_i(\mathbf{r}) \tag{5.16}$$

Focusing on the derivative of the Hartree exchange-correlation energy,

$$\begin{aligned} \frac{\delta E_{\mathrm{Hxc}}[n]}{\delta \phi_i^*(\mathbf{r})} &= \int \frac{\delta E_{\mathrm{Hxc}}[n]}{\delta n\left(\mathbf{r}'\right)} \frac{\delta n\left(\mathbf{r}'\right)}{\delta \phi_i^*(\mathbf{r})} \mathrm{d}\mathbf{r}' = \int \frac{\delta E_{\mathrm{Hxc}}[n]}{\delta n\left(\mathbf{r}'\right)} \phi_i(\mathbf{r}) \delta(\mathbf{r} - \mathbf{r}') \mathrm{d}\mathbf{r}' \\ &= \frac{\delta E_{\mathrm{Hxc}}[n]}{\delta n(\mathbf{r})}(\mathbf{r}) \phi_i(\mathbf{r}) = v_{\mathrm{Hxc}}(\mathbf{r}) \phi_i(\mathbf{r}) \end{aligned} \tag{5.17}$$

where v_{Hxc} is the Hartree exchange-correlation potential. From the last result and Eqn.5.16, the KS equations are found, in which ϕ_i are called KS orbitales and ϵ_i the KS energies

$$\begin{aligned} \left(-\frac{1}{2}\nabla^2 + v_{\mathrm{ne}}(\mathbf{r}) + v_{\mathrm{Hxc}}(\mathbf{r}) \right) \phi_i(\mathbf{r}) &= \varepsilon_i \phi_i(\mathbf{r}) \\ \left(-\frac{1}{2}\nabla^2 + v_{\mathrm{ne}}(\mathbf{r}) + v_{\mathrm{H}}(\mathbf{r}) + v_{\mathrm{xc}}(\mathbf{r}) \right) \phi_i(\mathbf{r}) &= \varepsilon_i \phi_i(\mathbf{r}) \\ \left(-\frac{1}{2}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r}) \right) \phi_i(\mathbf{r}) &= \varepsilon_i \phi_i(\mathbf{r}) \\ \hat{H}_{KS}\ \phi_i(\mathbf{r}) &= \varepsilon_i \phi_i(\mathbf{r}) \end{aligned} \tag{5.18}$$

In the previous equations, the Hartree and the exchange-correlation potentials appear

$$v_H(\mathbf{r}) = \frac{\delta E_{\mathrm{H}}[n]}{\delta n(\mathbf{r})} = \int \frac{n\left(\mathbf{r}'\right)}{|\mathbf{r} - \mathbf{r}'|} \mathrm{d}\mathbf{r}' \ , \ v_{xc}(\mathbf{r}) = \frac{\delta E_{\mathrm{xc}}[n]}{\delta n(\mathbf{r})} \tag{5.19}$$

The problem now is that the exact form of the exchange-correlation energy is unknown. This energy represents all terms beyond the single-particle kinetic energy $T_c[n]$ and E_H, namely the electron-electron interactions beyond the Hartree approximation. The latter, in absence of the $v_{xc}(\mathbf{r})$ in 5.18, defines the so called Hartree equations and mean-field approximation, in which each electron moves in the average field created by all other electrons, plus the Coulomb attraction exerted by the nucleus. Considering back instead the exchange-correlation potential, one could consider the different contributions as part of an effective potential $v_{eff}(\mathbf{r}) = v_{\mathrm{ne}}(\mathbf{r}) + v_{\mathrm{H}}(\mathbf{r}) + v_{\mathrm{xc}}(\mathbf{r})$, and therefore an energy functional $E_{KS} = T_s[n] + E_{eff}[n]$. The advantage of the KS strategy is that the ground state energy $n_0(\mathbf{r})$, associated to such one-electron problem, is the same as in the real interacting system. The exchange-correlation term represents the bridge from the true n-electron system to the fictitious one-electron one, and its preparation is essential. Even though it usually represents less than 10% of the total energy [119], the quality of a DFT calculation depends on the quality of its approximation. For a description of the available approximations of this term, the reader is invited to follow the rich literature on the topic. Here the author limits the discussion to show one popular example, the Generalized Gradient Approximation (GGA), able to capture inhomogeneities in the electron density with the inclusion of the density gradient:

$$E_{xc}^{GGA}[\rho(\boldsymbol{r})] = \int \rho(\boldsymbol{r})\varepsilon_{xc}^{GGA}[\rho(\boldsymbol{r}), \nabla\rho(\boldsymbol{r})]d\boldsymbol{r} \tag{5.20}$$

being ε_{xc}^{GGA} the exchange-correlation energy density at the GGA level.

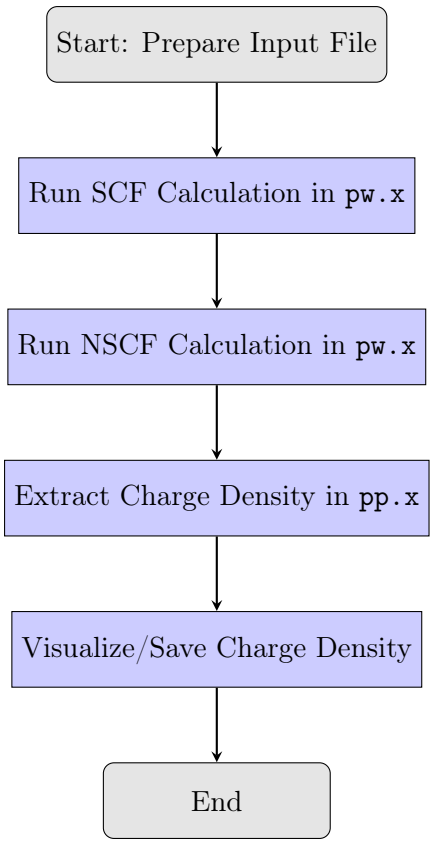
The Hartree equations, just as the KS ones, must be solved *self-consistently*, because the potential depends on the orbitals via the electron density distribution, which in turn depends on the solution of the secolar problem. Self-consistency in principle will be reached when a set of KS orbitals leads to an hamiltonian whose solutions are the same orbitals. In practice an iterative diagonalization of the hamiltonian takes place, following the scheme:

1. A superposition of the atomic electron densities in their isolated states is used (as instructed by a pseudopotential) to build an initial guess for the electron density of the entire system;
2. The hamiltonian terms are calculated;
3. The set of KS equations is diagonalized to obtain new KS orbitals and the associated electron density;
4. Mix the initial and current density and repeat the steps for iterative refinement of the hamiltonian terms and of the density;
5. Check whether the energy difference falls within a stopping criterion for self-consistency, achieving the ground state energy and density.

5.2 ECDs: methods

In Sec.(2.1) and Sec.(2.2), the Manuscripts "*Alloy informatics through ab initio charge density profiles: Case study of hydrogen effects in face-centred cubic crystals*" and "*Transfer Learning in Materials Informatics: structure-property relationships through minimal but highly informative multimodal input*" are introduced, together with their methods and results. Both papers have in common the use of data-driven methods on electron charge densities of bulk crystals. In this section, a closer look will be given to computational methods that can be used to obtain electron charge densities and the related charge properties from quantum simulations.

Within the rich landscape of DFT simulation suits, QUANTUM ESPRESSO (QE) [88–90] is certainly one of the most popular. The diagram in Fig.(5.1) explains the needed steps in order to obtain the electron charge density of a bulk crystal in QE.




```
cat > $name.supercell.
    pp_diff.in << here

&inputpp
outdir='./tmp_noH_$name',
prefix='$name',
plot_num=9
/
&plot
iflag=3,
output_format=5,
fileout='$name.
    supercell_charge_diff.xsf
    ',
nx=300,
ny=300,
nz=400,
/

here
```

Figure 5.1: Workflow for extracting electron charge density of a unit cell in QE, with corresponding example of `pp.x` input file creation commands in a high-throughput framework.

In this procedure, some attention should be paid to details of each step depending on what is the aim of the calculation. In particular, it is important to remember the need of a finer k-mesh in the `nscf` step, so to allow a better sampling of the electron charge density. Also, during extraction of the charge density with `pp.x`, is important to decide whether the interest is on the total or differential field: the latter represents the charge density minus the superposition of atomic densities, and in this sense stands as a preferable choice when the aim is to evaluate

interactions. Fig.(5.1) shows `pp.x` input file creation commands in a high-throughput framework, in which the `plot_num=9` flag requests for the differential field. Once the electron charge densities are acquired, many post-processing options are available depending on the desired analysis.

In the work of Sec.(2.1), differential electron charge densities of hydrogen defect bulk crystals are extracted, both in the presence and absence of the defect, with the aim of performing a difference among the two resulting fields and highlighting in this way the defect-induced perturbation. Substraction of fields is a procedure which requires special attention to regridding strategies, so to ensure equal dimensions and correct operations. Beyond operations between fields, the extracted fields can be used to acquire knowledge on atomic charges. Bader analysis [121] is a common approach, solving a long standing problem of how to *partition* electrons among the constitutive elements of a system, atoms or molecules. The solution of R. F. W. Bader was to identify regions bounded by surfaces running through charge density minima, and estimating the atomic charge as the integral of the electronic density inside the region where the atom is found and, if present, from nearby atom-free regions [122]. From a computational point of view, the code from G. Henkelman group [122] allows for an efficient application of Bader decomposition of charge densities. In order to apply this analysis the total electron charge densities are needed, which can be obtained in QE with the `pp.x` flag `plot_num=0`.

5.3 Machine Learning

5.3.1 Unsupervised Learning

Unsupervised learning techniques help us quantitatively analyze the existence of hidden patterns in the data. Let us suppose the example of a database $\mathbf{D}$ with $f > 3$ features for N samples. The visualization of how the samples distribute in the features space is limited to three-dimensions only, implying the need of a recursive feature selection. In the example of Materials Informatics applications, we could imagine these features to be atomic, structural or even electronic properties of the system of interest. In this sense, databases of real case applications can be very extensive, and include properties among which relations are well known, and properties among which relations are a potentially meaningful and useful discovery. In order to efficiently navigate the complex multi-dimensional space of features holding the samples, it is often useful to rely to unsupervised learning techniques including dimensionality reduction and clustering, highlighting the most important features and their correlations, as well as their driving on the samples distribution and aggregation.

Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction method, allowing to transform the dataset into a set of uncorrelated variables, called principal components (PCs), retaining as much variance from the data as possible [123]. Each eigenvector of the database $N \times N$ covariance matrix $\mathbf{C}$ indicates the direction of a principal component, and its corresponding eigenvalue quantifies the variance captured along that direction in the multi-variate space. The eigenvalue problem for PCA can be expressed as:

$$\mathbf{Cw} = \lambda \mathbf{w}, \tag{5.21}$$

where $\mathbf{w}$ represents the eigenvector and λ the eigenvalue. The previous equation can be found starting from the definition of a linear combination of the original dataset features and its variance, namely

$$\sum_{i=1}^{f} v_i \mathbf{d}_i = \mathbf{Dv} \quad , \quad var(\mathbf{Dv}) = \mathbf{v}^T \mathbf{Cv} \tag{5.22}$$

, with $\mathbf{v}$ being a vector of constants, and requiring a maximization of the variance as in

$$max\Big[\mathbf{v}^T \mathbf{Cv} - \lambda(\mathbf{v}^T \mathbf{v} - 1)\Big] \tag{5.23}$$

With a hierarchy of directions that maximize variance, one can now inspect the data distribution along the main axes, and reconstruct in this reduced space the influence of the original features by looking at *loadings*, the covariances between the PCs and the original dataset variables.

Clustering

Clustering methods help in studying and visualizing aggregation patterns in the data basing on their vicinity (implying similarity) in the features space. Different algorithms are available to solve the clustering problem [124–127], but we here focus on the approximate solution of S. Lloyd [127], also known as *K-Means*. The aim is to minimize the function

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{5.24}$$

representing the sum, over all K clusters, of squared distances between each data point x assigned to a cluster C_i and its centroid μ_i. The solution implemented in automatic packages like `scikit-learn` is based on:

- A set of random arbitrary centers $\mu_1^{(t=0)} \dots \mu_k^{(t=0)}$;
- A cluster C_i is defined as the set of points closer to μ_i than μ_j, for each $j \neq i$;

- The new k-centers (k-means) are $\mu_i^{(t+1)} = \frac{1}{|C_i|} \sum_{x \in C_i} x$;
- Repeat the last two steps until convergence

In many real use cases, the actual number of clusters in which the data is expected to classify is among the targets of study. When no ground truth is available regarding such information, the Elbow analysis [128] can be performed to find the optimal number of clusters, based on the squared distances sum: it is by construction a monotonic decreasing function of the number of clusters, but highlights the optimal number as its decrease step starts to be negligible.

Use case

In the section *Main results of the Thesis - Manuscript I* [4], I present a joint use of the discussed unsupervised learning techniques in Materials Informatics.

5.3.2 Convolutional Neural Networks

This Section aims at delivering a brief, schematic and intuitive presentation of Convolutional Neural Networks (CNNs). The reader is invited to deepen the reading into more detailed presentations which inspired this short one [129–132].

CNNs represent a specific class of deep learning architectures developed for pattern recognition during visual tasks. Starting from the seminal work of Le Cun et al. [133] for handwritten digit recognition, all the well known deep CNNs, e.g. ResNet [134] or VGGNet [135], have been developed basing on the same foundational criteria and components. Among the key constitutents of a CNN are:

- Input Layer: contains the image in the form of a tensor. In the case of a two-dimensional image, the tensor is of order 3, having the last dimension hosting the RGB color channels;
- Convolutional layers: they use the concept of *kernels*, filtering matrices sliding over the input dimensions to extract specific features through convolution operations. Fig. (5.2) represents an example of how a kernel works when applied on an input matrix;

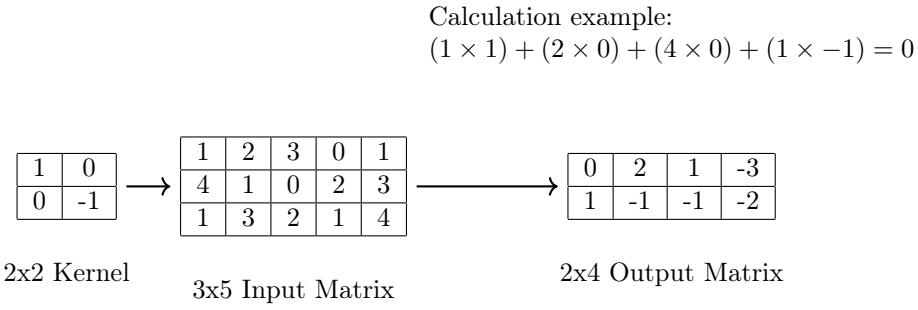


Figure 5.2: Example of convolution with a 2×2 kernel.

- Activation function: increase the non-linearity of the CNN, accordingly with the non-linear mapping represented by the image with respect to the content it represents. An example is the Rectified Linear Unit (ReLU), which truncates the input elements when negative, implying activation only for the input elements which display specific patters in the inspected region;

- Pooling layer: down-samples the input, with either average- or max-value selection within its sub-regions, trying to achieve shift-invariance [131] and reduction of complexity;

- Flattening: the extracted feature maps can be *flattened* to a one-dimensional vector, ready for the last stages of the CNN;

- Fully connected layer: usually at the last layer stage of the architecture, its output requires all elements from the previous layer;

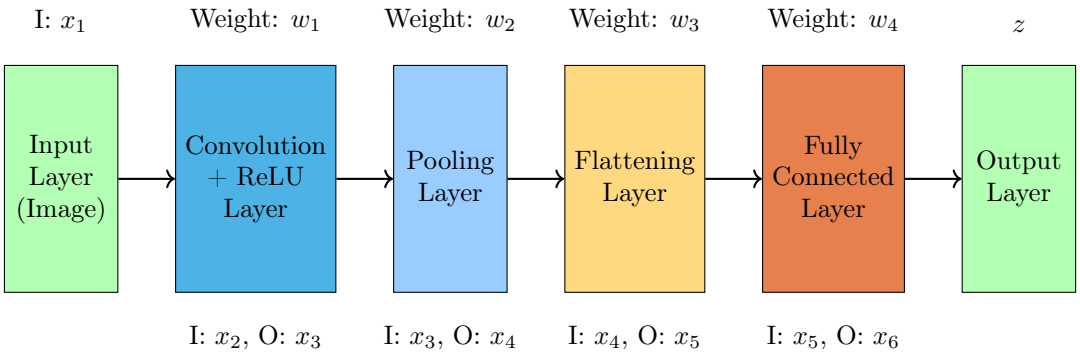


Figure 5.3: The key constituents of a CNN architecture. To each layer, corresponds an input (I), an output (O), and the learnable weights, w_i

In supervised learning, the ground-truth $\mathbf{y_t}$ is used to evaluate the accuracy of the prediction $\mathbf{z}$ by computing a loss function, e.g. :

$$L = \frac{1}{2} \parallel \mathbf{y_t} - \mathbf{z} \parallel^2 \tag{5.25}$$

The weights of the CNN model are optimized to minimize the loss through stochastic gradient descent (SGD):

$$\mathbf{w}_i \leftarrow \mathbf{w}_i - \eta \frac{\partial L}{\partial \mathbf{w}_i} \tag{5.26}$$

being $\mathbf{w}_i$ the weigth tensor of the i-th layer and η the *learning rate*. When all training samples, typically in small batches, have been used to provide the weights a small ($\eta \geq 0$) update along the opposite direction with respect to the loss gradient, the *epoch* is complete.

Use case

In Sec.(2.2), the use of CNNs on electron charge densities for the direct regression of the materials properties is presented. From the three-dimensional charge fields, a two-dimensional slice is extracted from an optimal plane basing on the crystal symmetry, and is used as an image for the algorithm.

5.3.3 Large Language Models

As for the previous Section, far from claiming a complete description of what has distinguished itself as a self-consistent field of study in Computer Science, this Section aims at introducing the reader of this Thesis to Large Language Models (LLMs) in a simple and schematic way, inspired from the literature [136–140].

It is out of doubt that the advent of LLMs [141, 142] has revolutionized natural language processing. Their architectures and the vast amounts of textual training data, enabled them to reach human level conversational capabilities. These models leverage the Transformer architecture [140], known for its efficiency and scalability due to its ability to process data in parallel [138].

Before going into a brief and schematic presentation of the key components of a transformer, it might be useful to consider the training task of LLMs. The core pre-training task for most LLMs is language modeling, where models predict the next token in a sequence. This task helps LLMs learn to generate coherent, contextually appropriate text. The objective during training is to maximize the likelihood of each token based on preceding tokens [138]. The larger architectures, like the latest in the GPT-Series, are also called *foundation* models, and their pre-training procedure would be enough in principle to enable them with conversational capabilities without fine-tuning. However, without proper fine-tuning with supervised instruction datasets to improve reliability and responsiveness to user prompts [137]] they would have limitations, such as difficulty following detailed user instructions and a tendency to generate biased content.

Different transformer architectures are available, and we here limit ourselves to describe the Encoder-Decoder one, advising the reader to explore other types [138]. This model includes both an encoder and a decoder stack. As shown in Fig.(5.4), the encoder generates hidden representations of the input sequence using multi-head self-attention, which enables tokens in a sequence to dynamically interact with one another and is essential for capturing complex relationships across long text. On the extracted representations, the decoder applies cross-attention allowing to auto-regressively produce the output sequence. Since Transformers lack inherent sequence order, positional encodings are added to embeddings, giving the model sequential context. To address training instability, normalization techniques like LayerNorm are applied, helping to

stabilize learning in deep neural architecture.

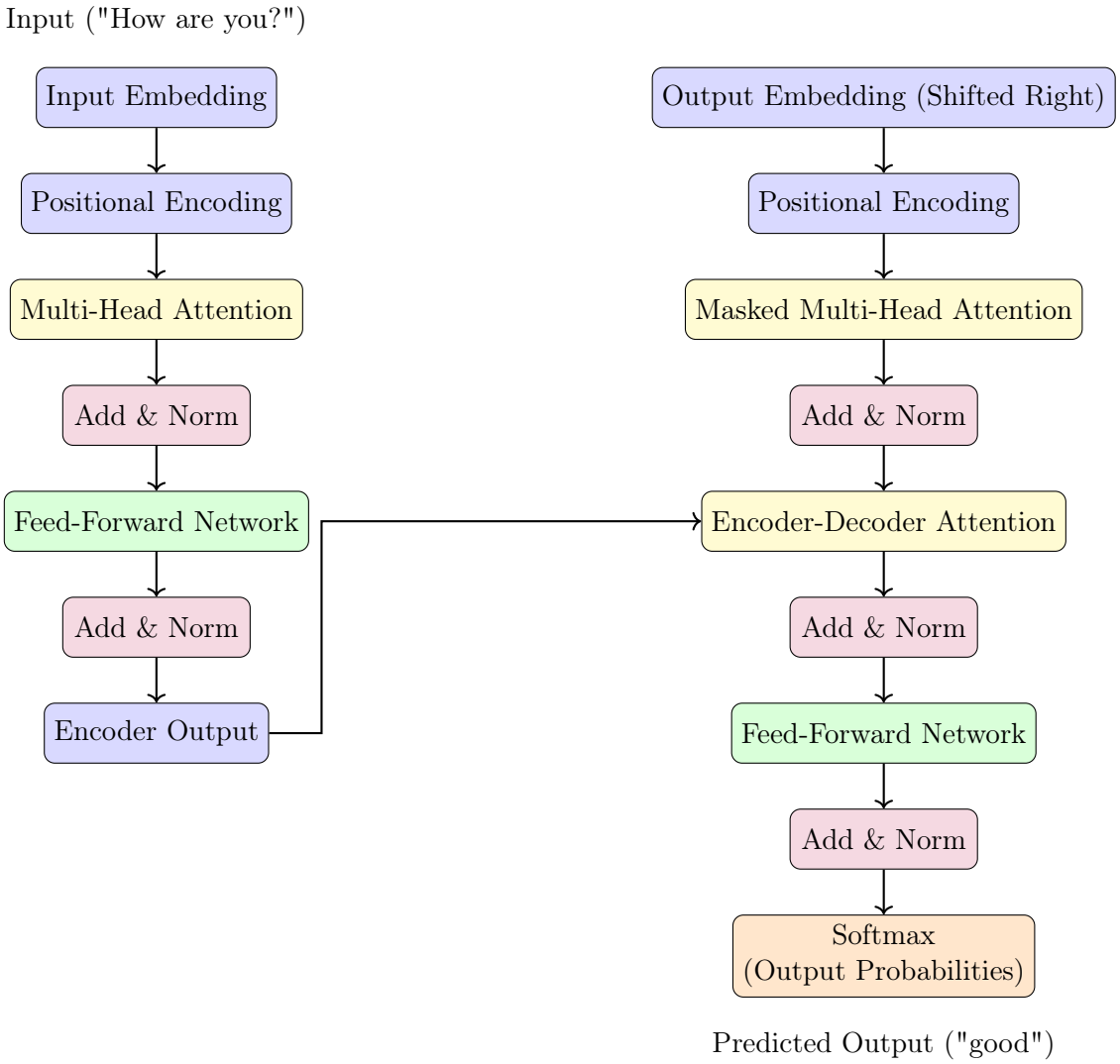


Figure 5.4: The key constituents of an Encoder-Decoder architecture.

Use case

In Sec.(2.2), LLMs are used on crystallographic information of the collected bulk crystals to perform regression of their scalar properties. In particular, their use is discussed also in the context of multi-modal strategies, in which the text embeddings extracted by the language model are combined with the visual embeddings extracted by a CNN.

Bibliography

[1] Amir H. Naghdi, Dario Massa, Kamran Karimi, and Stefanos Papanikolaou. High entropy alloy composition design for mechanical properties. In Dr. Yu Yin, Prof. Han Huang, Dr. Mingxing Zhang, and Prof. Libo Zhou, editors, *High Entropy Alloys - Composition and Microstructure Design*, chapter 1. IntechOpen, Rijeka, 2024.

[2] Dario Massa, Daniel Cieśliński, Amirhossein Naghdi, and Stefanos Papanikolaou. Substitutional alloying using crystal graph neural networks. *AIP Advances*, 14(1), 2024.

[3] Andrea Angeletti, Luca Leoni, Dario Massa, Luca Pasquini, Stefanos Papanikolaou, and Cesare Franchini. Hydrogen diffusion in magnesium using machine learning potentials: a comparative study. (last review stages in npj comp. mat.). 2407.21088 arxiv, 2024.

[4] Dario Massa, Efthimios Kaxiras, and Stefanos Papanikolaou. Alloy informatics through ab initio charge density profiles: Case study of hydrogen effects in face-centred cubic crystals. *Acta Materialia*, 268:119773, 2024.

[5] Dario Massa, Grzegorz Kaszuba, Stefanos Papanikolaou, and Piotr Sankowski. Transfer learning in materials informatics: structure-property relationships through minimal but highly informative multimodal input. (accepted in neurips2024 - ml4ps workshop). 2401.09301 arxiv, 2024.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[7] Ryszard Stanislaw Michalski, Jaime Guillermo Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[8] Stefanos Papanikolaou. Learning local, quenched disorder in plasticity and other crackling noise phenomena. *npj Computational Materials*, 4(1):27, 2018.

[9] Krishna Rajan. Materials informatics. *Materials Today*, 8(10):38–45, 2005.

[10] Logan Ward and Chris Wolverton. Atomistic calculations and materials informatics: A review. *Current Opinion in Solid State and Materials Science*, 21(3):167–176, 2017.

[11] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3(1):54, 2017.

[12] Keisuke Takahashi and Yuzuru Tanaka. Materials informatics: a journey towards material design and synthesis. *Dalton Transactions*, 45(26):10497–10499, 2016.

[13] Seeram Ramakrishna, Tong-Yi Zhang, Wen-Cong Lu, Quan Qian, Jonathan Sze Choong Low, Jeremy Heiarii Ronald Yune, Daren Zong Loong Tan, Stéphane Bressan, Stefano Sanvito, and Surya R Kalidindi. Materials informatics. *Journal of Intelligent Manufacturing*, 30:2307–2326, 2019.

[14] Li-Chiang Lin, Adam H Berger, Richard L Martin, Jihan Kim, Joseph A Swisher, Kuldeep Jariwala, Chris H Rycroft, Abhoyjit S Bhown, Michael W Deem, Maciej Haranczyk, et al. In silico screening of carbon-capture materials. *Nature materials*, 11(7):633–641, 2012.

[15] Christopher E Wilmer, Michael Leaf, Chang Yeon Lee, Omar K Farha, Brad G Hauser, Joseph T Hupp, and Randall Q Snurr. Large-scale screening of hypothetical metal–organic frameworks. *Nature chemistry*, 4(2):83–89, 2012.

[16] Wesley T Hong, Roy E Welsch, and Yang Shao-Horn. Descriptors of oxygen-evolution activity for oxides: a statistical evaluation. *The Journal of Physical Chemistry C*, 120(1):78–86, 2016.

[17] Jeff Greeley, Thomas F Jaramillo, Jacob Bonde, IB Chorkendorff, and Jens K Nørskov. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nature materials*, 5(11):909–913, 2006.

[18] Stefano Curtarolo, Wahyu Setyawan, Gus LW Hart, Michal Jahnatek, Roman V Chepulskii, Richard H Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, et al. Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012.

[19] E. Saal et al. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd). *JOM*, 65(1501), 2013.

[20] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

[21] Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea

Centrone, et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj computational materials*, 6(1):173, 2020.

[22] A. Jain et al. The materials project: A materials genome approach to accelerating materials innovation. *APL Mater*, 1(011002), 2013.

[23] Claudia Draxl and Matthias Scheffler. Nomad: The fair concept for big data-driven materials science. *Mrs Bulletin*, 43(9):676–682, 2018.

[24] CW Andersen, R Armiento, E Blokhin, GJ Conduit, S Dwaraknath, ML Evans, A Fekete, A Gopakumar, S Grazulis, A Merkys, et al. Optimade, an api for exchanging materials data. sci data 8 (1): 217, 2021.

[25] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

[26] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

[27] Andrew S Rosen, Shaelyn M Iyer, Debmalya Ray, Zhenpeng Yao, Alan Aspuru-Guzik, Laura Gagliardi, Justin M Notestein, and Randall Q Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

[28] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications*, 8(1):15679, 2017.

[29] D Xue, PV Balachandran, J Hogden, J Theiler, D Xue, and T Lookman. Accelerated search for materials with targeted properties by adaptive design. *Nat Commun*, 7:11241, 2016.

[30] Atsuto Seko, Atsushi Togo, Hiroyuki Hayashi, Koji Tsuda, Laurent Chaput, and Isao Tanaka. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. *Physical review letters*, 115(20):205901, 2015.

[31] Quan Zhou, Peizhe Tang, Shenxiu Liu, Jinbo Pan, Qimin Yan, and Shou-Cheng Zhang. Learning atoms for materials discovery. *Proceedings of the National Academy of Sciences*, 115(28):E6411–E6417, 2018.

[32] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

[33] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.

[34] Kamal Choudhary, Brian DeCost, and Francesca Tavazza. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Physical review materials*, 2(8):083801, 2018.

[35] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

[36] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B—Condensed Matter and Materials Physics*, 87(18):184115, 2013.

[37] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7), 2011.

[38] Logan Ward, Ruoqian Liu, Amar Krishna, Vinay I Hegde, Ankit Agrawal, Alok Choudhary, and Chris Wolverton. Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations. *Physical Review B*, 96(2):024104, 2017.

[39] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

[40] Z. Wu et al. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn Syst.*, 32:4–24, 2021.

[41] J. Jorgensen et al. Neural message passing with edge updates for predicting properties of molecules and materials. *arXiv.org e-Print archive*, page arXiv:1806.01261, 2018.

[42] K. Schütt et al. Schnet: a continuous-filter convolutional neural network for modeling quantum interactions. *NeurIPS*, 30:991–1001, 2017.

[43] D. K. Duvenaud et al. Convolutional networks on graphs for learning molecular fingerprints. *NeurIPS*, page 2224–2232, 2015.

[44] Z. Wu et al. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530, 2018.

[45] S. Kearnes et al. Molecular graph convolutions: Moving beyond fingerprints. *J. Comput.-Aided Mol. Des.*, 30(595–608), 2016.

[46] C. W. Coley et al. Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Info. Model.*, 57(17571772), 2017.

[47] T. Xie and J. C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, 2018.

[48] T. Xie and J. C. Grossman. Hierarchical visualization of materials space with graph convolutional neural networks. *J. Chem. Phys.*, 149:174111, 2018.

[49] C. Chen et al. Graph net- works as a universal machine learning framework for molecules and crystals. *Chem. Mater.*, 31(9):3564, 2019.

[50] A. Dunn et al. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *Npj Comput. Mater.*, 6:138, 2020.

[51] S.-Y. Louis et al. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.*, 22:18141–18148, 2020.

[52] C. W. Park and C. Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.*, 4:063801, 2020.

[53] W. Gong and Q. Yan. Graph-based deep learning frameworks for molecules and solid-state materials. *Computational Materials Science*, 195:110332, 2021.

[54] P Reiser et al. Graph neural networks for materials science and chemistry. *Communications Materials*, 3:93, 2022.

[55] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

[56] Cheol Woo Park and Chris Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials*, 4(6):063801, 2020.

[57] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

[58] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018.

[59] Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.

[60] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020.

[61] Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential, 2024.

[62] Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C. Lawrence Zitnick, and Zachary W. Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models, 2024.

[63] Yutack Park, Jaesun Kim, Seungwoo Hwang, and Seungwu Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 20(11):4857–4868, May 2024.

[64] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Samuel M. Blau, Vlad Cărare, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord,

Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry, 2024.

[65] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. Chgnet: Pretrained universal neural network potential for charge-informed atomistic modeling, 2023.

[66] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, November 2022.

[67] Cheol Woo Park and Chris Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials*, 4(6):063801, 2020.

[68] Andrew S Rosen, Shaelyn M Iyer, Debmalya Ray, Zhenpeng Yao, Alan Aspuru-Guzik, Laura Gagliardi, Justin M Notestein, and Randall Q Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

[69] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon JL Billinge, et al. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1):59, 2022.

[70] Jens K Nørskov, Frank Abild-Pedersen, Felix Studt, and Thomas Bligaard. Density functional theory in surface chemistry and catalysis. *Proceedings of the National Academy of Sciences*, 108(3):937–943, 2011.

[71] Qiu He, Bin Yu, Zhaohuai Li, and Yan Zhao. Density functional theory for battery materials. *Energy & Environmental Materials*, 2(4):264–279, 2019.

[72] RAP Ribeiro, SR De Lazaro, and SA Pianaro. Density functional theory applied to magnetic materials: Mn3o4 at different hybrid functionals. *Journal of Magnetism and Magnetic Materials*, 391:166–171, 2015.

[73] F Illas, I de PR Moreira, JM Bofill, and M Filatov. Extent and limitations of density-functional theory in describing magnetic systems. *Physical Review B—Condensed Matter and Materials Physics*, 70(13):132414, 2004.

[74] Riaz Hussain, Muhammad Saeed, Muhammad Yasir Mehboob, Saif Ullah Khan, Muhammad Usman Khan, Muhammad Adnan, Mahmood Ahmed, Javed Iqbal, and Khurshid

Ayub. Density functional theory study of palladium cluster adsorption on a graphene support. *RSC advances*, 10(35):20595–20607, 2020.

[75] Sergey A Katsyuba, Elena E Zvereva, Ana Vidiš, and Paul J Dyson. Application of density functional theory and vibrational spectroscopy toward the rational design of ionic liquids. *The Journal of Physical Chemistry A*, 111(2):352–370, 2007.

[76] Kamal Choudhary, Irina Kalish, Ryan Beams, and Francesca Tavazza. High-throughput identification and characterization of two-dimensional materials using density functional theory. *Scientific reports*, 7(1):5179, 2017.

[77] Camilla Pellegrini and Antonio Sanna. Ab initio methods for superconductivity. *Nature Reviews Physics*, 6(8):509–523, 2024.

[78] Yusuf Shaidu and Omololu Akin-Ojo. First principles predictions of superconductivity in doped stanene. *Computational Materials Science*, 118:11–15, 2016.

[79] Divya G Ladha. A review on density functional theory–based study on two-dimensional materials used in batteries. *Materials Today Chemistry*, 11:94–111, 2019.

[80] Valeria Butera. Density functional theory methods applied to homogeneous and heterogeneous catalysis: a short review and a practical user guide. *Physical Chemistry Chemical Physics*, 2024.

[81] David Sibanda, Sunday Temitope Oyinbo, and Tien-Chien Jen. A review of atomic layer deposition modelling and simulation methodologies: Density functional theory and molecular dynamics. *Nanotechnology Reviews*, 11(1):1332–1363, 2022.

[82] Nicola Marzari, Andrea Ferretti, and Chris Wolverton. Electronic-structure methods for materials design. *Nature materials*, 20(6):736–749, 2021.

[83] Johanna I Fuks. Time-dependent density functional theory for charge-transfer dynamics: review of the causes of failure and success. *The European Physical Journal B*, 89:1–19, 2016.

[84] Emilio Artacho, Eduardo Anglada, Oswaldo Diéguez, Julian D Gale, Alberto García, Javier Junquera, Richard M Martin, Pablo Ordejón, José Miguel Pruneda, Daniel Sánchez-Portal, et al. The siesta method; developments and applicability. *Journal of Physics: Condensed Matter*, 20(6):064208, 2008.

[85] Jens Jørgen Mortensen, Ask Hjorth Larsen, Mikael Kuisma, Aleksei V Ivanov, Alireza Taghizadeh, Andrew Peterson, Anubhab Haldar, Asmus Ougaard Dohn, Christian Schäfer,

Elvar Örn Jónsson, et al. Gpaw: An open python package for electronic structure calculations. *The Journal of Chemical Physics*, 160(9), 2024.

[86] Xavier Gonze, Bernard Amadon, Gabriel Antonius, Frédéric Arnardi, Lucas Baguet, Jean-Michel Beuken, Jordan Bieder, François Bottin, Johann Bouchet, Eric Bousquet, et al. The abinit project: Impact, environment and recent developments. *Computer Physics Communications*, 248:107042, 2020.

[87] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications*, 180(11):2175–2196, 2009.

[88] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, Andrea Dal Corso, Stefano de Gironcoli, Stefano Fabris, Guido Fratesi, Ralph Gebauer, Uwe Gerstmann, Christos Gougoussis, Anton Kokalj, Michele Lazzeri, Layla Martin-Samos, Nicola Marzari, Francesco Mauri, Riccardo Mazzarello, Stefano Paolini, Alfredo Pasquarello, Lorenzo Paulatto, Carlo Sbraccia, Sandro Scandolo, Gabriele Sclauzero, Ari P Seitsonen, Alexander Smogunov, Paolo Umari, and Renata M Wentzcovitch. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter*, 21(39):395502 (19pp), 2009.

[89] P Giannozzi, O Andreussi, T Brumme, O Bunau, M Buongiorno Nardelli, M Calandra, R Car, C Cavazzoni, D Ceresoli, M Cococcioni, N Colonna, I Carnimeo, A Dal Corso, S de Gironcoli, P Delugas, R A DiStasio Jr, A Ferretti, A Floris, G Fratesi, G Fugallo, R Gebauer, U Gerstmann, F Giustino, T Gorni, J Jia, M Kawamura, H-Y Ko, A Kokalj, E Küçükbenli, M Lazzeri, M Marsili, N Marzari, F Mauri, N L Nguyen, H-V Nguyen, A Otero de-la Roza, L Paulatto, S Poncé, D Rocca, R Sabatini, B Santra, M Schlipf, A P Seitsonen, A Smogunov, I Timrov, T Thonhauser, P Umari, N Vast, X Wu, and S Baroni. Advanced capabilities for materials modelling with quantum espresso. *Journal of Physics: Condensed Matter*, 29(46):465901, 2017.

[90] Paolo Giannozzi, Oscar Baseggio, Pietro Bonfà, Davide Brunato, Roberto Car, Ivan Carnimeo, Carlo Cavazzoni, Stefano de Gironcoli, Pietro Delugas, Fabrizio Ferrari Ruffino, Andrea Ferretti, Nicola Marzari, Iurii Timrov, Andrea Urru, and Stefano Baroni. Quantum espresso toward the exascale. *The Journal of Chemical Physics*, 152(15):154105, 2020.

[91] Georg Kresse and Jürgen Hafner. Ab initio molecular dynamics for liquid metals. *Physical review B*, 47(1):558, 1993.

[92] Georg Kresse and Jürgen Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational materials science*, 6(1):15–50, 1996.

[93] Georg Kresse and Jurgen Hafner. Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements. *Journal of Physics: Condensed Matter*, 6(40):8245, 1994.

[94] Jimmy-Xuan Shen, Jason M Munro, Matthew K Horton, Patrick Huck, Shyam Dwaraknath, and Kristin A Persson. A representation-independent electronic charge density database for crystalline materials. *Scientific data*, 9(1):661, 2022.

[95] Pol Febrer, Peter Jørgensen, Miguel Pruneda, Alberto Garcia, Pablo Ordejon, and Arghya Bhowmik. Graph2mat: Universal graph to matrix conversion for electron density prediction. 2024.

[96] Peter Bjørn Jørgensen and Arghya Bhowmik. Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids. *npj Computational Materials*, 8(1):183, 2022.

[97] Zhuoran Qiao, Anders S Christensen, Matthew Welborn, Frederick R Manby, Anima Anandkumar, and Thomas F Miller III. Informing geometric deep learning with electronic interactions to accelerate quantum chemistry. *Proceedings of the National Academy of Sciences*, 119(31):e2205221119, 2022.

[98] Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

[99] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

[100] Peter Bjørn Jørgensen and Arghya Bhowmik. QM9 Charge Densities and Energies Calculated with VASP, 2022.

[101] Peter Bjørn Jørgensen and Arghya Bhowmik. NMC Li-ion Battery Cathode Energies and Charge Densities, 2022.

[102] Taoyuze Lv, Zhicheng Zhong, Yuhang Liang, Feng Li, Jun Huang, and Rongkun Zheng. Deep charge: Deep learning model of electron density from a one-shot density functional theory calculation. *Physical Review B*, 108(23):235159, 2023.

[103] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[104] Tsz Wai Ko and Shyue Ping Ong. Recent advances and outstanding challenges for machine learning interatomic potentials. *Nature Computational Science*, 3(12):998–1000, 2023.

[105] Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35:11423–11436, 2022.

[106] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

[107] Janosh Riebesell, Rhys EA Goodall, Anubhav Jain, Philipp Benner, Kristin A Persson, and Alpha A Lee. Matbench discovery–an evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920*, 2023.

[108] Janosh Riebesell, Rhys Goodall, Anubhav Jain, Kristin Persson, et al. Matbench discovery-can machine learning identify stable crystals? In *Workshop on"Machine Learning for Materials"ICLR 2023*.

[109] Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential. *arXiv preprint arXiv:2410.22570*, 2024.

[110] Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.

[111] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

[112] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

[113] Félix Musil, Michael J Willatt, Mikhail A Langovoy, and Michele Ceriotti. Fast and accurate uncertainty estimation in chemical machine learning. *Journal of chemical theory and computation*, 15(2):906–915, 2019.

[114] Maksim Kulichenko, Kipton Barros, Nicholas Lubbers, Ying Wai Li, Richard Messerly, Sergei Tretiak, Justin S Smith, and Benjamin Nebgen. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nature Computational Science*, 3(3):230–239, 2023.

[115] Justin S Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E Roitberg. Less is more: Sampling chemical space with active learning. *The Journal of chemical physics*, 148(24), 2018.

[116] Ryosuke Jinnouchi, Kazutoshi Miwa, Ferenc Karsai, Georg Kresse, and Ryoji Asahi. On-the-fly active learning of interatomic potentials for large-scale atomistic simulations. *The Journal of Physical Chemistry Letters*, 11(17):6946–6955, 2020.

[117] Ryosuke Jinnouchi, Ferenc Karsai, and Georg Kresse. On-the-fly machine learning force field generation: Application to melting points. *Physical Review B*, 100(1):014105, 2019.

[118] Julien Toulouse. Introduction to density-functional theory. *Sorbonne Université and CNRS*, 2019.

[119] June Gunn Lee. *Computational materials science: an introduction*. CRC press, 2016.

[120] Walter Kohn, Axel D Becke, and Robert G Parr. Density functional theory of electronic structure. *The journal of physical chemistry*, 100(31):12974–12980, 1996.

[121] Richard FW Bader. Atoms in molecules. *Accounts of chemical research*, 18(1):9–15, 1985.

[122] Graeme Henkelman, Andri Arnaldsson, and Hannes Jónsson. A fast and robust algorithm for bader decomposition of charge density. *Computational Materials Science*, 36(3):354–360, 2006.

[123] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[124] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114, 1996.

[125] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[126] Daniel Defays. An efficient algorithm for a complete link method. *The computer journal*, 20(4):364–366, 1977.

[127] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[128] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[129] Jianxin Wu. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5(23):495, 2017.

[130] Jake Bouvrie. Notes on convolutional neural networks. 2006.

[131] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.

[132] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[133] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[134] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[135] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[136] Timm Teubner, Christoph M Flath, Christof Weinhardt, Wil van der Aalst, and Oliver Hinz. Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering*, 65(2):95–101, 2023.

[137] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

[138] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[139] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[140] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[141] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[142] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.