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

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 $\bar{\mathbf{v}}_i^e = \frac{1}{N^e}\sum_{k=1}^{N_i^e}\left\{\mathbf{e}'_k\right\}_{r_k=i}$, 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 $\bar{\mathbf{u}}^e = \frac{1}{N^e}\sum_{k=1}^{N^e}\left\{\mathbf{e}'_k\right\}$ and $\bar{\mathbf{u}}^v = \frac{1}{N^v}\sum_{i=1}^{N^v}\left\{\mathbf{v}'_i\right\}$.

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

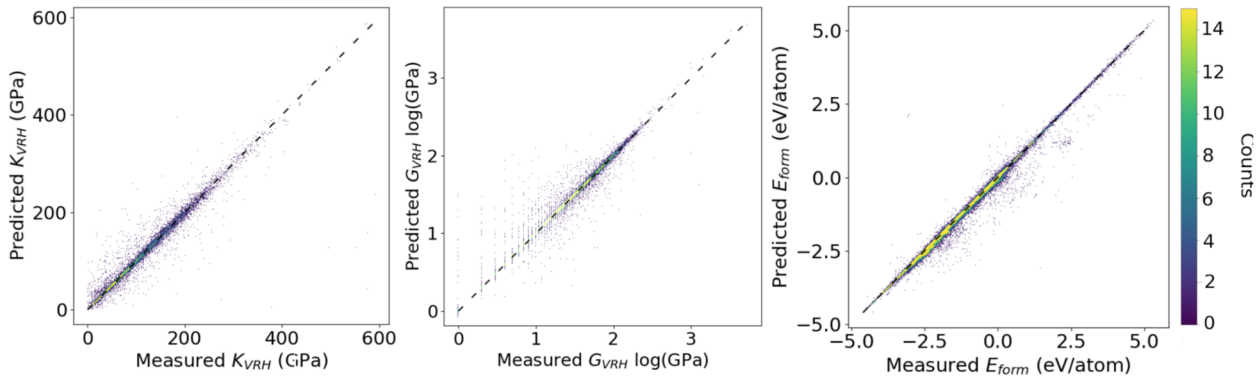


FIG. 1. Parity plots for the pre-trained model on the MP dataset. The plots involve the predictions on (a) the bulk modulus (K_{VRH}), (b) the shear modulus (G_{VRH}), and (c) the formation energy (E_{form}).

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:

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

where x is the atomic fraction of substitutional defects, $E(\mathrm{M}_{1-x}\mathrm{D}_x)$ 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

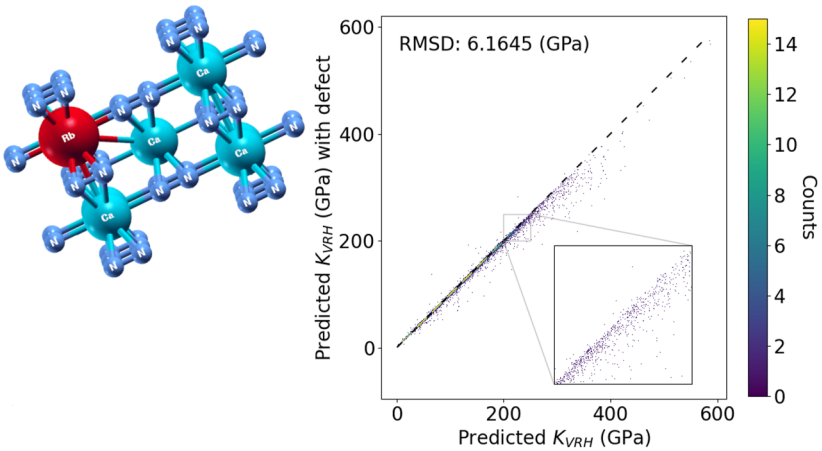


FIG. 2. Predicted bulk moduli in the Rb-defected MP crystals with respect to their prediction in non-defected ones (b). An example structure from the MP dataset is shown, CaN_2, in which one atom of Ca has been replaced with a Rb atom (a). Here, only the case of Rb-defected systems is shown due to its 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

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

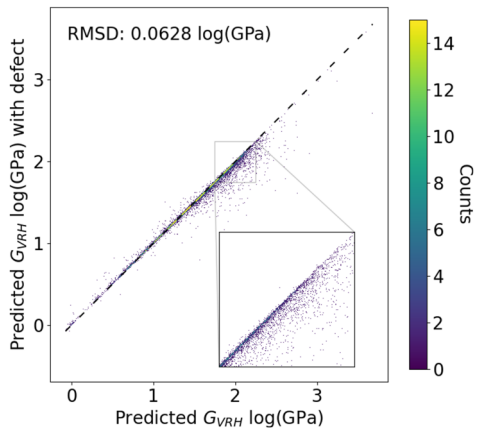


FIG. 3. Predicted shear modulus in the Rb-defected MP crystals with respect to the prediction in non-defected ones. Here, only the case of Rb-defected systems is shown due to its largest RMSD among the set of considered defects. Similar plots for the Mn- and H-defected systems are shown in the Appendix.

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

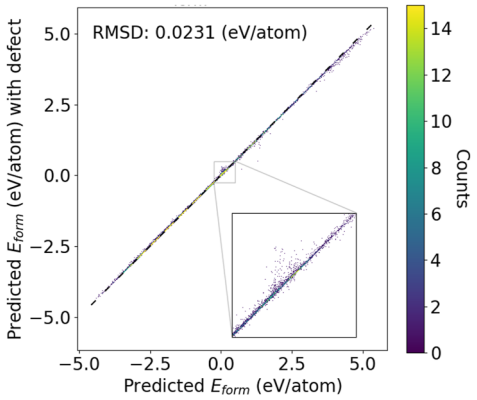


FIG. 4. Predicted formation energy in the H-defected MP crystals with respect to the prediction in non-defected ones. Here, only the case of H-defected systems is shown due to its largest RMSD among the set of considered defects. Similar plots for the Rb- and Mn-defected systems are shown in the Appendix.

deviations are all in the order of 0.02 eV/atom and are progressively less prominent going from H to Rb and Mn.

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

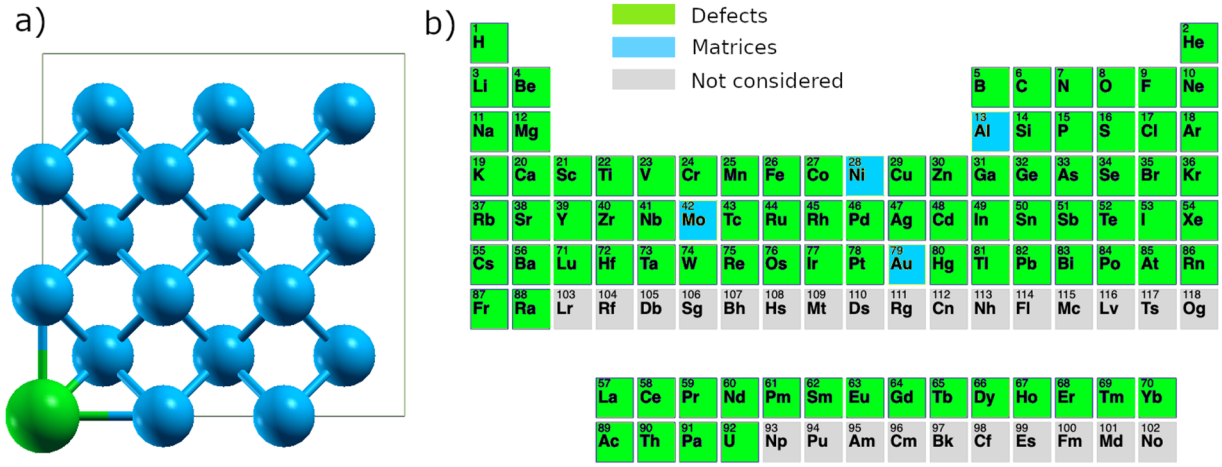


FIG. 5. Periodic Table plot (b) explaining the process of selecting a set of host matrices (in light blue) and substitutional defects (in green). To a given selected host matrix, the non-selected ones represent defects too. The host matrices are 3 × 3 × 3 supercells of the highlighted species (a).

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

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

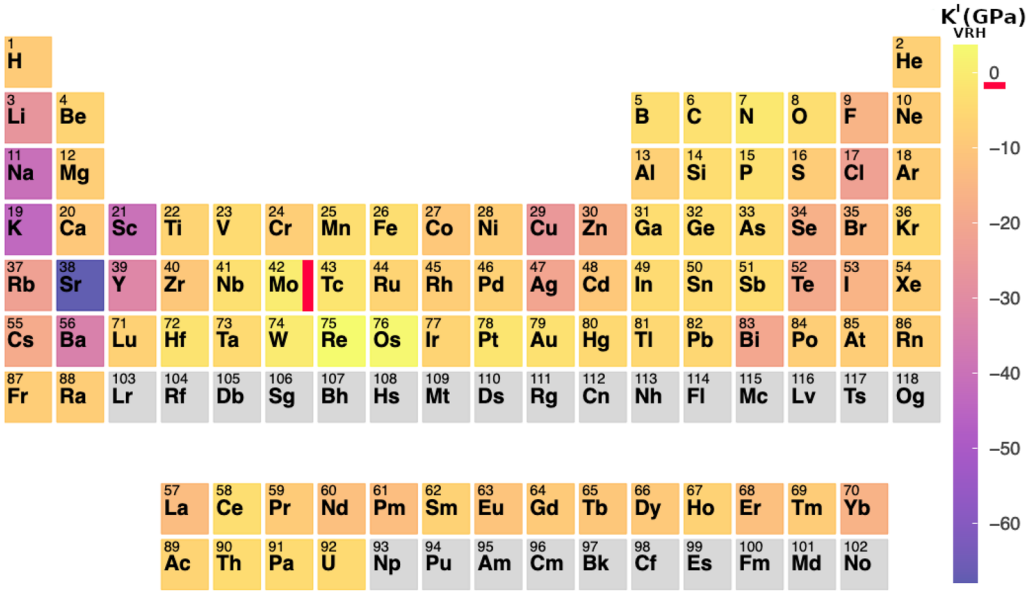


FIG. 6. Predicted bulk modulus variation (K'_{VRH}) for a single-atom substitutionally defected Mo supercell with respect to the undefected one, for each possible defect atomic species from the provided Periodic Table. Similar plots for Al, Au, and Ni supercells are provided in the Appendix. The red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

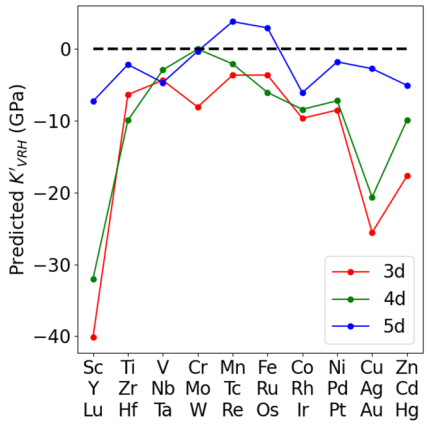


FIG. 7. Predicted bulk modulus variation (K'_{VRH}) for a single-atom substitutionally defected Mo supercell with respect to the undefected one along the 3d, 4d, and 5d series of the Periodic Table. The black dashed line highlights the pure Mo case.

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

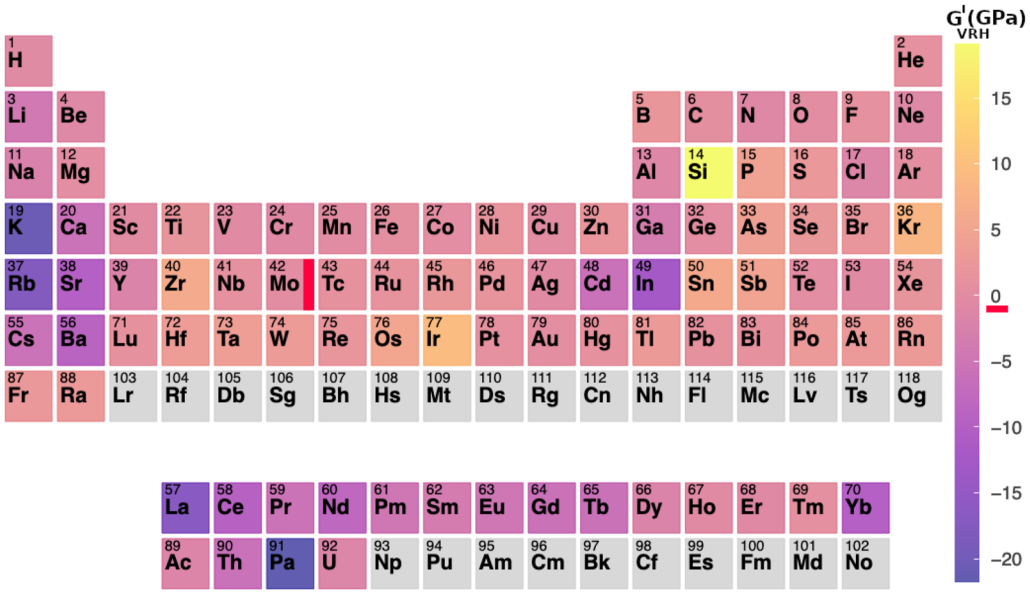


FIG. 8. Predicted shear modulus variation (G'_{VRH}) for a single-atom substitutionally defected Mo supercell with respect to the undefected one for each possible defect atomic species from the provided Periodic Table. Similar plots for Al, Au, and Ni supercells are provided in the Appendix. The red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

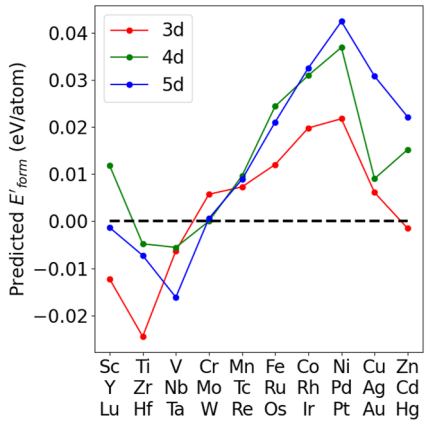


FIG. 10. Predicted formation energy variation (E'_{form}) for a single-atom substitutionally defected Mo supercell with respect to the undefected one along the 3d, 4d, and 5d series of the Periodic Table. The black dashed line highlights the pure Mo case.

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

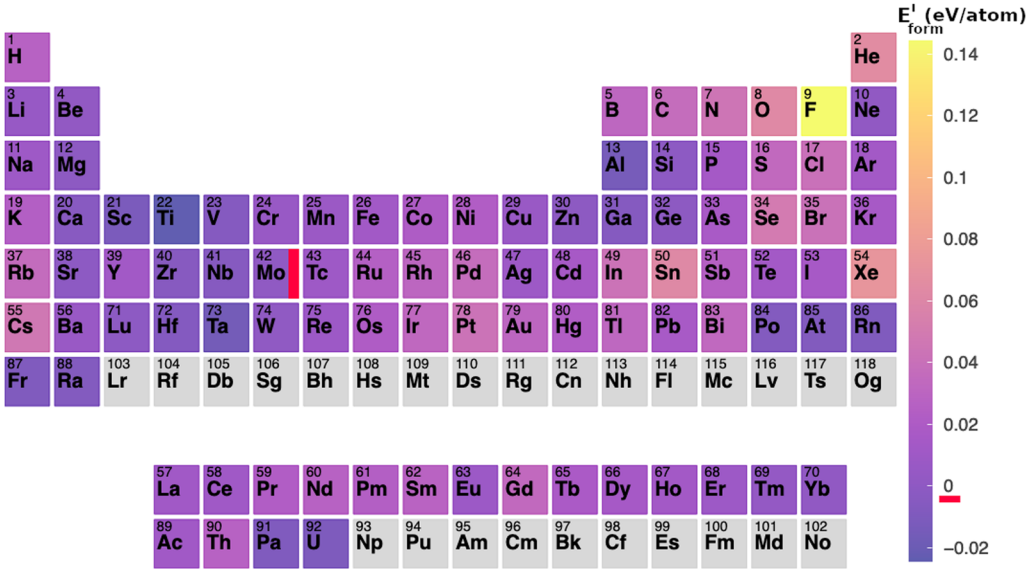


FIG. 9. Predicted formation energy variation (E'_{form}) for a single-atom substitutionally defected Mo supercell with respect to the undefected one for each possible defect atomic species from the provided Periodic Table. Similar plots for Al, Au, and Ni supercells are provided in the Appendix. The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

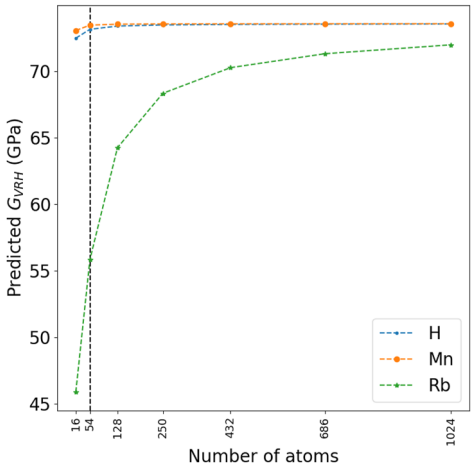


FIG. 12. Saturation plot of a Mo host matrix G_{VRH} when substitutionally defected with H, Mn, or Rb for different supercell sizes.

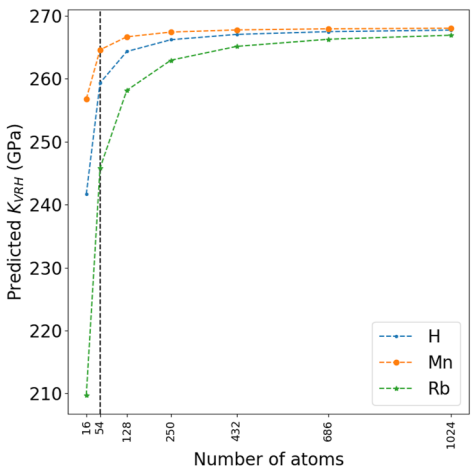


FIG. 11. Saturation plot of a Mo host matrix K_{VRH} when substitutionally defected with H, Mn, or Rb for different supercell sizes.

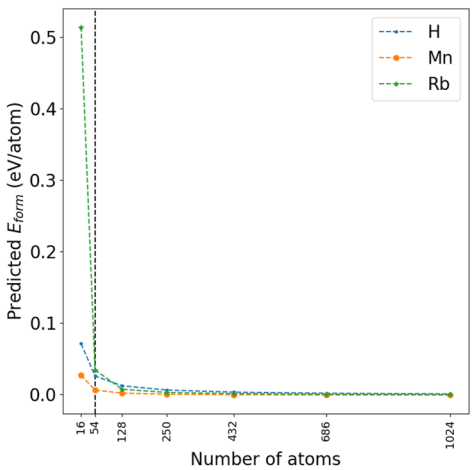


FIG. 13. Saturation plot of a Mo host matrix E_{form} when substitutionally defected with H, Mn, or Rb for different supercell sizes.

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,

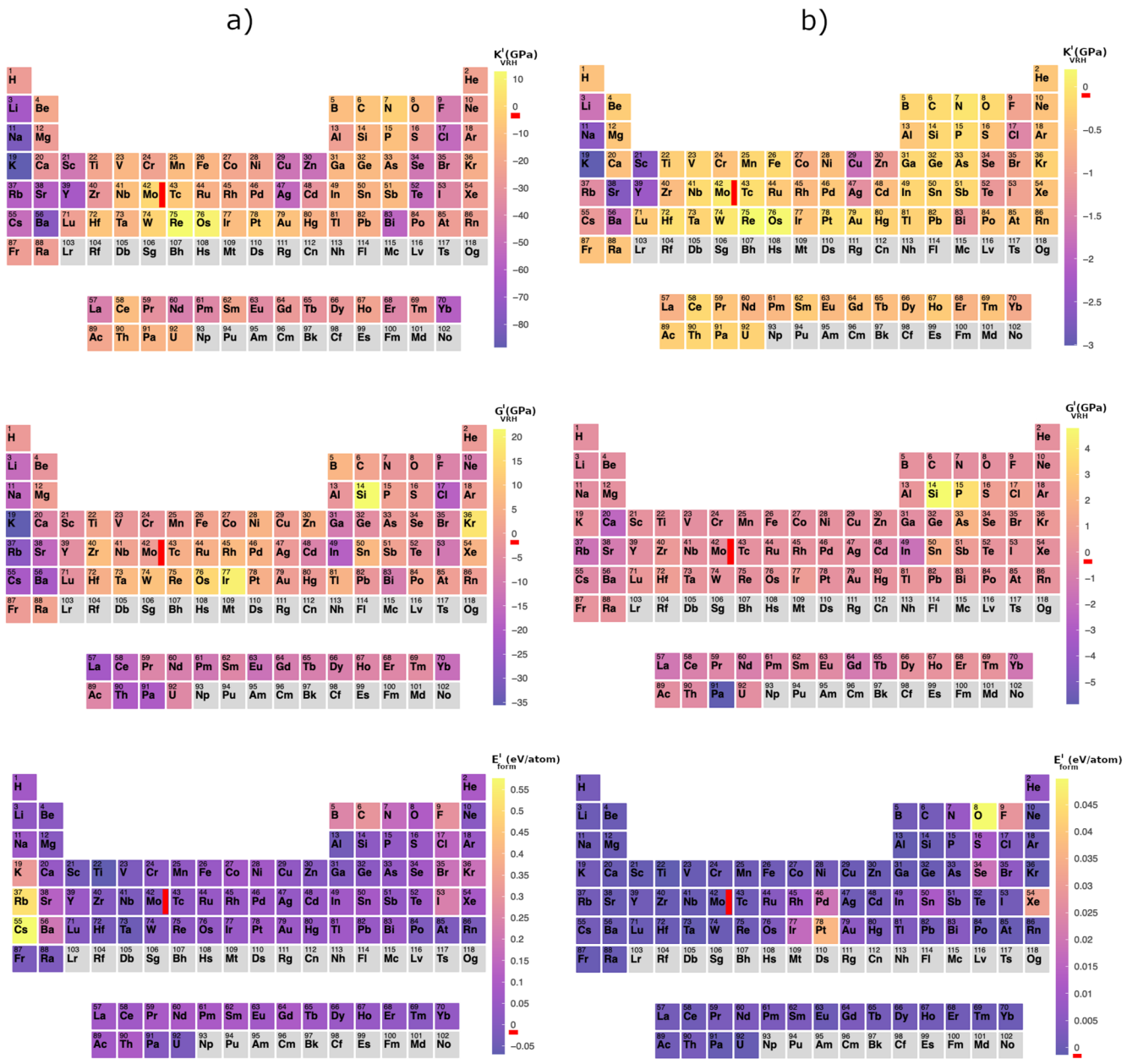
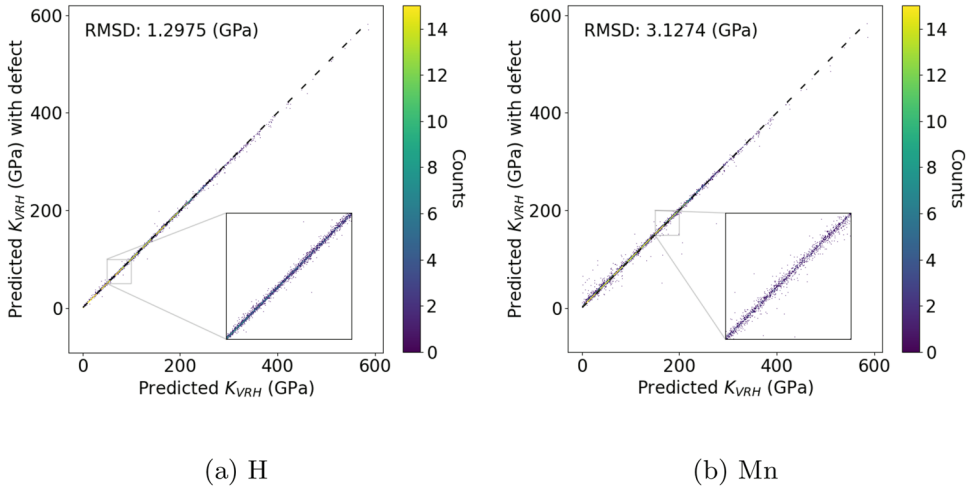


FIG. 14. Predicted variations K'_{VRH} [(a) and (d)], G'_{VRH} [(b) and (e)], and E'_{form} [(c) and (f)] for a single-atom substitutionally defected Mo supercell with respect to the undefected one for each possible defect atomic species from the provided Periodic Table. In plots (a)–(c), a $2 \times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

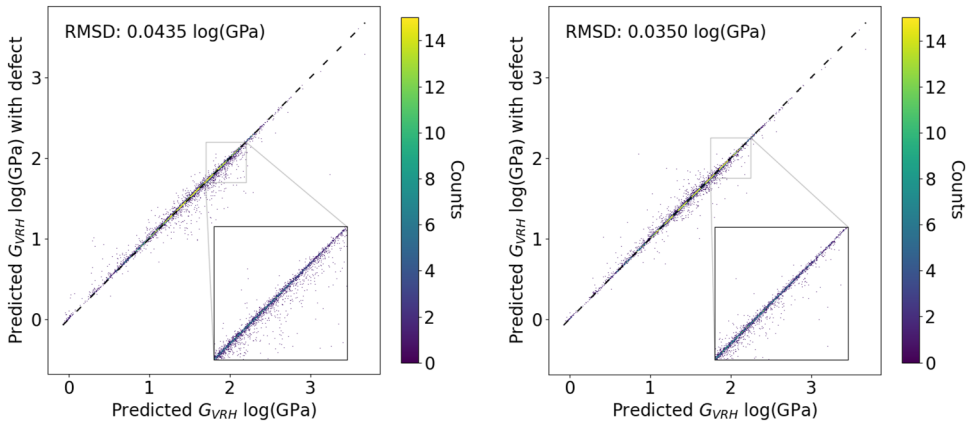
IV. CONCLUSIONS

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



(a) H (b) Mn

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



(a) H (b) Mn

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.

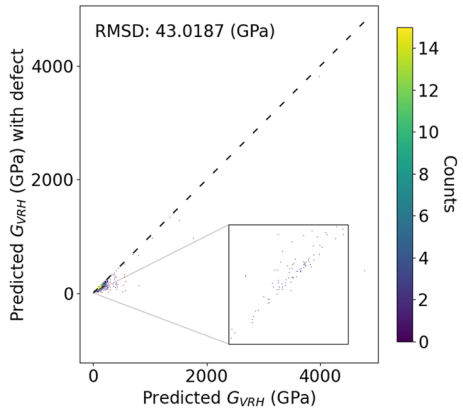


FIG. 18. Predicted formation energy in the Rb-defected MP crystals with respect to its prediction in non-defected ones. The plot has the aim to display the need of a log–log scale to appreciate the deviations.

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

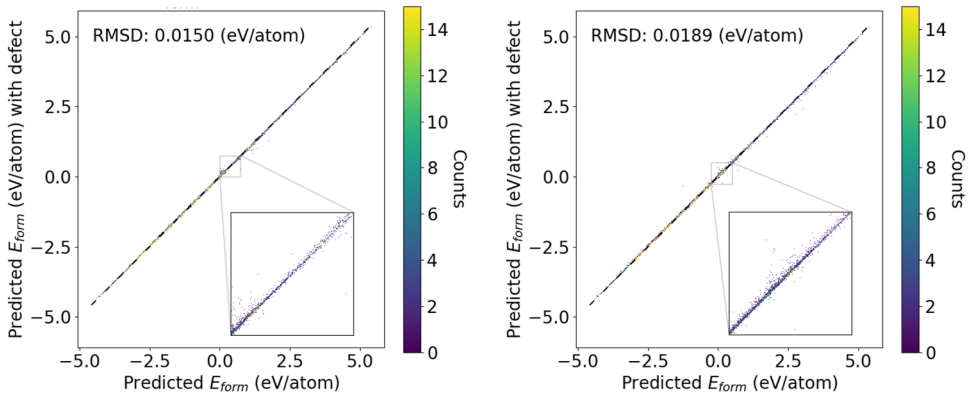
The authors have no conflicts to disclose.

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,

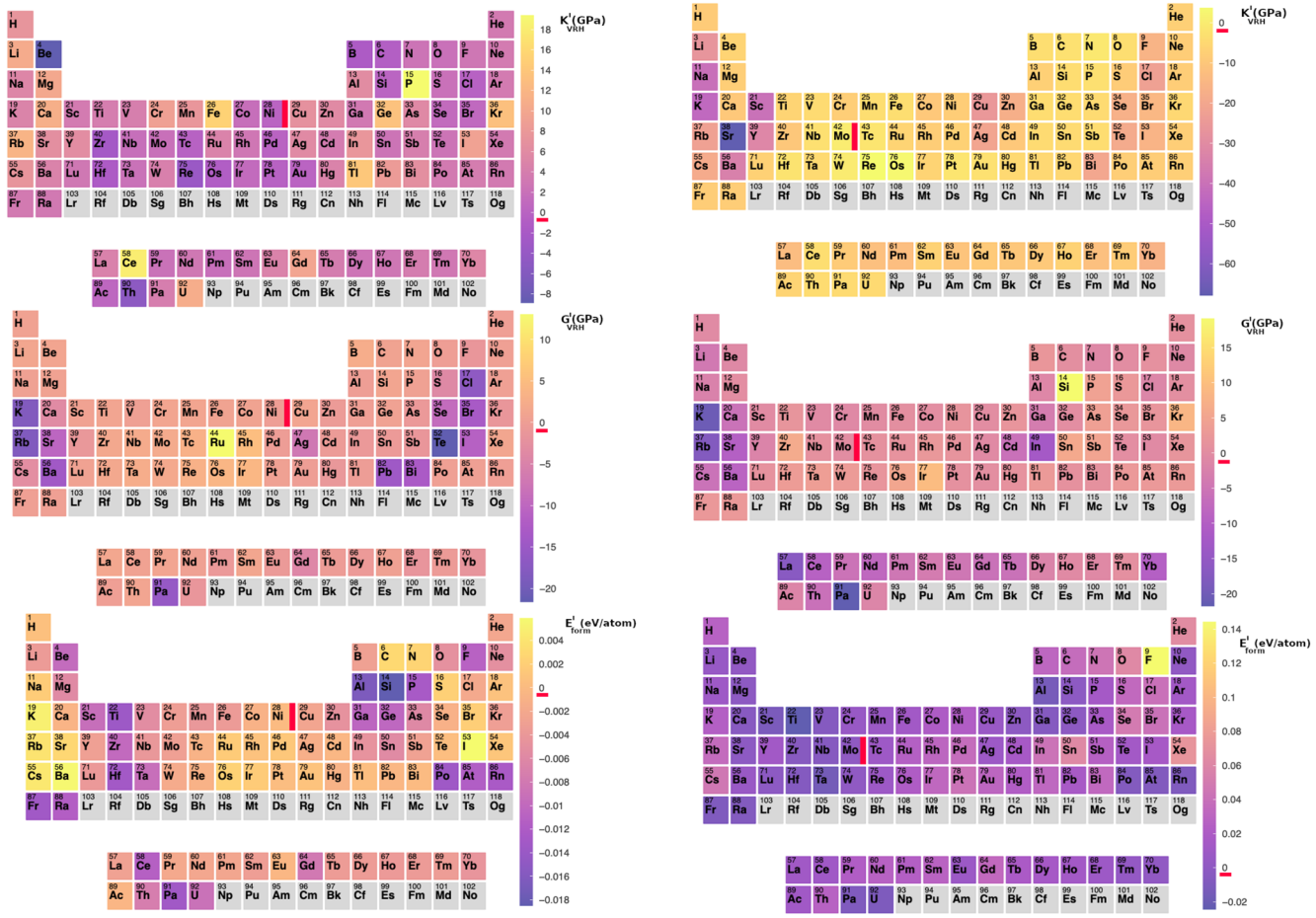


FIG. 19. Comparison between the Periodic Table plots of the predicted bulk modulus, shear modulus, and formation energy for a single-atom substitutionally defected Ni [(a)–(c)] and Mo [(d)–(f)] supercell with respect to the undefected one for each possible atomic species from the provided Periodic Table.

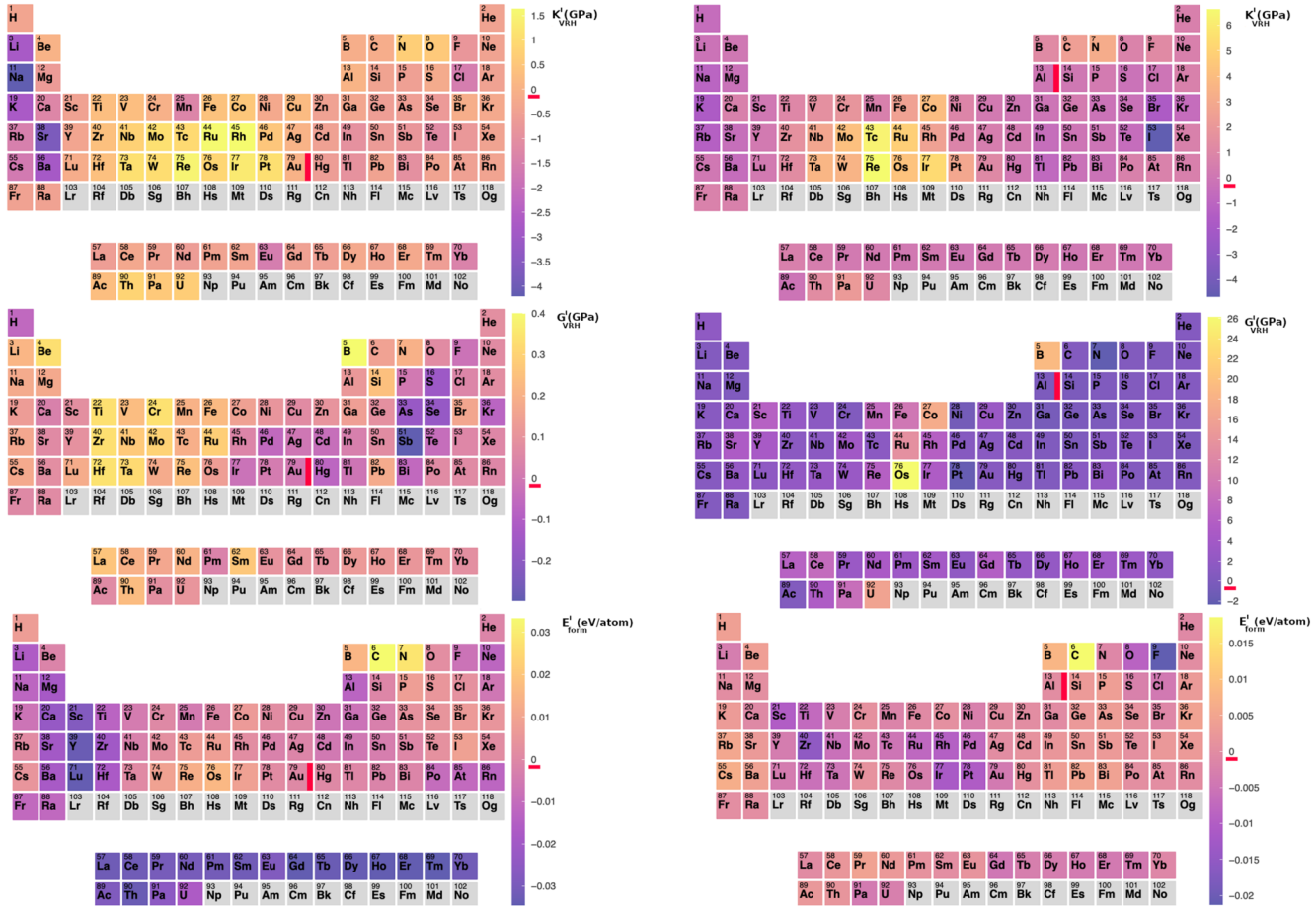
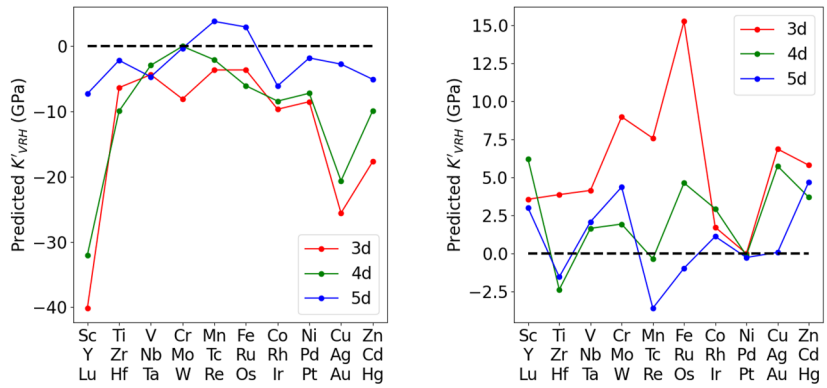
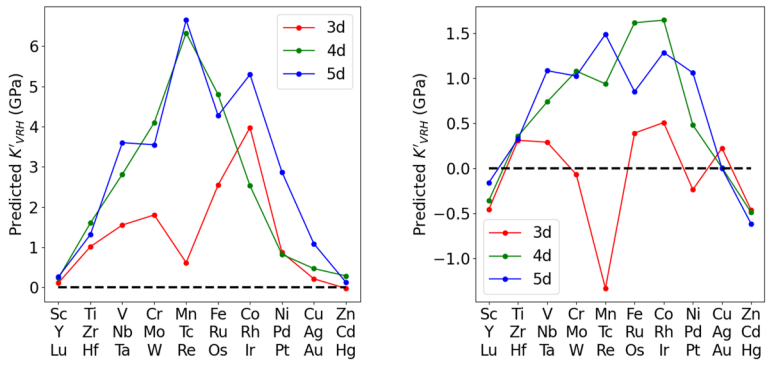


FIG. 20. Comparison between the Periodic Table plots of the predicted bulk modulus, shear modulus, and formation energy for a single-atom substitutionally defected Au [(a)–(c)] and Al [(d)–(f)] supercell with respect to the undefected one for each possible atomic species from the provided Periodic Table.



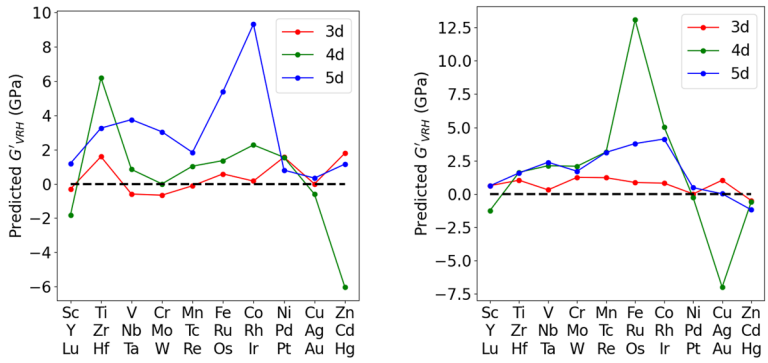
(a) Mo (b) Ni

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.



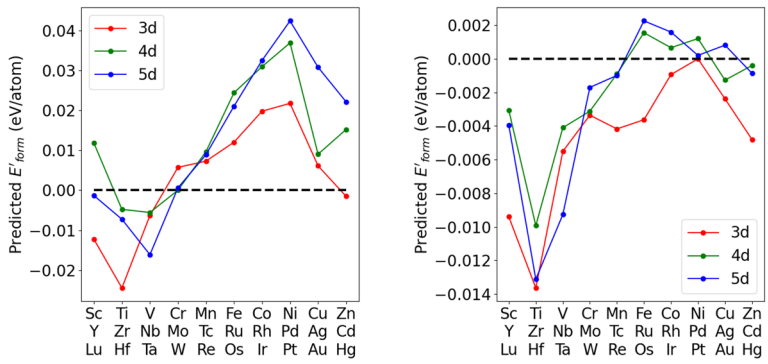
(a) Al (b) Au

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.



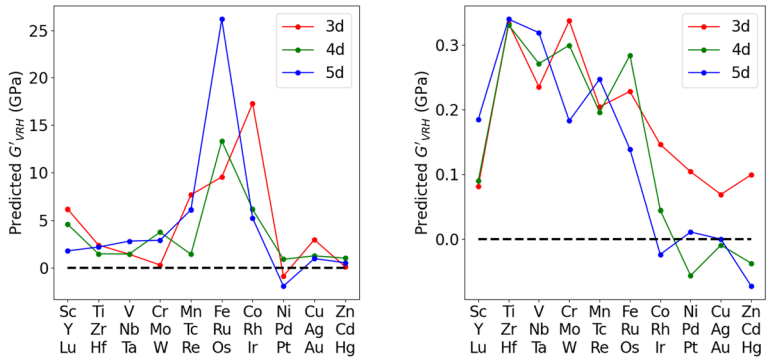
(a) Mo (b) Ni

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.



(a) Mo (b) Ni

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.



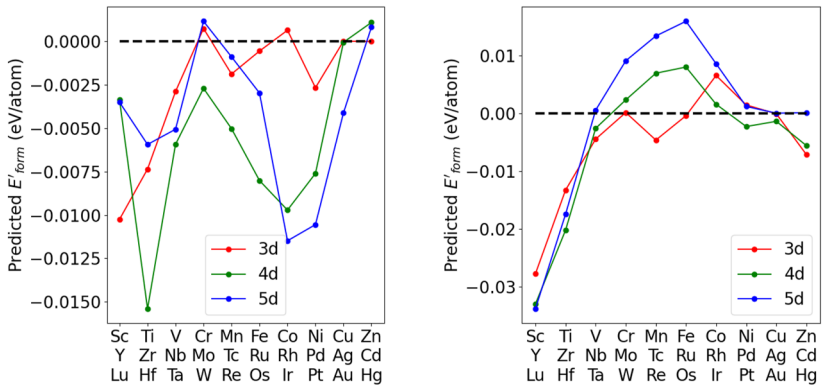
(a) Al (b) Au

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)



(a) Al (b) Au

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