Materials property prediction using symmetry-labeled graphs as atomic-position independent descriptors

Peter Bjørn Jørgensen, Estefanía Garijo del Río, Mikkel N. Schmidt, and Karsten Wedel Jacobsen

Computational materials screening studies require fast calculation of the properties of thousands of materials. The calculations are often performed with Density Functional Theory (DFT), but the necessary computer time sets limitations for the investigated material space. Therefore, the development of machine learning models for prediction of DFT calculated properties are currently of interest. A particular challenge for *new* materials is that the atomic positions are generally not known. We present a machine learning model for the prediction of DFT-calculated formation energies based on Voronoi quotient graphs and local symmetry classification without the need for detailed information about atomic positions. The model is implemented as a message passing neural network and tested on the Open Quantum Materials Database (OQMD) and the Materials Project database. The test mean absolute error is 20 meV on the OQMD database and 40 meV on Materials Project Database. The possibilities for prediction in a realistic computational screening setting is investigated on a dataset of 5976 $ABSe_3$ selenides with very limited overlap with the OQMD training set. Pretraining on OQMD and subsequent training on 100 selenides result in a mean absolute error below 0.1 eV for the formation energy of the selenides.

I. INTRODUCTION

Over the last decades, high-throughput computational screening studies have been employed to identify new materials within different areas such as (photo-)electrochemistry [1–3], batteries [4, 5], catalysis [6, 7], and more [8–10]. Such studies are typically based on Density Functional Theory [11, 12] and because of computational requirements they are usually limited to some thousands or tens of thousands of materials. In order to investigate larger parts of the huge space of possible materials, new methods are needed to perform faster calculations or to guide the search in the material space in a more informed way.

One way to circumvent the computationally demanding DFT calculations is to use machine learning (ML) techniques to predict materials properties, and this approach has been explored intensively the last years. Several descriptors or fingerprints to characterize the atomic structure of a material have been suggested including the partial radial distribution function [13] and the Coulomb matrix [14]. More involved fingerprints combining many atomic properties and crystal structure attributes based on Voronoi graphs have also been developed [15, 16], along with graph representations, which are directly mapped onto convolutional neural networks [17–19].

The use of ML to speed up DFT calculations may have several goals in a computational screening setting. If the atomic structure (i.e. the positions of all the atoms) of a material is known, ML may in principle provide the same information about the material as a DFT calculation would: structural stability, phonon dispersion relations, elastic constants etc. It might even in principle provide data of a better quality than standard (semi-)local DFT calculations, comparable to more advanced DFT calculations with hybrid functionals or even higher-level methods as recently demonstrated for molecules [20].

However, the atomic positions of *new* materials will generally not be known. If the atomic positions are known from experiment, the material is not really new (even though many of its properties might be unknown) and if the positions are obtained from a DFT calculations there is no need to use a ML prediction of already calculated properties.

Our focus here will be the prediction of properties of *new* materials where the detailed atomic positions are unknown, and since the most crucial property of a new material is its stability we shall concentrate on prediction of formation energies.

The obvious question of course then is, how we can describe or classify a crystalline material without knowing the explicit positions of the atoms. The most fundamental property of a material is its chemical composition, i.e. for a ternary material $A_xB_yC_z$, the identity of the elements A, B, and C and their relative appearance $x : y : z$. It turns out that based on this information alone a number of predictions about material stability can be made. Meredig et al. [21] demonstrated that it is possible to predict thermodynamic stability of new compounds with reasonable accuracy based on composition alone, and a number of new compound compositions were predicted and their structures subsequently determined. However, this approach of course has its limitations as it cannot distinguish between materials with the same composition but different crystal structures.

A rigorous classification of a crystalline material comes from its symmetry. Any periodic material belongs to one of the 230 space groups, and this puts restrictions on the possible atomic positions. In the simplest cases of, say, a unary material with one atom in the unit cell with space group Fm-3m (an fcc crystal), all atomic positions are determined up to a scaling of the volume. Similarly, the fractional positions (i.e. relative to the unit cell) of the atoms in materials with several elements can be determined entirely by symmetry as for example shown for $BaSnO_3$ in the cubic perovskite structure in Figure 1. More generally, scaled atomic positions may be fully or

partially determined depending on their symmetry, and the symmetry properties can be expressed using the so-called Wyckoff sites. This classification was recently used by Jain and Bligaard [22] to build a machine learning model based on only composition and the Wyckoff positions, i.e. without any detailed information about the atomic positions. They were able to achieve a mean absolute errors of about 0.07 eV/at on the prediction of the formation energy on a test dataset of more than 85000 materials.

Here, we shall develop a machine learning model, which does not require knowledge of the detailed atomic positions. However, unlike the model proposed by Jain and Bligaard, it will be based on local information about interatomic bonds and the symmetry of their environments. The bonds will be identified using Voronoi graphs and the symmetry will be classified using the Voronoi facets. The resulting model has a mean absolute error on the heats of formation for the OQMD database of only 21 meV and for the ICSD part of OQMD it is 38 meV.

In section II we describe the proposed graph representation based on quotient graphs and the classification of Voronoi facet point symmetry and in section III we investigate the relation between quotient graphs and prototypes based on data from OQMD. This is followed by an introduction of the machine learning model and the datasets in section IV and V respectively. The numerical results are presented in section VI and followed by the conclusions in section VII.

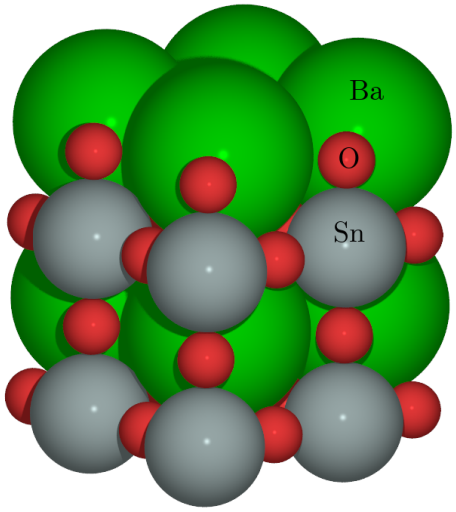


Figure 1. Structure of BaSnO3. The unit cell contains one Ba atom (green), one Sn atom (grey) and three O atoms (red).

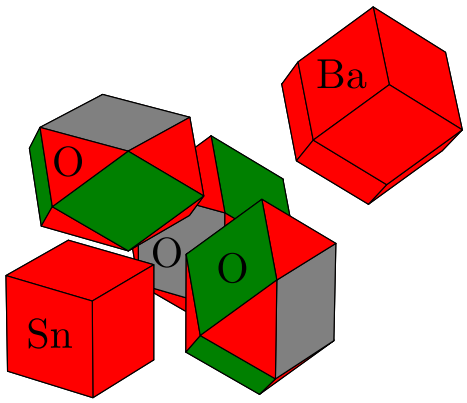


Figure 2. Voronoi cells of BaSnO3. The cells have been displaced for the visualization. The color of the facets corresponds to the atomic species of the neighbouring atom.

II. GRAPH REPRESENTATION

As representation for the machine learning algorithm we use the quotient graph as introduced by [23] and also used in [19]. The quotient graph is a finite graph representation of the infinite periodic network of atoms. Every atom in the unit cell corresponds to a vertex of the quotient graph. We denote the graph G and the set of N vertices $\{v_i\}_{i=1}^{N}$. When two atoms are connected in the network we draw an edge between the atoms in the quotient graph. In this work we use the Voronoi diagram to decide when two atoms are connected, specifically a pair of atoms are connected if they share a facet in the Voronoi diagram. Due to periodic boundary conditions a pair of atoms may share several facets and in this case there will be several edges between the atoms. When interatomic distances are available the edges are labeled with the distance between the atoms.

As an example we look at BaSnO3 in the perovskite structure as shown in Figure 1. This material has five atoms in the unit cell. After performing Voronoi tessellation we get a Voronoi cell for each atom in the unit cell as shown in Figure 2. The Voronoi diagram defines the edges in the quotient graph which is illustrated in Figure 3.

The Voronoi construction may result in the appearance of quite small facets. This is for example often the case for structures with high symmetry, where small displacements of the atoms introduce new facets. We remove these small facets and the corresponding connections in the graph by introducing a cutoff in the solid angle of the facet Ω_{cut}. We use $\Omega_{cut} = 0.2$, but as we shall see later the results are surprisingly stable with regards to increasing this value.

The graph is annotated with the symmetry group of each of the Voronoi facets. In the following section we describe this symmetry classification in more detail.

A. Symmetry Group Classification

To characterize the symmetry of an atomic environment we classify the symmetry of each Voronoi facet into the 9 non-trivial two-dimensional point groups $(C_2, C_3, C_4, C_6, D_1, D_2, D_3, D_4, D_6)$. The classification

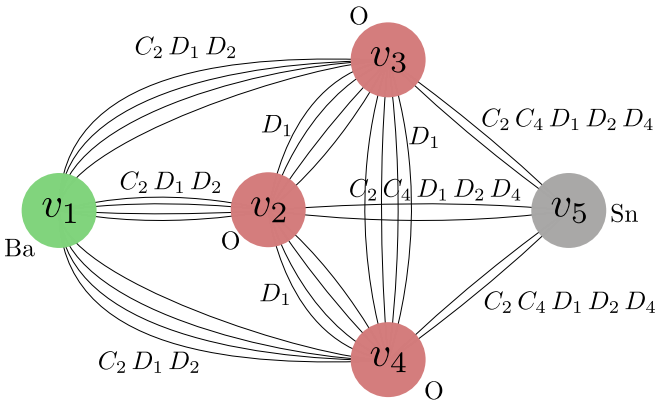


Figure 3. Quotient graph for BaSnO3. The edge labels shows the point groups of the corresponding facets of the Voronoi diagram. For this particular case, the repeated edges between vertices all have the same point groups, but in general they could be different.

method is inspired by the symmetry measure introduced by Heijmans and Tuzikov [24]. Given the vertices of the two-dimensional Voronoi facet we go through the following procedure

1. Compute centroid and center the shape.
2. Search for mirror axis and align it with the x-axis if it exists.
3. For each point symmetry group apply all elements of the group and calculate the area of the convex hull of the new points generated by this procedure.

The symmetry measure is then the ratio between the area of the original shape and the area defined by the convex hull of the new vertices. When the symmetry measure for a given group is close to unity we label the facet as having this symmetry. See Figure 5 for an example shape and its symmetry measure for each group. The search for mirror axis in step 2 is done by computing the moment of inertia and test the two principal axes for mirror symmetry. When the moments of inertia are the same, for example when the shape is a regular polygon, the principal axes are arbitrary and we fall back to testing for mirror symmetry at all axes going through the centroid and either a vertex or a midpoint of a line segment. For a regular hexagon these axes are illustrated in Figure 4.

III. GRAPH REPRESENTATION AND PROTOTYPES

In many applications prototypes are used as a descriptor for the overall structure of a material and as part of a computational screening procedure some of the atoms of the prototypes may be swapped with other elements. We want to assess whether there is a correspondence between the prototypes and Voronoi graphs, i.e. do two materials with the same prototype have the same Voronoi graph and do two materials with the same Voronoi graph have the same prototype? The question cannot be ultimately answered because prototype naming is not completely well-defined: in some cases several different prototypes are used to describe the same material, and many materials may not have prototypes attached to them. But we can show to which extent Voronoi graphs is aligned with the use of prototypes.

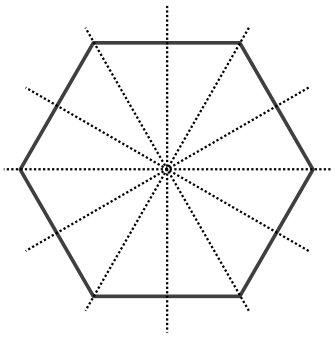


Figure 4. Mirror axes of a hexagon.

For this analysis we use the OQMD database and look at all unary, binary and ternary compounds that are labeled with a prototype. For each of these sets we want to look at the link between the graphs G and the prototypes P, i.e. if we know that a given structure has a specific prototype do we then also know which graph it has and vice versa. One way of measuring this is through the mutual information $I(G;P)$ of G and P. The mutual information is symmetric and can be computed as

$$I(G;P) = H(G) - H(G|P) \quad (1)$$
$$= H(P) - H(P|G), \quad (2)$$

where H denotes the entropy. The mutual information is thus the average decrease in entropy we get from knowing the other variable. We also compute the normalized mutual information known as the uncertainty coefficient $U(X|Y) = I(X;Y)/H(X)$ which can be seen as given Y what fraction of bits of X can we predict. To compute these quantities we need the distribution over graphs and we obtain these distributions approximately by comparing graph fingerprints.[25] The quantities for OQMD are shown for the unlabeled graph in Table I and for the graph labeled with rotation symmetries in Table II.

The uncertainty coefficient is close to 90% in most cases except for the Unary compounds $U(P|G)$. In this case structures with different prototypes map to the same graph and we may be discarding important structural information. Including symmetry information increases the number of unique graphs significantly, which implies that the uncertainty coefficient $U(G|P)$ decreases while $U(P|G)$ increases.

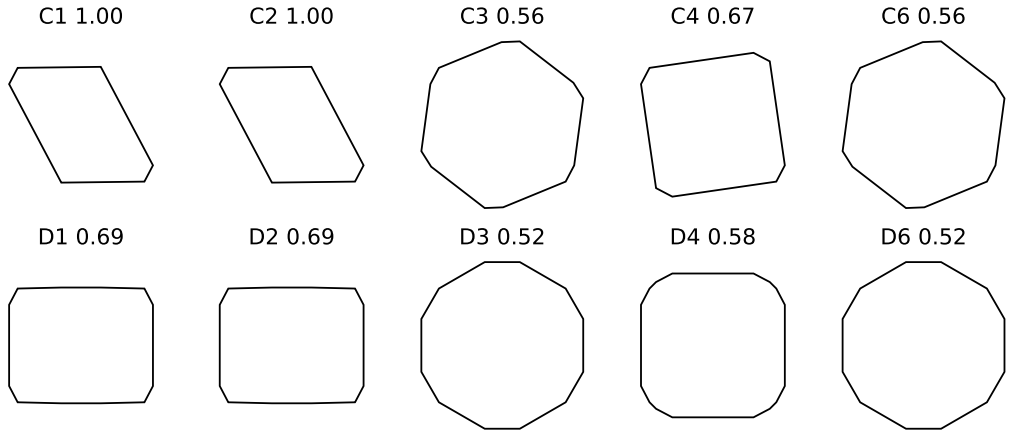


Figure 5. Convex hull of the shape in the top left corner after the symmetry operations of the corresponding groups has been applied. The label above each shape denotes the point group and the symmetry measure for that group.

	N	$\|G\|$	$\|P\|$	$H(G)$	$H(P)$	$I(G,P)$	$U(G\|P)$	$U(P\|G)$
Unary	1487	85	67	4.4	4.7	3.7	0.84	0.80
Unary ICSD	196	46	49	4.2	4.2	3.8	0.90	0.90
Binary	53528	1318	871	4.3	4.5	3.8	0.90	0.86
Binary ICSD	5862	1219	850	8.2	8.0	7.6	0.92	0.95
Ternary	339960	4006	1754	2.0	1.9	1.8	0.91	0.98
Ternary ICSD	11500	3487	1740	10.0	9.1	8.8	0.88	0.97

Table I. Correspondence between Voronoi graphs and prototypes in OQMD without symmetry labels. N denotes the number of entries, $|G|$ the number of unique Voronoi graphs and $|P|$ the number of different prototypes.

IV. NEURAL MESSAGE PASSING MODEL

In this section we introduce the machine learning model which takes the labeled graph as input and outputs an energy prediction as a scalar. We describe the model as a message passing framework on a graph, similarly to Gilmer *et al.* [26]. Denote the graph G with vertex features x_v and edge features ε_{vw} for an edge from vertex v to vertex w. Each vertex has a hidden state h_v^t at "time" (or layer) t and we denote the edge hidden state e_{vw}^t. The hidden states of vertices and edges are updated in a number of interaction steps T. In each step the hidden state of vertices are updated in parallel by receiving and aggregating messages from neighbouring vertices. The messages are computed by the message function $M_t(\cdot)$ and the vertex state is updated by a state transition function $S_t(\cdot)$, i.e.

$$m_v^{t+1} = \sum_{w \in N(v)} M_t\left(h_v^t, h_w^t, e_{vw}^t\right), \tag{3}$$

$$h_v^{t+1} = S_t\left(h_v^t, m_v^{t+1}\right), \tag{4}$$

where $N(v)$ denotes the neighborhood of v, i.e. the vertices in the graph that has an edge to v. The edge hidden states are also updated by an edge update function $E_t(\cdot)$ that depends on the previous edge state as well as the vertices that the edge connects:

$$e_{vw}^{t+1} = E_t(h_v^t, h_w^t, e_{vw}^t). \tag{5}$$

After T interaction steps the vertex hidden state represents the atom type and its chemical environment. We then apply a readout function $R(\cdot)$ which maps the set of vertex states to a single entity

$$\hat{y} = R\left(\left\{h_v^T \in G\right\}\right) \tag{6}$$

The readout function operates on the set of vertices and must be invariant to the ordering of the set. This is often achieved simply by summing over the vertices. In some architectures the final edge states are also included as an argument to the readout function. The message function $M_t(\cdot)$, state transition function $S_t(\cdot)$, edge update function $E_t(\cdot)$ and readout function $R(\cdot)$ are implemented as neural networks with trainable weight matrices. To fully define the model we just need to define these functions and a number of models can be cast into this framework. We use different weight matrices for each time step t, however in some architectures the weights are shared between layers to reduce the number of parameters.

In this work we use the model proposed in our prior work [27]. The model can be seen as an extension of the SchNet model [18], with the addition of an edge update

	N	$\lvert G\rvert$	$\lvert P\rvert$	$H(G)$	$H(P)$	$I(G,P)$	$U(G\lvert P)$	$U(P\lvert G)$
Unary	1487	316	67	6.6	4.7	4.4	0.67	0.94
Unary ICSD	196	90	49	5.4	4.7	4.2	0.77	0.99
Binary	53528	2491	871	5.6	4.5	4.3	0.77	0.96
Binary ICSD	5862	1997	850	9.1	8.0	7.8	0.86	0.98
Ternary	339960	6927	1754	2.1	1.9	1.8	0.86	0.99
Ternary ICSD	11500	5169	1740	10.9	9.1	10.0	0.82	0.99

Table II. Correspondence between Voronoi graphs and prototypes in OQMD with symmetry labels. N denotes the number of entries, $|G|$ the number of unique Voronoi graphs and $|P|$ the number of different prototypes.

network. The message function is only a function of the sending vertex and can be written as

$$M_t(h_w^t, e_{vw}^t) = (W_1^t h_w^t) \odot g(W_3^t g(W_2^t e_{vw}^t)), \quad (7)$$

where $\odot$ is element-wise multiplication and $g(x)$ is the activation function, more specifically the shifted soft-plus function $g(x) = \ln(\mathrm{e}^x + 1) - \ln(2)$. It can be seen as a smooth version of the more popular rectified linear unit. As an edge update network we use a two layer neural network and the input is a concatenation of the sending and receiving vertex states and the current edge state.

$$e_{vw}^{t+1} = E_t(h_v^t, h_w^t, e_{vw}^t) = g(W_{\mathrm{E2}}^t g(W_{\mathrm{E1}}^t (h_v^t; h_w^t; e_{vw}^t))), \quad (8)$$

where $(\cdot;\cdot)$ denotes vector concatenation. This choice of edge update network means that the edge state for each of the two different directions between a pair of vertices become different after the first update. The state transition function is also a two layer neural network. It is applied to the sum of incoming messages and the result is added to the current hidden state as in Residual Networks [28]:

$$S_t\left(h_v^t, m_v^{t+1}\right) = h_v^t + W_5^t g(W_4^t m_v^{t+1}), \quad (9)$$

After a number of interaction steps T we apply a readout function for which we use a two layer neural network that maps the vertex hidden representation to a scalar and finally we average over the contribution from each atom, i.e.

$$R\left(\left\{h_v^T \in G\right\}\right) = \frac{1}{N} \sum_{h_v^T \in G} W_7 g(W_6 h_v^T), \quad (10)$$

In other words an atom and its chemical environment is mapped to an energy contribution.

A. Initial Vertex and Edge Representation

The initial vertex hidden state h_v^0 depends on the atomic number of the corresponding atom. The atomic number is used to look up a vector representation for that atom. Using a hidden representation of size 256 the initial hidden state is thus the result of a lookup function $\ell(x) : \mathbb{N} \to \mathbb{R}^{256}$. The weights in the vector representation are also trained during the optimization.

We use the model on three different levels of available information. In the most ignorant scenario we have no labels on the edges of the graph and in this case the edge update function (8) just ignores the edge representation on the first layer, i.e. e_{vw}^0 is a "vector" of length 0 and $e_{vw}^t, t \in 1, \ldots, T$ are vectors of length 256. The next level of information is to include the point group information as described in section II A. There are 9 non-trivial point groups and we encode this information as an indicator vector of length 9, where 1 means that the corresponding facet belongs to the given point group. Finally we also run numerical experiments with the full spatial information for which the edges of the quotient graph are labeled with the interatomic distance. The distances are encoded by expanding them in a series of exponentiated quadratic functions as also done in [17, 18, 27]:

$$(e_{vw}^0)_k = \exp\left(-\frac{(d_{vw} - (-\mu_{\min} + k\Delta))^2}{2\Delta^2}\right), k = 0 \ldots k_{\max} \quad (11)$$

where $\mu_{\min}$, Δ, and $k_{\max}$ are chosen such that the centers of the functions covers the range of the input features. This can be seen as a soft 1-hot-encoding of the distances, which makes it easier for a neural network to learn a function where the input distance is uncorrelated with the output. In the experiments we use $\mu_{\min} = 0$, $\Delta = 0.1$, and $k_{\max} = 150$.

V. DATASETS

For the numerical experiments we use two publicly available datasets and one dataset we generate.

a. The Materials Project [29] This dataset contains geometries and formation energies of 86 680 inorganic compounds with input structures primarily taken from the The Inorganic Crystal Structure Database (ICSD) [30]. We use the latest version of the database (version 2018.11). The number of examples is reduced to 86 579 after we exclude all materials with noble gases (He, Ne, Ar, Kr, Xe) because they occur so infrequently in the dataset that we consider them as outliers. This brings the number of different elements in the dataset down to 84.

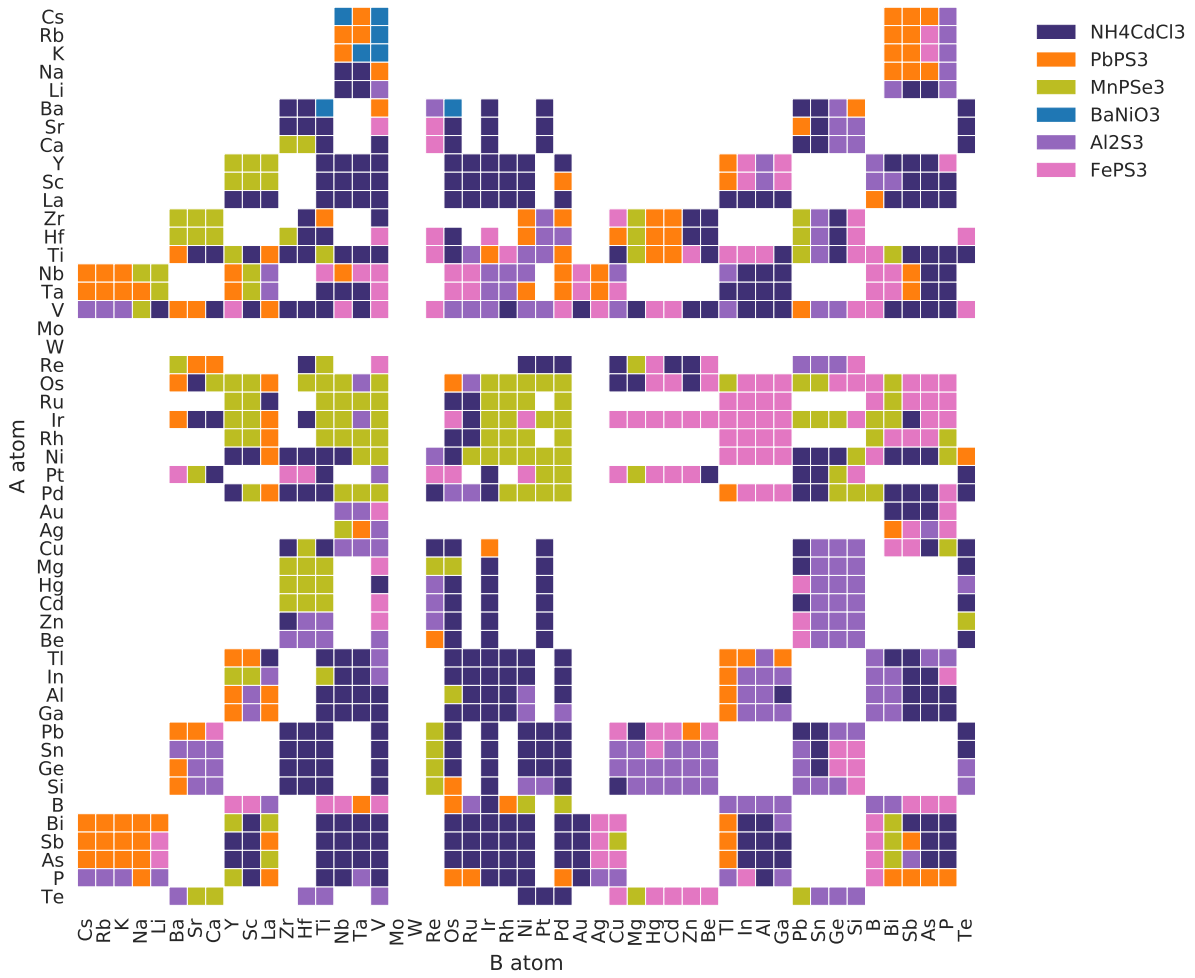


Figure 6. Map of the most stable prototype for each composition $ABSe_3$. The compositions that do not fulfill the valence rule have not been studied and thus, they are not colored.

b. Open Quantum Materials Database (OQMD) [31, 32] Is also a database of inorganic structures and we use the currently latest version (OQMD v1.2) available on the project's website. Again we consider materials with noble gases as outliers and we also exclude highly unstable materials with a heat of formation of more than 5 eV/atom. Some entries in the database are marked as duplicates and we filter them in the following way: when a set of duplicates is encountered we use the first entry of the database, but if the standard deviation of the calculated heat of formation exceeds 0.05 eV/atom we discard the whole set of duplicates. This leaves us with a total of 562 134 entries.

For both datasets we split the entries into five parts of equal size to be used for 5-fold cross-validation, where the machine is trained on 4/5 of the data, and the remaining 1/5 is used for testing. For OQMD we also distribute the entries of OQMD that originates from ICSD equally between the five folds.

c. Ternary Selenides $ABSe_3$ For further testing, we have developed a third dataset of selenides. The intention behind this set is to test the ability of the model in a realistic computational screening setting. This dataset has only very limited overlap with OQMD, and predictions are made exclusively based on the symmetry labeled graphs of the new materials without any detailed information about the atomic coordinates.

The dataset contains the structures and formation energies of 5976 ternary selenides with stoichiometries $ABSe_3$, where A and B are different transition metals in six different prototypes.

The procedure for generating this dataset resembles the one presented in reference [3]. We start by looking up the $ABSe_3$ compounds reported in ICSD [30], and selecting the 6 prototypes that appear more than once: hexagonal $P6_3/mmc$ structure of $BaNiO_3$, orthorhombic $Pnma$ structure of NH_4CdCl_3/Sn_2S_3, monoclinic $C2/m$ $FePS_3$, monoclinic Pc structure of $PbPS_3$, trigonal $R\bar{3}$ structure of $MnPSe_3$ and hexagonal $P6_1$ structure of Al_2S_3.

These structures are then used as templates, and we substitute the transition metal atoms A and B by 49 transition metals. Here, we avoid for simplicity Cr, Mn, Fe and Co, which usually lead to structures with large magnetic moments. We also limit ourselves to those combinations $ABSe_3$ for which the valences of cations and anions add up to zero. This leads to a total of 512 $ABSe_3$ compounds: 484 ternaries, which are then studied in 12 structures (6 for the $ABSe_3$ and 6 for the $BASe_3$) and 28 binaries, for which we study 6 different structures. A map to the compositions and structures studied can be found in figure 6.

The resulting 5976 structures have then been relaxed using Density Functional Theory (DFT) as implemented in the codes ASE [33] and GPAW [34]. We perform two different kinds of electronic structure calculations: a coarse-grained calculation with the exchange-correlation functional PBEsol [35] for the steps of the optimization and a fined-grained at the relaxed structure with the PBE exchange-correlation functional [36]. The cuoff energy for the plane wave basis set used to expand the wave functions is 800 eV in both cases. For the sampling of the Brillouin zone we use a Monkhorst–Pack mesh [37] with a density of $5.0/(\text{Å}^{-1})$ k-points in each direction for the relaxation steps and of $8.0/(\text{Å}^{-1})$ k-points for the refined calculation at the relaxed structure. All structures have been relaxed until the forces on the atoms have reached at least 0.05 eV/Å.

VI. NUMERICAL RESULTS AND DISCUSSION

To assess the loss in accuracy going from full spatial information to unlabeled quotient graph we train/test the model in three different settings as mentioned in section IV A. In the most ignorant setting the quotient graph has only unlabeled edges. On the next level we label the edges with the symmetry of the corresponding Voronoi facet. With full spatial information the edges of the quotient graph are labeled with the distance between the atoms. The model is trained with the Adam optimizer [38] for up to 10×10^6 steps using a batch size of 32. The initial learning rate is 1×10^{-4} and it is decreased exponentially so at step s the learning rate is $10^{-4} \cdot 0.96^{\frac{s}{10^5}}$. When training on OQMD and Materials Project we use 5000 examples from the training data for early stopping. More specifically this validation set is evaluated every 50 000 steps and if the mean absolute error (MAE) has not improved for 1×10^6 steps the training is terminated. When training on the ternary selenides $ABSe_3$ dataset the 10% of the training data is used as a validation set and the validation set is evaluated every training epoch. In some of the results we use a model that has been pretrained on OQMD. In that case the model is trained on 4 out of 5 OQMD folds until the stopping criterion is met and the weights of the model are then used as initialization for training on the selenides dataset.

	Dist.	Sym	No sym	V-RF
OQMD all	14	20	26	85
OQMD unary	58	108	128	85
OQMD binary	30	39	60	86
OQMD ternary	14	19	23	80
ICSD all	24	38	45	113
ICSD unary	56	75	119	72
ICSD binary	32	47	58	118
ICSD ternary	22	34	39	109
Matproj all	26	40	43	84
Matproj unary	96	144	179	127
Matproj binary	48	69	73	99
Matproj ternary	27	40	43	87

Table III. MAE of test set energy predictions in meV/atom. The ICSD results are for the model trained on OQMD and tested only on the ICSD part of OQMD.

	Dist.	Sym	No sym	V-RF
OQMD all	54	70	80	173
OQMD unary	184	257	342	190
OQMD binary	89	98	138	162
OQMD ternary	52	68	71	131
ICSD all	81	98	111	188
ICSD unary	262	222	353	180
ICSD binary	73	111	129	202
ICSD ternary	88	85	102	182
Matproj all	72	115	122	172
Matproj unary	246	349	467	289
Matproj binary	120	180	192	203
Matproj ternary	65	108	111	181

Table IV. RMSE of test set energy predictions in meV/atom. The ICSD results are for the model trained on OQMD and tested only on the ICSD part of OQMD.

A. OQMD

The mean absolute errors (MAE) and root mean squared errors (RMSE) of the test set predictions are shown in Table III and the MAE is further visualized in Figure 7. As expected, the lowest prediction errors are obtained with the model where distance information is provided. If we focus on the OQMD the overall MAE is as low as 14 meV with distance information. This is lower than the SchNet-model [18] by almost a factor of two because of the edge updates as discussed in Ref. [27]. Two versions of the models without distance information are also shown. In one of them the symmetry information has not been used, but in the other one the symmetry classification of the Voronoi facets has been included as edge information. These two models do of course have larger errors than the one benefiting from the distance information, but still the error is surprisingly small. The MAE is only 20 meV for the model using symmetry information. For comparison the results for the model proposed by Ward et al. [15] is also shown in the figures (labeled V-RF for Voronoi - random forest). This model

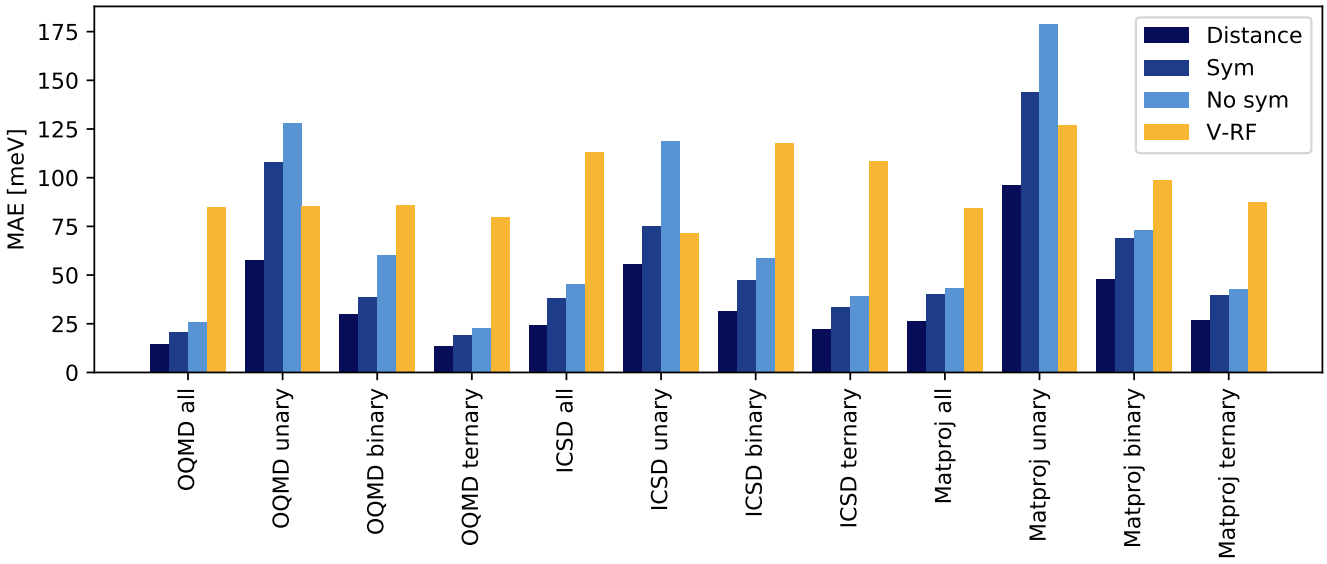


Figure 7. The figure shows the data in Table III

also builds on a Voronoi graph construction, but since the fractional areas of the Voronoi cells are provided, information about the distances are provided. Furthermore, many other attributes are added as information to the random forest algorithm applied. When this machine is applied to OQMD (using the same 5-fold splitting of the data as applied to the other algorithms) the resulting error is considerably larger, 85 meV, for all of OQMD.

To understand more about the behavior of the ML algorithms investigated here, we have considered the test errors on different subsets of OQMD and also on the Material Project database [29]. Let us first note that the OQMD contains two different types of structure sources. One type, which gives rise to the largest number of materials, consists of a number of fixed crystal structures or prototypes decorated by the different chemical elements. There are 16 elemental prototypes, 12 binary ones, and 3 ternary ones. For two of the ternary ones, one of the elements is predefined to be oxygen. This generates a very large number of materials of varying composition and stability, but in a fairly small number of different crystal structures. The other type of structures comes from materials from the experimental ICSD database. This group is characterized by a much greater variation in the crystal structures, but is naturally limited to mostly stable materials, since they have been experimentally synthesized.

We first consider the test error on the subsets of OQMD consisting of the unary, binary, and ternary systems, and we shall focus on the model where the symmetry information is included, but the distances are not. As can be seen from Table III, the test error is considerably larger on the unary systems (108 meV) than on the database as a whole. This also holds for the binary ones but to a smaller degree (39 meV). It is not clear to us at the moment exactly why this is so, but we shall discuss some possible explanations. The unary and binary systems only constitute a fairly small part of the total database, and the weight of these systems during the training is therefore also limited. Another factor may be that a large fraction of the unary and binary systems belong to the group of materials where the crystal structures are systematically generated as mentioned above. This means that many rather "artificial" and unstable materials are generated, where the atoms are situated in environments, which do not occur in reality, and the resulting energies may be far above more stable structures. This could potentially be difficult for the machine to learn.

B. ICSD/OQMD

Table III also shows the results for the ICSD subset of the OQMD database. The results shown are for the model trained on all of OQMD but tested only on the ICSD subset. The overall MAE is seen to be roughly a factor of two larger than for all of OQMD. This is probably due to the fact that the ICSD is a subset with a large variation of structures and this makes prediction more difficult on average. We see the same trend as for all of OQMD, that the error decreases going from unary to binary to ternary systems. For the unary systems the test error is in fact lower for the ICSD subset than for all of OQMD, which may be due to the fact that physically artificial high-energy systems appear in OQMD but not in ICSD. For the binary systems there is a balance: the ICSD does not contain so many high-energy systems, which could make predictions better, but on the other hand the larger variation of crystal structures is more difficult to predict.

C. Materials Project Database

The models have also been trained and tested on the Materials Project dataset [29]. The overall error is fairly similar to the one obtained for the ICSD subset of OQMD as might be expected since the Materials Project is also based on mostly materials from the ICSD. The errors for the unary and binary subsets are somewhat larger for the Materials Project database. This might be due to the fact that the machine trained on OQMD benefits from the larger number of systematically generated unary and binary systems in that database.

D. RMSE vs. MAE

The root-mean-square-errors are shown in Table IV. In all cases the values are considerable higher than the MAE. This is an indication that the distribution of the errors have heavier tails than a Gaussian, and as we shall see in the following examples that a significant number of outliers exist. The outliers might be due to limitations of the model but could also appear because of problematic entries in the database as also discussed by Ward et al. [15].

E. Solid angle cutoff of Voronoi facets

The above results are all calculated using a cutoff of the Voronoi facet solid angle of $\Omega_{cut} = 0.2$. However, the results are almost independent of the value as shown in Figure 8, where the MAE on all of OQMD is shown for the model where symmetry but no distance information is included. We see that the error decreases slightly when small facets are removed with $\Omega_{cut} = 0.2$, and increases only slowly when Ω_{cut} is further increased. We take this as an indication that the connectivity of the material is well described even when the graph is reduced to essentially include only nearest neighbor bonds.

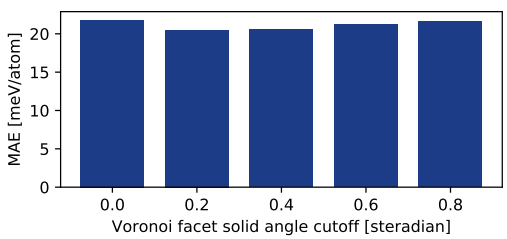


Figure 8. Prediction error vs Voronoi facet solid angle cutoff Ω_{cut} for the model using symmetry labels.

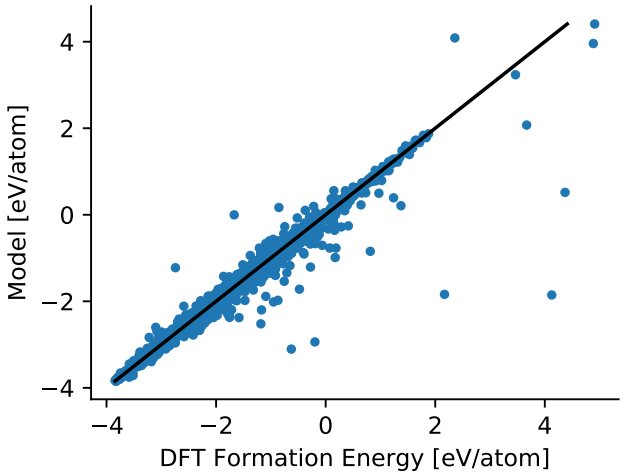


Figure 9. Predictions on 12395 ABO_3 structures of OQMD (MAE=35, RMSE=110 meV/atom)

F. ABO_3 materials in OQMD

We now consider the subset of all oxides in the OQMD with the composition ABO_3. We shall investigate to which extent the model is able to predict the right ground state structure for a given composition. We first show the overall prediction for the 12935 materials of this type in OQMD in Figure 9. We again use the model with symmetry-labeled graphs without distance information. The MAE is 35 meV, which is about the same value as the one for the subset of ternaries in ICSD (34 meV). The RMSE is again significantly higher (110 meV) because of severe outliers as can be seen in the plot.

We now ask the following question: given a composition (A,B) the model predicts a ground state structure G_{ML}. If we are going to investigate this structure and other low energy structures with DFT, how high up in energy (as predicted by the model) do we have to go before we find the DFT ground state structure G_{DFT}? We only include entries for which there is more than one structure (12329/12395) and the average number of structures per composition is 4.7 The energy difference $\Delta E = E^{ML}(G_{\mathrm{DFT}}) - E^{ML}(G_{\mathrm{ML}})$ of course varies from system to system, and the distribution is shown in Figure 10. The mean absolute difference (MAD) of this distribution is very small, only 8 meV, and the maximum error is a clear outlier at 0.98 eV. The reason for the small MAD is that for 2113 out of the 2646 compositions the correct ground state is predicted, however, in many cases because only two structures exist in the database for a given composition. If we only look at the 533 compositions for which the ML model predicts the wrong ground state the MAD is 42 meV/atom. For comparison the energy prediction for the ground state structures has an MAE of 28 meV/atom. The low MAD value of 42 meV is promising for applications to computational screening.

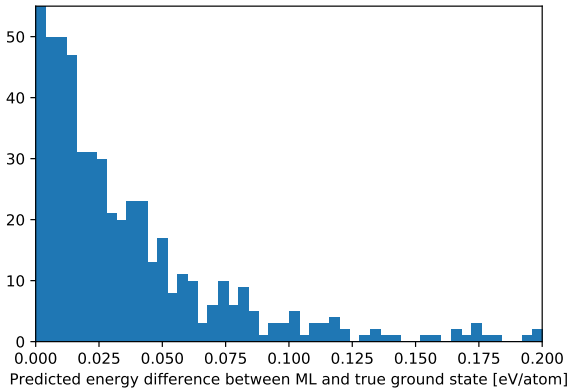


Figure 10. Predicted energy difference between the DFT ground state and the ML ground state: $\Delta E = E^{ML}(G_{\mathrm{DFT}}) - E^{ML}(G_{\mathrm{ML}})$ for the ABO_3 materials in OQMD. The total number of compositions is 2646. The peak at zero is much higher than shown in the graph. It corresponds to the 2113 compositions, where the right ground state is predicted. For the remaining 533 compositions the mean absolute difference is 42 meV/atom.

It sets an energy scale for how many structures have to be investigated by DFT to identify the DFT ground state after the model predictions have been generated.

G. $ABSe_3$ selenides

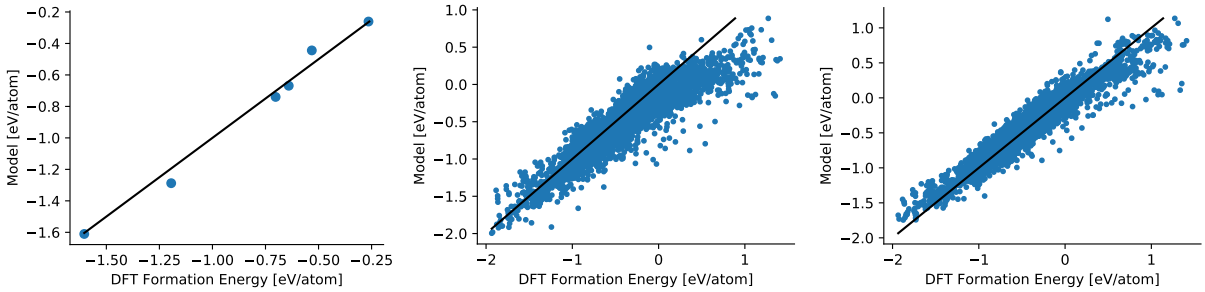
The last dataset we shall consider consists of selenides with the $ABSe_3$ composition as discussed in the section about the datasets. This dataset is considerably more challenging for two reasons. Firstly, there is very little overlap between this dataset and the training dataset OQMD. Only 6 materials are shared between the two datasets, and the test predictions for these are shown in Figure 11a. The MAE is 43 meV, and the RMSE is also low, only 56 meV. The second challenge is, that we shall now use the model to make predictions based on the initial graph before relaxations. The 6 different prototypes in the dataset each have a graph in the original material giving rise to the naming of the prototype. For example, one of the types is hexagonal $P6_3/mmc$ structure of $BaNiO_3$, so for predictions in this structure we shall use the graph of $BaNiO_3$. Some of the prototype structures have a fair number of atoms in the unit cell (up to 20) and a low symmetry (monoclinic), which means that there are many free atomic coordinates that are optimized during relaxation. This leads to frequent modifications of the graph during relaxation.

Figure 11b shows the model predictions based on the initial prototype graphs versus the DFT energies of the resulting optimized structures. The MAE is 165 meV, which is considerably higher than the value for the oxides. Particularly large deviations are seen for large and positive heats of formation. In a computational screening setting this might not be an issue because the high-energy materials are going to be excluded anyway. The RMSE is only a factor 222/165 = 1.35 larger than the MAE, which is due to the small number of outliers compared to for example the oxides (Figure 9).

The prediction quality can be significantly improved by additional training on the selenide dataset. Even a limited number of data points have a considerable effect. This is to be expected since the overlap between the selenide dataset and the OQMD is only 6 materials as mentioned above. Figure 11c shows the model-DFT comparison if the model is first trained on the OQMD dataset and then subsequently trained on 100 materials out of the 5976 selenides in the database. The MAE is reduced from 165 meV to 96 meV bringing the error down to a value comparable to the error between DFT and experiment [32].

The effect of additional training on the selenide dataset is shown as a function of training set size on a logarithmic scale in Figure 12. The points on the y-axis correspond to the situation without any additional training. As can be seen, a small amount of additional training leads to significant reduction of the prediction error. The red curve corresponds to the situation discussed above where the model is first trained on OQMD, and then further trained on the initial graphs (but relaxed energies) for part of the selenides. For comparison, the green curve shows the prediction error, when the training and prediction is based on the final graph. Using the initial graphs instead of the final graphs gives rise to only a slightly higher MAE. This is encouraging for the potential use of the approach in computational screening studies, where predictions have to be based on the initial prototype structures to avoid the computationally costly DFT calculations.

As a baseline we also show the results of the model if it is trained exclusively on the selenide dataset (orange curve). As expected the MAE is much larger than for the pretrained model for small amounts of data. For larger training sets the MAE drops gradually and with a dataset size of about 500 materials the prediction error is comparable to the one for the OQMD-pretrained model, which is trained on an additional 50 selenides. We ascribe the rather successful performance of the model without pretraining at large training set sizes to the systematic character of the dataset: only 6 different crystal structures are represented and they are systematically decorated with a particular subset of atoms. The last model (blue curve) is again only trained on the selenide dataset, but now only one interaction step ($T = 1$) is performed in the message passing neural network in contrast to the three iterations used otherwise. The performance is seen to be rather similar to the model with $T = 3$ up to a training dataset size of 300. With only one iteration in the network information about the identity of neighboring atoms is exchanged, and this is apparently sufficient to roughly characterize the 6 crystal structures. At larger training set sizes, where the prediction error is smaller,



(a) Predicted energy for selenides in OQMD (MAE=43 RMSE=56 meV/atom)

(b) Predicted energy for selenides (initial structures) using model trained on OQMD (MAE=165 RMSE=222 meV/atom)

(c) Predicted energy for selenides (initial structures) using model pretrained on OQMD and then on 100 selenides (MAE=96 RMSE=133 meV/atom)

Figure 11. Model predictions on $ABSe_3$ structures.

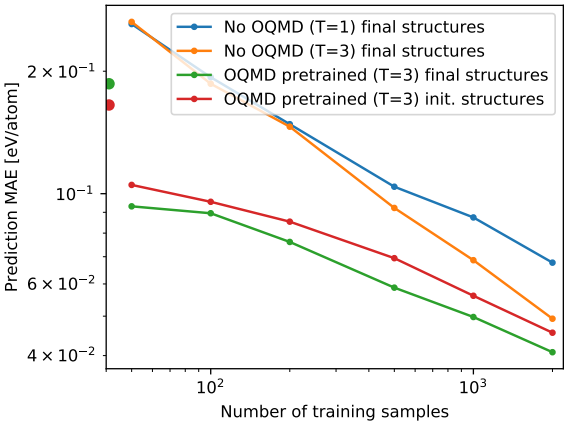


Figure 12. Predictions on $ABSe_3$ structures with increasing number of training samples. The unconnected points correspond to the model only trained on OQMD final structures, i.e. the pretrained model.

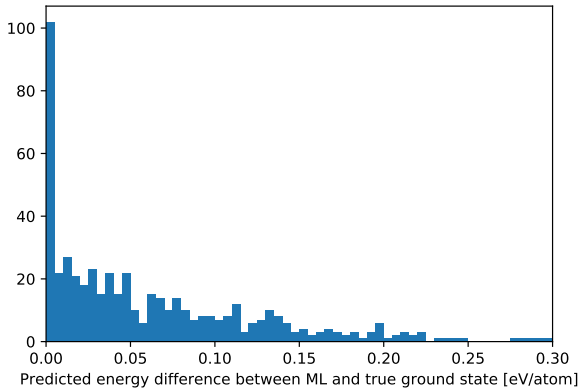


Figure 13. Predicted energy difference between the ML ground state and the true ground state, $\Delta E = E^{ML}(G_{\mathrm{DFT}}) - E^{ML}(G_{\mathrm{ML}})$, for the selenide dataset. The mean absolute difference is 62 meV/atom.

the network with three iterations outperforms the one with only one iteration.

Figure 13 shows the distribution of the predicted energy difference between the DFT ground state structure and the ML predicted ground state structure, $\Delta E = E^{ML}(G_{\mathrm{DFT}}) - E^{ML}(G_{\mathrm{ML}})$ for the selenide dataset. Only in 79 out of the 512 compositions, the model predicts the DFT ground state. This in not particularly impressive, since random prediction of a structure would give roughly $512/12 \approx 43$ correct predictions. However, the dataset have many low-lying energy structures, where even full DFT calculations cannot be expected to necessarily predict the correct ground state structure. This was investigated in more detail in a similarly generated dataset of ABS_3-sulfides used for computational screening of water-splitting materials [3]. The mean absolute difference is only 62 meV/atom with a maximum error of 0.3 eV/atom. The low mean value is clearly promising for future applications to computational materials screening.

VII. CONCLUSIONS

In summary, we have proposed a ML model for the prediction of the formation energy of crystalline materials based on Voronoi quotient graphs and a local symmetry description. It uses a message passing neural network with edge updates. The model is independent of the detailed atomic positions and can therefore be used to predict the formation energy of new materials, where the detailed structure is unknown.

The model test MAE is very small (20 meV) on the OQMD dataset, and a factor of two larger (38 meV) on the ICSD subset of OQMD. To test the model in a realistic materials screening setting, we created a dataset of 6000 selenides with very small overlap with the OQMD. The model pretrained on OQMD and applied to the selenides shows an MAE of 165 meV. This value can be lowered to 96 meV with an additional training on 100 selenides. Further training can lower the MAE to below

50 meV.

Based on the results we conclude, that is possible to develop ML models with position-independent descriptors, which are useful for realistic materials screening studies. However, extrapolation from OQMD to other datasets is challenging. One reason for this may be, as pointed out before, that the OQMD is composed of materials of two types: Some are generated systematically in rather few predefined crystal structures while others come from ICSD. (There is of course a significant overlap between the two types). The first type is characterized by a large variation in stability, but low variation in crystal structures, while the second type is the opposite: the experimentally observed materials in ICSD exhibit a large variation in crystal structures, but they are all (except for some high-pressure entries) stable low-energy materials. This bias might limit the extrapolation to datasets which contain structures weakly represented in OQMD and with element combinations, which are far from stable. One way forward could be to create datasets with less bias so that unstable materials are represented in a greater variety of structures.

We see a number of potential improvements of the proposed model. More symmetry information could be included using for example Wyckoff positions [22] or additional graph edges describing symmetry relations. Furthermore, it is possible to label the quotient graphs with crystal translation information so that the infinite graph can be reconstructed [39]. This would make the crystal description more unique.

Perhaps the model could also learn the atomic positions from the graph representation. The latest developments in generative models have succeeded in generating small molecules in 3D space [40]. By combining this kind of model with the restrictions imposed by the connectivity and symmetries described by the quotient graph (see for example [41, 42]) it might be possible to directly predict the atomic positions without running DFT relaxations.

Another useful extension would be to model uncertainties in the prediction. Even though the datasets used here have a relatively high number of entries they only contain a tiny fraction of the chemical space. If the model could learn what it does not know it would be very useful in an active learning setting where DFT calculations could be launched by the model to explore areas of the chemical space with high uncertainty. A promising direction for uncertainty modeling is to use ensembles of neural networks where different techniques can be considered to ensure diversity between ensemble members [43–46].

VIII. ACKNOWLEDGMENTS

We would like to thank Peter Mahler Larsen for helpful discussions. We acknowledge support from the VILLUM Center for Science of Sustainable Fuels and Chemicals which is funded by the VILLUM Fonden research grant (9455) and thanks to Nvidia for the donation of one Titan X GPU.

[1] I. E. Castelli, T. Olsen, S. Datta, D. D. Landis, S. Dahl, K. S. Thygesen, and K. W. Jacobsen, Energy Environ. Sci. **5**, 5814 (2012).

[2] Y. Wu, P. Lazic, G. Hautier, K. Persson, and G. Ceder, Energy & Environmental Science **6**, 157 (2012).

[3] K. Kuhar, A. Crovetto, M. Pandey, K. S. Thygesen, B. Seger, P. C. K. Vesborg, O. Hansen, I. Chorkendorff, and K. W. Jacobsen, Energy Environ. Sci. **10**, 2579 (2017).

[4] A. Urban, D.-H. Seo, and G. Ceder, npj Computational Materials **2**, 16002 (2016).

[5] M. Aykol, S. Kim, V. Hegde, D. Snydacker, Z. Lu, S. Hao, S. Kirklin, D. Morgan, and C. Wolverton, Nature Communications **7** (2016), 10.1038/ncomms13779, cited By 38.

[6] M. Andersson, T. Bligaard, A. Kustov, K. Larsen, J. Greeley, T. Johannessen, C. Christensen, and J. K. Nørskov, Journal of Catalysis **239**, 501 (2006).

[7] J. K. Nørskov, T. Bligaard, J. Rossmeisl, and C. H. Christensen, Nature Chemistry **1**, 37 (2009).

[8] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti, G. Pizzi, and N. Marzari, Nature Nanotechnology **13**, 246 (2018).

[9] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nature Materials **12**, 191 (2013).

[10] K. Alberi, M. B. Nardelli, A. Zakutayev, L. Mitas, S. Curtarolo, A. Jain, M. Fornari, N. Marzari, I. Takeuchi, M. L. Green, M. Kanatzidis, M. F. Toney, S. Butenko, B. Meredig, S. Lany, U. Kattner, A. Davydov, E. S. Toberer, V. Stevanovic, A. Walsh, N.-G. Park, A. Aspuru-Guzik, D. P. Tabor, J. Nelson, J. Murphy, A. Setlur, J. Gregoire, H. Li, R. Xiao, A. Ludwig, L. W. Martin, A. M. Rappe, S.-H. Wei, and J. Perkins, Journal of Physics D: Applied Physics **52**, 013001 (2018).

[11] P. Hohenberg and W. Kohn, Physical Review **136**, 864 (1964).

[12] W. Kohn and L. J. Sham, Physical Review **140**, 1133 (1965).

[13] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, Physical Review B **89**, 205118 (2014).

[14] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, International Journal of Quantum Chemistry **115**, 1094 (2015).

[15] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton, Phys. Rev. B Condens. Matter **96**, 024104 (2017).

[16] C. Oses, C. Toher, E. Gossett, A. Tropsha, O. Isayev, and S. Curtarolo, Nature communications **8**, 1 (2017).

[17] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Nat. Commun. **8**, 13890 (2017).

[18] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).

[19] T. Xie and J. C. Grossman, Physical Review Letters **120**, 145301 (2018).

[20] P. Zaspel, B. Huang, H. Harbrecht, and O. A. von Lilienfeld, Journal of Chemical Theory and Computation , acs.jctc.8b00832 (2019).

[21] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Phys. Rev. B Condens. Matter **89**, 094104 (2014).

[22] A. Jain and T. Bligaard, Phys. Rev. B Condens. Matter **98**, 214112 (2018).

[23] S. J. Chung, T. Hahn, and W. E. Klee, Acta Crystallogr. A **40**, 42 (1984).

[24] H. J. A. M. Heijmans and A. V. Tuzikov, IEEE Transactions on Pattern Analysis and Machine Intelligence **20**, 980 (1998).

[25] The graph fingerprints are computed using the neural message passing model with random weight initialization. We use two instances of neural network weight initialization and six different atomic embedding instances, thus having 12 models in total. The fingerprint is then a vector where each entry is the scalar output of one of these models.

[26] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, in *International Conference on Machine Learning* (2017) pp. 1263–1272.

[27] P. B. Jørgensen, K. W. Jacobsen, and M. N. Schmidt, (2018), arXiv:1806.03146 [stat.ML].

[28] K. He, X. Zhang, S. Ren, and J. Sun, (2015), arXiv:1512.03385 [cs.CV].

[29] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, APL Materials **1**, 011002 (2013).

[30] G. Bergerhoff, R. Hundt, R. Sievers, and I. D. Brown, J. Chem. Inf. Comput. Sci. **23**, 66 (1983).

[31] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).

[32] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, npj Computational Materials **1**, 15010 (2015).

[33] A. H. Larsen, J. J. Mortensen, *et al.*, Journal of Physics: Condensed Matter **29**, 273002 (2017).

[34] J. Enkovaara, C. Rostgaard, J. J. Mortensen, *et al.*, Journal of Physics: Condensed Matter **22**, 253202 (2010).

[35] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **100**, 136406 (2008).

[36] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[37] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).

[38] D. Kingma and J. Ba, (2014), arXiv:1412.6980 [cs.LG].

[39] W. E. Klee, Crystal Research and Technology **39**, 959 (2004).

[40] N. W. A. Gebauer, M. Gastegger, and K. T. Schütt, (2018), arXiv:1810.11347 [stat.ML].

[41] G. Thimm, Acta Crystallogr. A **65**, 213 (2009).

[42] J.-G. Eon, Acta Crystallogr. A **67**, 68 (2011).

[43] A. A. Peterson, R. Christensen, and A. Khorshidi, Phys. Chem. Chem. Phys. **19**, 10978 (2017).

[44] B. Lakshminarayanan, A. Pritzel, and C. Blundell, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017) pp. 6402–6413.

[45] T. Pearce, N. Anastassacos, M. Zaki, and A. Neely, (2018), arXiv:1805.11324 [stat.ML].

[46] I. Osband, J. Aslanides, and A. Cassirer, (2018), arXiv:1806.03335 [stat.ML].