

# General machine-learning surrogate models for materials prediction

Chandramouli Nyshadham,<sup>1</sup> Matthias Rupp,<sup>2</sup> Brayden Bekker,<sup>1</sup> Alexander V. Shapeev,<sup>3</sup> Tim Mueller,<sup>4</sup> Conrad W. Rosenbrock,<sup>1</sup> Gábor Csányi,<sup>5</sup> David W. Wingate,<sup>6</sup> and Gus L. W. Hart<sup>1</sup>

<sup>1</sup>*Department of Physics and Astronomy, Brigham Young University, Provo, UT 84602, USA*

<sup>2</sup>*Fritz Haber Institute of the Max Planck Society, Faradayweg 4–6, 14195 Berlin, Germany*

<sup>3</sup>*Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Building 3, Moscow, 143026, Russia*

<sup>4</sup>*Department of Materials Science and Engineering, Johns Hopkins University, Baltimore, MD 21218, USA*

<sup>5</sup>*Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom*

<sup>6</sup>*Computer Science Department, Brigham Young University, Provo, UT 84602, USA*

Surrogate machine-learning models are transforming computational materials science by predicting properties of materials with the accuracy of ab initio methods at a fraction of the computational cost. We demonstrate surrogate models that simultaneously interpolate energies of different materials on a dataset of 10 binary alloys (AgCu, AlFe, AlMg, AlNi, AlTi, CoNi, CuFe, CuNi, FeV, NbNi) with 10 different species and all possible fcc, bcc and hcp structures up to 8 atoms in the unit cell, 15 950 structures in total. We find that prediction errors remain unaffected when increasing the number of simultaneously modeled alloys. Several state-of-the-art materials representations and learning algorithms were found to qualitatively agree, with prediction errors as low as 1 meV/atom.

## I. INTRODUCTION

Advances in computational power and electronic structure methods have enabled large materials databases<sup>1–4</sup>. Using high-throughput approaches,<sup>5</sup> these databases have proven a useful tool to predict properties of materials. However, given the combinatorial nature of materials space,<sup>6,7</sup> it is infeasible to compute properties for more than a tiny fraction of all possible materials using electronic structure methods such as density functional theory<sup>8,9</sup>. A potential answer to this challenge lies in a new paradigm: surrogate machine-learning models for accurate materials predictions.<sup>10–12</sup>

The key idea is to use machine learning to rapidly and accurately interpolate between reference simulations, effectively mapping the problem of numerically solving for the electronic structure of a material onto a statistical regression problem.<sup>13</sup> Such fast surrogate models could be used to filter the most suitable materials from a large pool of possible materials and then validate the found subset by electronic structure calculations. Such an “accelerated high-throughput” (AHT) approach (Figure 1) could potentially increase the number of investigated materials by several orders of magnitude.

Traditionally, empirical interatomic potentials were used to reproduce macroscopic properties of materials faster than DFT. Well-known empirical interatomic potentials for periodic solids include Lennard-Jones potentials, the Stillinger-Weber potential and Embedded Atom Methods (EAM) for alloys. A problem with empirical interatomic potentials is that they are designed with a fixed number of parameters and cannot be systematically improved. In contrast, surrogate models which are empirical interatomic models based on machine learning, systematically improve with additional data.

We demonstrate the feasibility of accurate *general* (simultaneously interpolate energies of multiple different materials) surrogate machine-learning models for en-

thalpies of formation of materials across the composition, lattices, and structures. We find that five combinations of state-of-the-art representations and regression methods (Table I) all yield consistent predictions with errors in the low meV/atom range.

## II. BACKGROUND

A surrogate machine-learning model replaces ab initio simulations by mapping a crystal structure to properties such as enthalpies, elastic constants or band gaps. Their importance lies in the fact that once the model is trained, properties of new materials can be predicted in time either constant or linear in the size of the material, with low pre-factor, typically in milliseconds.

The two major parts of a surrogate machine-learning model are the numerical representation of the input data<sup>11,21</sup> and the learning algorithm. We use the term “representation” for a set of features (as opposed to a collection of unrelated or only loosely related descriptors) that satisfies certain physical requirements<sup>12,13,18,22</sup> such as invariance to translation, rotation, and permutation of atoms, uniqueness<sup>23</sup>, differentiability, and computational efficiency. The role of the representation is akin to that of a basis set in that the predicted property is expanded in terms of a set of reference structures.

The materials community has proposed several representations<sup>10–12,18,21,24–27</sup> for crystal structures. Many do not fulfill the above properties or are restricted in practice to materials with a few different chemical elements. Consequently, surrogate models based on these representations are limited in their accuracy, due to the violation of the uniqueness requirement<sup>22,28</sup>.

We explore three state-of-the-art representations that fulfill above properties for construction of general surrogate models: Many-body tensor representation<sup>12</sup> (MBTR), smooth overlap of atomic positions<sup>10,18</sup> (SOAP) and moment tensor potentials<sup>11</sup> (MTP). Each

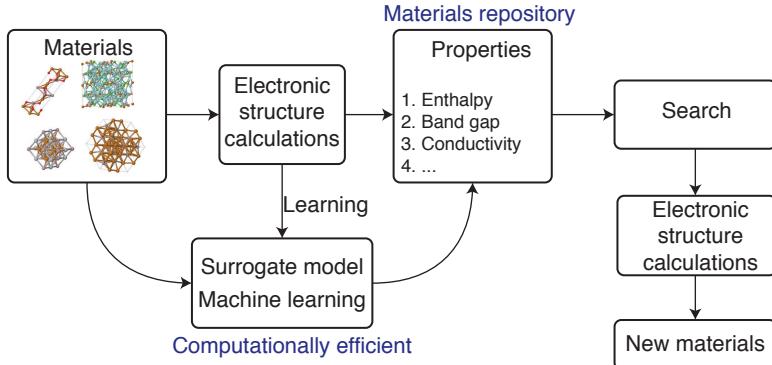


Figure 1. *The accelerated high-throughput approach.* Candidate structures and properties are generated by surrogate machine-learning models based on reference electronic structure calculations in a materials repository. Selected structures are validated by electronic structure calculations, preventing false positive errors.

Table I. *State-of-the-art surrogate machine-learning models* investigated in this work.

| Abbrv.       | Surrogate model  | Description  |
|--------------|--|--|
| CE           | Cluster Expansion <sup>14–17</sup>   | One of the early successful surrogate models developed in the materials community. A material’s ground state energy is expanded as an Ising-type model with constant expansion coefficients. |
| MBTR<br>+KRR | Many-Body Tensor Representation <sup>12</sup> + Kernel Ridge Regression                      | Materials are expanded in distributions of $k$ -body terms stratified by chemical element species, using non-linear regression.  |
| MBTR<br>+DNN | Many-Body Tensor Representation + Deep Neural Network  | MBTR is used as input for DNN to learn a new representation and predict using a parametric deep regression method.   |
| SOAP<br>+GPR | Smooth Overlap of Atomic Positions <sup>18</sup> + Gaussian Process Regression <sup>10</sup> | Atomic environments represented as smoothed Gaussian densities of neighboring atoms expanded in a spherical harmonics basis, using non-parametric regression.                                |
| MTP          | Moment Tensor Potentials (MTP) <sup>11,19,20</sup> + Polynomial Regression                   | Atomic environments expanded in a tailored polynomial basis, computed via contractions of moment tensors.  |

representation is employed as proposed and implemented by its authors, including the regression method: Kernel ridge regression<sup>13</sup> (KRR) for MBTR, Gaussian process regression regression<sup>29</sup> (GPR) for SOAP, and polynomial regression<sup>11</sup> for MTP. Since predictions (but not necessarily other properties) of the kernel-based KRR and GPR are identical, we will use the two terms interchangeably here. We also employed cluster expansion<sup>14–17</sup> (CE) and Deep Neural Network<sup>30,31</sup> (DNN) models.

CE models have been used for three decades to efficiently model ground state energies of metal alloys, but require that the atomic structure can be mapped to site occupancies on a fixed lattice. They are therefore less suited to model different materials. In this work, we use them as a baseline and build a separate CE model for each alloy.

DNNs are essentially recursively stacked layers of functions, the large number of layers being the major difference to conventional neural networks. They have been used to predict energies<sup>32–36</sup> and to learn representations.<sup>37,38</sup> While DNNs can learn representations (“end-to-end learning”, here from nuclear charges, atom positions and unit cell basis vectors to enthalpy of formation), this requires substantially more data than

starting with a representation as input<sup>24–26</sup>. We, therefore, provide the DNN with MBTR as input. The idea of using MBTR+DNN is to explore whether a representation-learning technique can improve upon a manually designed representation in conjunction with the standard Gaussian kernel (MBTR+KRR).

### III. MATERIALS AND METHODS

#### A. Data

We created a dataset (DFT-10B) containing structures of the 10 binary alloys AgCu, AlFe, AlMg, AlNi, AlTi, CoNi, CuFe, CuNi, FeV, NbNi. Each alloy system includes all possible unit cells with 1–8 atoms for face-centered cubic (fcc) and body-centered cubic (bcc) crystal types, and all possible unit cells with 2–8 atoms for the hexagonal close-packed (hcp) crystal type. This results in 631 fcc, 631 bcc and 333 hcp structures, yielding  $1595 \cdot 10 = 15\,950$  unrelaxed structures in total. Crystal structures were generated using the enumeration algorithm by Hart and Forcade<sup>39</sup>.

Lattice parameters for each crystal structure were

set according to Vegard's law.<sup>40,41</sup> Total energies were computed using density functional theory (DFT) with projector-augmented wave (PAW) potentials<sup>42–44</sup> within the generalized gradient approximation (GGA) of Perdew, Burke, and Ernzerhof<sup>45</sup> (PBE) as implemented in the Vienna Ab Initio Simulation Package<sup>46,47</sup> (VASP). The  $k$ -point meshes for sampling the Brillouin zone were constructed using generalized regular grids.<sup>48,49</sup> The details of the  $k$ -point density for all 10 alloys is mentioned in the appendix (table IV).

## B. Models

All single-alloy surrogate models were trained using 1000 randomly selected crystal structures and validated on the remaining 595 structures. Models trained on multiple alloys use the union of the individual alloy's splits. Parametrization details of all surrogate models used in this work can be found in the appendix.

## IV. RESULTS AND DISCUSSION

### A. Energy predictions for single alloys

Prediction errors for enthalpies of formation of each of the five surrogate models on each binary alloy subset of the data are presented in Figure 2. Prediction errors of all surrogate models agree qualitatively on all subsets of the data. We interpret this consistency to be indicative of the validity of the machine learning approach to surrogate models of materials properties, independently of the parametrization details of the models.

For four binary systems (AgCu, AlMg, CoNi, CuNi) predictions errors are at or below 3 meV/atom. The prediction errors of all surrogate models on the remaining six systems (AlFe, AlNi, AlTi, CuFe, FeV, NbNi) are consistent, and it is not obvious as to why these systems are harder to learn. When generating the data, the same methodology and parameters were used for all alloys, and similar fitting procedures were employed for each surrogate model.

We point out that whenever the elements that constitute a binary alloy system belong to the same column of the periodic table or are close to each other in the periodic table in terms of atomic number, the surrogate models' predictions are good and vice versa. Indeed, together these numbers explain 80 % of the variance in prediction errors (Figure 5). A complementary observation is that while absolute errors vary from alloy to alloy, relative errors ( $\delta_{\text{RMSE}}$ ), expressed as a percentage of the range of energies of an alloys' subset of the data, remains essentially constant at around 1 %.

### B. General models trained on all alloys

Surrogate models based on analytic expressions need to be fitted to each alloy system separately. For CE, the representation is naturally tied to a particular lattice (e.g. fcc, bcc), making it difficult to train on multiple alloy systems with different lattices at the same time. Here we train a cluster expansion on all alloys by constraining all 30 systems to use a single set of hyperparameters for regularization (i.e. all use the same prior probability distribution of ECI values). The machine-learning surrogate models based on MBTR, SOAP, and MTP do not suffer from the problem of representation being tied to a particular lattice. They express energy as a continuous function of distances and can be trained on multiple materials simultaneously.

We trained four of the five investigated surrogate models simultaneously on all 10 alloy systems and compared the mean absolute error (MAE) of these combined models with the average MAE when trained on each alloy system separately (Table II; note that RMSE would differ due to its non-linear nature). The quantitative agreement indicates that prediction errors remain unaffected when training on multiple systems. For the cluster expansion, these results suggest that there is a single set of parameters for generating a prior probability distribution over ECI values (provided in the supporting information) that works well across a variety of chemistries and lattice types.

We investigate simultaneous training of alloys in more detail for the MBTR+KRR model. Figure 3 presents deviations of the MAE of a single model trained on  $k$  alloy systems from the average MAE when the model is trained on each alloy system separately. In all of the possible  $\sum_{k=1}^{10} \binom{10}{k} = 1023$  cases the deviation is below 1 meV/atom. These deviations are on the order one would expect from minor differences in hyperparameter values. We conclude that prediction errors remain consistently unaffected when increasing the number of simultaneously modeled alloys.

Table II. *Performance of general models.* Shown are mean absolute errors (MAE) of models trained on all 10 alloy systems simultaneously (right column) versus the average MAE of models trained on individual alloy systems. The combined fit using SOAP+GP was not performed in this work.

| Surrogate Model | Mean Absolute Errors (meV/atom) |                |
|-----------------|---------------------------------|----------------|
|                 | Average of Separate Models      | Combined Model |
| CE              | 4.7                             | 4.8            |
| MBTR+KRR        | 5.1                             | 5.3            |
| MBTR+DNN        | 5.1                             | 4.6            |
| SOAP+GP         | 4.5                             | —              |
| MTP             | 3.1                             | 3.4            |

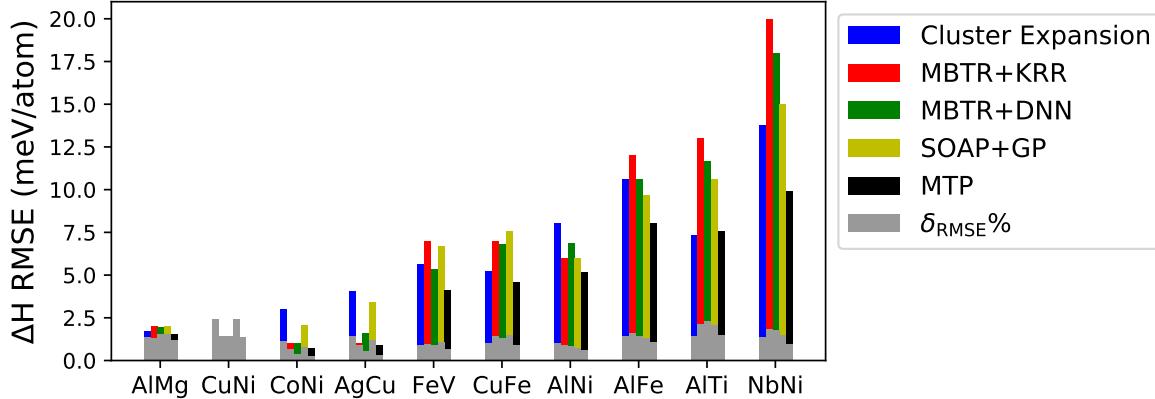


Figure 2. Comparison of five machine learning surrogate models on the DFT10-B dataset. Shown are the root mean squared error (RMSE) of predicted enthalpies of formation of each surrogate model on each binary alloy subset in meV/atom (colored bars) and as a percentage of energy range (gray bars). RMSE for MTP results is computed using pure atom total energies obtained from DFT. The consistency of errors across models indicates the validity of machine learning surrogate models to predict properties of materials—prediction errors are similar, independent of details of model parametrization. Note that relative error varies around 1 % for all datasets.

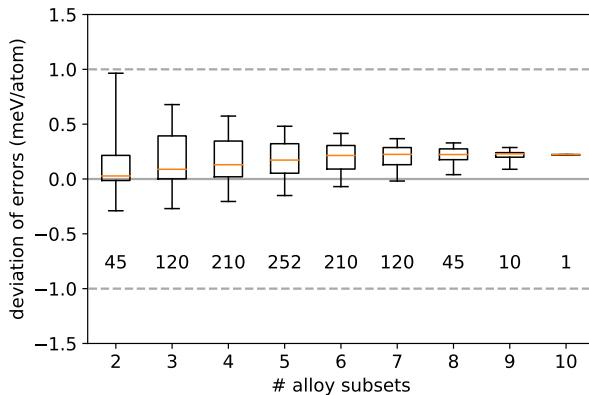


Figure 3. Performance of MBTR+KRR model for multiple alloy systems. Shown are deviation of mean absolute error (MAE, vertical axis) of an MBTR+KRR surrogate model trained on  $k$  (horizontal axis) alloy systems simultaneously from the average MAE of  $k$  models trained on each alloy subsystem separately. Whiskers, boxes, horizontal line and numbers inside the plot show the range of values, quartiles, median and sample size, respectively. Difference in error between individual and combined models is always less than 1 meV/atom.

### C. Caveat emptor

Are reported errors reliable estimates of future performance in applications? It depends. We discuss the role of training and validation set composition as an example of the intricacies of statistical validation of machine learning models.

In the limit of infinite independent and identically distributed data, one would simply sample a large enough

validation set and measure prediction errors, with the law of large numbers ensuring the reliability of the estimates. Here, however, data are limited due to the costs of generating them via ab initio simulations, and are neither independent nor identically distributed. In such a setting, part of the available data is used for validation, either in the form of a hold-out set (as in this work) or via cross-validation, suited for even smaller datasets.

Prediction errors in machine learning models improve with data (otherwise it would not be machine *learning*). This implies that if only few training samples exist for a “subclass” of structures, prediction errors for similar structures will be high. For example, consider the number of atoms per unit cell in the DFT-10B dataset used here: There are only 11 structures for each alloy that have 1 or 2 atoms in the unit cell. Consequently, prediction errors are high for those (Figure 6).

For combinatorial reasons, the number of possible structures increases strongly with the number of atoms in the unit cell (Table III). This biases error statistics in two ways: As discussed, prediction errors will be lower for classes with more samples. At the same time, because these classes have more samples, they will contribute more to the measured errors, dominating averages such as the RMSE.

Table III. Size distribution in the DFT-10B dataset. Shown are the number of structures with  $k$  atoms in the unit cell,  $k \leq 10$  (per alloy; multiply by 10 for the total dataset).

| atoms/unit cell | 1 | 2 | 3  | 4  | 5  | 6   | 7   | 8    |
|-----------------|---|---|----|----|----|-----|-----|------|
| #structures     | 4 | 7 | 12 | 48 | 56 | 210 | 208 | 1050 |

Figure 4 presents MBTR+KRR prediction errors (RMSE in meV/atom) for different but same-size splits

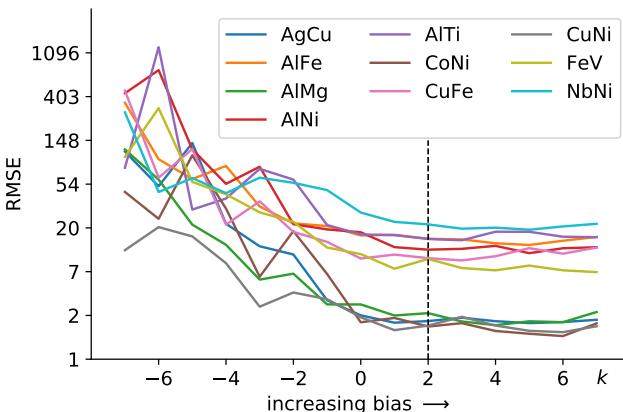


Figure 4. *Influence of biased training and validation sets.* Shown are the root mean squared errors (meV/atom) as a function of training and validation set composition obtained using MBTR+KRR model. See main text for discussion.

of the data into training and validation sets. On the left, all structures with  $|k|$  or fewer atoms in the unit cell are excluded from the training set (and therefore included in the validation set). This results in many high-error structures in the validation set, with the effect decreasing for smaller  $|k|$ . For  $k = 0$ , size does not influence the split. On the right, structures with  $\leq k$  atoms are always included in the training set, resulting in fewer high-error structures in the validation set. The dashed line marks the value of  $k = 2$  recommended in this work (see also Figure 6).

Retrospective errors reported in the literature should, therefore, be critically assessed. The design of such studies should favor “representative” validation sets over those tweaked to yield lowest possible errors to report. For combinatorial datasets, the smallest structures (those that can be considered to be outliers) should be included in the training set.<sup>50</sup>

## V. CONCLUSION AND OUTLOOK

We showed that it is possible to use machine learning to build *general* surrogate models that can predict the enthalpy of formation of crystal structures across 10 different binary alloy systems, for three lattice types (fcc, bcc, hcp) and for structures not in their ground state. In this, we find that the concept of using machine learning to predict materials properties is independent of the details of the used surrogate models as predictions of several state-of-the-art materials representations and learning algorithms were found to be in qualitative agreement. This observation also seems to be congruent with recent efforts towards a unifying mathematical framework for some of the used representations.<sup>51</sup>

The ability to use a single general surrogate model has the potential to greatly simplify the use of surrogate models for exploration of materials spaces by avoiding the need to identify “homogeneous” subspaces and then building separate models for each of them. This also avoids problems such as discontinuities at the boundaries of separate models.

Is it possible to do better? Recent results suggest that it might be possible to exploit similarities between chemical element species to improve learning rates further.<sup>52</sup> This requires either to explicitly account for element similarities in the representations or to learn element similarities from the data, for example with a DNN. While such alchemical learning is outside of the scope of this work, we do observe an improvement in prediction errors for the general MBTR+DNN model (Table II, Figure 7).

## ACKNOWLEDGMENTS

CN is thankful to Kennedy Lincoln and Wiley Morgan for insightful discussions. CN, BB, CR, and GH acknowledge the funding from ONR (MURI N00014-13-1-0635). MR acknowledges funding from the EU Horizon 2020 program Grant 676580, The Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence. AS was supported by the Russian Science Foundation (Grant No 18-13-00479). TM acknowledges funding from the National Science Foundation under award number DMR-1352373 and computational resources provided by the Maryland Advanced Research Computing Center (MARCC).

<sup>1</sup> Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H. Taylor, Lance J. Nelson, Gus L.W. Hart, Stefano Sanvitto, Marco Buongiorno-Nardelli, Natalio Mingo, and Ohad Levy. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput *ab initio* calculations. *Comput. Mater. Sci.*, 58:227–235, 2012. doi:10.1016/j.

commatsci.2012.02.002.

<sup>2</sup> James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Chris Wolverton. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *J. Miner. Met. Mater. Soc.*, 65(11):1501–1509, 2013. doi:10.1007/s11837-013-0755-4.

- <sup>3</sup> Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.*, 1:011002, 2013. doi:10.1063/1.4812323.
- <sup>4</sup> Claudia Draxl and Matthias Scheffler. NOMAD: The FAIR concept for big-data-driven materials science. *MRS Bull.*, to appear, 2018. URL: <https://arxiv.org/abs/1805.05039>.
- <sup>5</sup> Stefano Curtarolo, Gus L. W. Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, and Ohad Levy. The high-throughput highway to computational materials design. *Nat. Mater.*, 12(3):191–201, 2013. doi:10.1038/nmat3568.
- <sup>6</sup> Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications*, 8:15679, 2017. doi:10.1038/ncomms15679.
- <sup>7</sup> Aron Walsh. Inorganic materials: The quest for new functionality. *Nature chemistry*, 7(4):274, 2015. doi:10.1038/nchem.2213.
- <sup>8</sup> P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964. URL: <https://link.aps.org/doi/10.1103/PhysRev.136.B864>, doi:10.1103/PhysRev.136.B864.
- <sup>9</sup> W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965. URL: <https://link.aps.org/doi/10.1103/PhysRev.140.A1133>, doi:10.1103/PhysRev.140.A1133.
- <sup>10</sup> Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104:136403, Apr 2010. URL: <https://link.aps.org/doi/10.1103/PhysRevLett.104.136403>, doi:10.1103/PhysRevLett.104.136403.
- <sup>11</sup> Alexander V. Shapeev. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.*, 14(3):1153–1173, 2016. doi:10.1137/15M1054183.
- <sup>12</sup> Haoyan Huo and Matthias Rupp. Unified representation for machine learning of molecules and materials. *arXiv*, page 1704.06439v3, 2017. URL: <https://arxiv.org/abs/1704.06439>.
- <sup>13</sup> Matthias Rupp. Machine learning for quantum mechanics in a nutshell. *Int. J. Quant. Chem.*, 115(16):1058–1073, 2015. doi:10.1002/qua.24954.
- <sup>14</sup> Juan M. Sanchez, François Ducastelle, and Denis Grattas. Generalized cluster description of multicomponent systems. *Phys. Stat. Mech. Appl.*, 128(1–2):334–350, 1984. doi:10.1016/0378-4371(84)90096-7.
- <sup>15</sup> Didier De Fontaine. Cluster approach to order-disorder transformations in alloys. In Henry Ehrenreich and David Turnbull, editors, *Solid State Physics*, volume 47, pages 33–176. Elsevier, 1994. doi:10.1016/S0081-1947(08)60639-6.
- <sup>16</sup> Chris G. van de Walle and Gerbrand Ceder. Automating first-principles phase diagram calculations. *J. Ph. Equilib.*, 23(4):348–359, 2002. doi:10.1361/105497102770331596.
- <sup>17</sup> Tim Mueller and Gerbrand Ceder. Bayesian approach to cluster expansions. *Phys. Rev. B*, 80(2):024103, 2009. doi:10.1103/PhysRevB.80.024103.
- <sup>18</sup> Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87(18):184115, 2013. doi:10.1103/PhysRevB.87.184115.
- <sup>19</sup> Konstantin Gubaev, Evgeny V Podryabinkin, and Alexander V Shapeev. Machine learning of molecular properties: Locality and active learning. *The Journal of Chemical Physics*, 148(24):241727, 2018.
- <sup>20</sup> Konstantin Gubaev, Evgeny V Podryabinkin, Gus LW Hart, and Alexander V Shapeev. Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. *arXiv preprint arXiv:1806.10567*, 2018.
- <sup>21</sup> Kristof T. Schütt, Henning Glawe, Felix Brockherde, Antonio Sanna, Klaus-Robert Müller, and Eberhard K.U. Gross. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B*, 89(20):205118, 2014. doi:10.1103/PhysRevB.89.205118.
- <sup>22</sup> O. Anatole von Lilienfeld, Raghunathan Ramakrishnan, Matthias Rupp, and Aaron Knoll. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *Int. J. Quant. Chem.*, 115(16):1084–1093, 2015. doi:10.1002/qua.24912.
- <sup>23</sup> Representation is variant against transformations changing the property, as systems with identical representation but differing in property would introduce errors<sup>28</sup>.
- <sup>24</sup> Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98(14):146401, 2007. doi:10.1103/PhysRevLett.98.146401.
- <sup>25</sup> Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.*, 134(7):074106, 2011. doi:10.1063/1.3553717.
- <sup>26</sup> Jörg Behler. Representing potential energy surfaces by high-dimensional neural network potentials. *J. Phys. Condens. Matter*, 26(18):183001, 2014. doi:10.1088/0953-8984/26/18/183001.
- <sup>27</sup> Felix Faber, Alexander Lindmaa, O. Anatole von Lilienfeld, and Rickard Armiento. Crystal structure representations for machine learning models of formation energies. *Int. J. Quant. Chem.*, 115(16):1094–1101, 2015. doi:10.1002/qua.24917.
- <sup>28</sup> Jonathan E. Moussa. Comment on “Fast and accurate modeling of molecular atomization energies with machine learning”. *Phys. Rev. Lett.*, 109(5):059801, 2012. doi:10.1103/PhysRevLett.109.059801.
- <sup>29</sup> Carl Rasmussen and Christopher Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, 2006. URL: <http://www.gaussianprocess.org/gpml>.
- <sup>30</sup> Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015. doi:10.1038/nature14539.
- <sup>31</sup> Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015. doi:10.1016/j.neunet.2014.09.003.
- <sup>32</sup> Kristof T. Schütt, Huziel E. Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet—a deep learning architecture for molecules and materials. *J. Chem. Phys.*, 148(24):241722, 2018. doi:10.1063/1.5019779.

- <sup>33</sup> Nicholas Lubbers, Justin S. Smith, and Kipton Barros. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.*, 148(24):241715, 2018. doi:[10.1063/1.5011181](https://doi.org/10.1063/1.5011181).
- <sup>34</sup> Kyle Mills, Michael Spanner, and Isaac Tamblyn. Deep learning and the schrödinger equation. *Phys. Rev. A*, 96(4):042113, 2017. doi:[10.1103/PhysRevA.96.042113](https://doi.org/10.1103/PhysRevA.96.042113).
- <sup>35</sup> Justin S. Smith, Olexandr Isayev, and Adrian E. Roitberg. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.*, 8(4):3192–3203, 2017. doi:[10.1039/C6SC05720A](https://doi.org/10.1039/C6SC05720A).
- <sup>36</sup> Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nat. Comm.*, 8:13890, 2017. doi:[10.1038/ncomms13890](https://doi.org/10.1038/ncomms13890).
- <sup>37</sup> Matthew K. Matlock, Na Le Dang, and S. Joshua Swamidas. Learning a local-variable model of aromatic and conjugated systems. *ACS Cent. Sci.*, 4(1):52–62, 2018. doi:[10.1021/acscentsci.7b00405](https://doi.org/10.1021/acscentsci.7b00405).
- <sup>38</sup> Xun Gao and Lu-Ming Duan. Efficient representation of quantum many-body states with deep neural networks. *Nat. Commun.*, 8:662, 2017. doi:[10.1038/s41467-017-00705-2](https://doi.org/10.1038/s41467-017-00705-2).
- <sup>39</sup> Gus L. W. Hart and Rodney W. Forcade. Algorithm for generating derivative structures. *Phys. Rev. B*, 77(22):224115, 2008. doi:[10.1103/PhysRevB.77.224115](https://doi.org/10.1103/PhysRevB.77.224115).
- <sup>40</sup> Lars Vegard. Die Konstitution der Mischkristalle und die Raumfüllung der Atome. *Z. Physik*, 5(1):17–26, 1921. doi:[10.1007/BF01349680](https://doi.org/10.1007/BF01349680).
- <sup>41</sup> Alan R. Denton and Neil W. Ashcroft. Vergard's law. *Phys. Rev. A*, 43(6):3161, 1991. doi:[10.1103/PhysRevA.43.3161](https://doi.org/10.1103/PhysRevA.43.3161).
- <sup>42</sup> Georg Kresse and Daniel Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B*, 59(3):1758, 1999. doi:[10.1103/PhysRevB.59.1758](https://doi.org/10.1103/PhysRevB.59.1758).
- <sup>43</sup> Peter E. Blöchl. Projector augmented-wave method. *Phys. Rev. B*, 50(24):17953, 1994. doi:[10.1103/PhysRevB.50.17953](https://doi.org/10.1103/PhysRevB.50.17953).
- <sup>44</sup> Georg Kresse and Jürgen Hafner. Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements. *J. Phys. Condens. Matter*, 6(40):8245–8257, 1994. doi:[10.1088/0953-8984/6/40/015](https://doi.org/10.1088/0953-8984/6/40/015).
- <sup>45</sup> John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77(18):3865–3868, 1996. doi:[10.1103/PhysRevLett.77.3865](https://doi.org/10.1103/PhysRevLett.77.3865).
- <sup>46</sup> Georg Kresse and Jürgen Furthmüller. Efficiency of *ab initio* total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.*, 6(1):15–50, 1996. doi:[10.1016/0927-0256\(96\)00008-0](https://doi.org/10.1016/0927-0256(96)00008-0).
- <sup>47</sup> Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54(16):11169, 1996. doi:[10.1103/PhysRevB.54.11169](https://doi.org/10.1103/PhysRevB.54.11169).
- <sup>48</sup> Pandu Wisesa, Kyle A. McGill, and Tim Mueller. Efficient generation of generalized Monkhorst-Pack grids through the use of informatics. *Phys. Rev. B*, 93(15):155109, 2016.
- <sup>49</sup> Wiley S Morgan, Jeremy J Jorgensen, Bret C Hess, and Gus LW Hart. Efficiency of generalized regular\kb-point grids. *arXiv*, page 1804.04741, 2018. URL: <https://arxiv.org/abs/1804.04741>.
- <sup>50</sup> Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108(5):058301, 2012. doi:[10.1103/PhysRevLett.108.058301](https://doi.org/10.1103/PhysRevLett.108.058301).
- <sup>51</sup> Michael J. Willatt, Félix Musil, and Michele Ceriotti. Theory and practice of atom-density representations for machine learning. *arXiv*, 2018. URL: <https://arxiv.org/abs/1807.00408>.
- <sup>52</sup> Felix A Faber, Anders S Christensen, Bing Huang, and O Anatole von Lilienfeld. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of Chemical Physics*, 148(24):241717, 2018. doi:<https://doi.org/10.1063/1.5020710>.
- <sup>53</sup> Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from [tensorflow.org](https://www.tensorflow.org/). URL: <https://www.tensorflow.org/>.
- <sup>54</sup> Noam Bernstein, Gábor Csányi, James Kermode, Albert Bartok-Partay, Livia Bartok-Partay, Federico Bianchini, Anke Butenuth, Marco Caccin, Silvia Cereda, Gabor Csanyi, Alessio Comisso, Tom Daff, ST John, Chiara Gattinoni, Gianpietro Moras, James Kermode, Letif Mones, Alan Nichol, Lars Pastewka David Packwood, Giovanni Peralta, Ivan Solt, Oliver Strickson, Wojciech Szlachta, Csilla Varnai, and Steven Winfield. Quip, 2008–2018. Software available for non-commercial use from [libatoms.org](http://www.libatoms.org/). URL: <http://www.libatoms.org/>.

## SUPPLEMENTARY MATERIAL

### A. Method details

#### 1. Cluster expansion

To determine the cutoff distances for the cluster expansions and determine the initial parameters for the prior probability distributions, we used a length scale in which the edge of a bcc unit cell is 1 unit of length and assumed the hcp, bcc, and fcc crystal structures all had the same nearest-neighbor distance. The cutoff distances used to determine the set of clusters included in the expansion are as follows:

| Number of sites in cluster | Maximum distance between sites |
|----------------------------|--------------------------------|
| 2                          | 8                              |
| 3                          | 4                              |
| 4                          | 2                              |
| 5                          | 1.5                            |
| 6                          | 1.5                            |

This resulted in a total of 791, 941, and 2870 distinct orbits of clusters in the bcc, fcc, and hcp expansions, respectively, including the empty cluster. These numbers were reduced after fitting by “trimming” the cluster expansions, in which cluster functions with very small ECI were removed from the expansion. To determine which clusters to remove, we used the fact that when the cluster functions are orthonormal, the expected squared error due to truncation,  $E(\text{error}^2)$ , is given by

$$E(\text{error}^2) = \sum_b V_b^2, \quad (1)$$

where  $V_b$  is the ECI for the  $b$ -th cluster function, the expectation of the squared error on the left is over all possible lattice decorations, and the sum on the right is over cluster functions excluded from the expansion. Thus removing an orbit of clusters with multiplicity  $m_b$  increases the expected squared error by  $m_b V_b^2$ . To trim clusters from the expansion with little loss of accuracy, we removed all orbits of cluster functions for which  $\sqrt{m_b V_b^2} < 10^{-5}$  eV. The trimming procedure changed the final average root-mean-squared prediction errors on the training sets by less than  $10^{-5}$  eV / atom and removed on average more than 70% of the ECIs in the expansions.

The ECIs for the cluster expansions were fit to the training data using the Bayesian approach with a multivariate Gaussian prior distribution<sup>17</sup>. The inverse of the covariance matrix for the prior,  $\Lambda$ , was diagonal, with elements given by

$$\lambda_{\alpha\alpha} = \begin{cases} 0 & \text{for } n_\alpha = 0 \\ e^{-\lambda_1} & \text{for } n_\alpha = 1 \\ e^{-\lambda_2} e^{-\lambda_3 r_\alpha} n_\alpha^{\lambda_4} & \text{for } n_\alpha > 1 \end{cases}, \quad (2)$$

where  $n_\alpha$  is the number of sites in cluster function  $\alpha$  and  $r_\alpha$  is the maximum distance between sites in Angstroms. The parameters  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$ , and  $\lambda_4$  were initially set to 10, 10, 5, and 5 respectively then optimized by using a conjugate gradient algorithm to minimize the root mean square leave-one-out cross-validation error, an estimate of prediction error<sup>16</sup>. For the combined fit, in which a single set of regularization parameters were used for all 30 cluster expansions, the optimized values of  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$ , and  $\lambda_4$  were 10.0, 20.8, 4.2, and 15.3 respectively.

#### 2. MBTR+KRR

The Many-Body Tensor Representation (MBTR) numerically represents atomistic systems as distributions of many-body terms, such as atom counts, distances, and angles, stratified (separated) by chemical elements. For details please consult Ref. 12. Kernel ridge regression<sup>13</sup> with a Gaussian kernel was employed throughout. In this work, we use the following parametrization:

| $k$ | geometry   | weighting   | discretization           | $\sigma$   |
|-----|------------|-------------|--------------------------|------------|
| 2   | 1/distance | identity^2  | (0, 0.005, 90)           | $2^{-17}$  |
| 2   | 1/distance | identity^2  | (0, 0.005, 90)           | $2^{-4.5}$ |
| 3   | angle      | 1/dotdotdot | (−0.15, $\pi/100$ , 100) | $2^{-14}$  |

We did not use 1-body terms as enthalpies of formation are the result of a linear operation in atom counts already. Values for the  $\sigma$  hyperparameter above refer to Figure 3, where we used fixed hyperparameter values (Gaussian kernel  $\sigma = 2^7$ , KRR regularization strength  $\lambda = 2^{-20}$ ). For individual models, hyperparameters were optimized on a base-2 logarithmic grid.

#### 3. MBTR+DNN

The mathematical details of the many-body tensor representation for the crystal structures are mentioned in ref.<sup>12</sup>. Each crystal structure is expanded in terms of distributions ( $k$ -body terms) of atom counts, (inverse) distances and angles. The Gaussian kernel with a variance ( $\sigma$ ) of 11.3 was used for fitting. Each MBTR vector is 1450 long and was optimized using a grid search. The details of the weighting functions, smearing parameters for each  $k$ -body term are as follows,

| $k$ | geometry   | weighting   | discretization   | $\sigma$  |
|-----|------------|-------------|------------------|-----------|
| 1   | atom count | 1/identity  | (0.5, 1, 25)     | $10^{-4}$ |
| 2   | 1/distance | identity^2  | (0.1, 0.005, 70) | $2^{-17}$ |
| 3   | angle      | 1/dotdotdot | (0.1, 0.05, 140) | $2^{-8}$  |

MBTR+DNN model uses the same parameters as MBTR+KRR model for generating the representation. The only difference between the representations is that the  $k$ -body terms in MBTR+DNN model are stratified by all 10 elements instead of just two. This results in a

representation vector which is 147100 long. The architecture of the convolution neural network used in this work is listed in the table below.

| Layer type                   | Specifications                    |
|------------------------------|-----------------------------------|
| Fully connected layer        | (Size: 2048)                      |
| Fully connected layer        | (Size: 1024)                      |
| Reshaping data               | (Size: 4 x 4 x 64)                |
| Convolution transposed layer | (Kernel: 5 x 5, 64 filters)       |
| Convolution layer            | (Kernel: 3 x 3, 64 filters)       |
| Max pooling layer            | (Pool size: 2 x 2; stride: 2 x 2) |
| Convolution layer            | (Kernel: 3 x 3, 32 filters)       |
| Reshaping data               | (Size: 1 x 1024)                  |
| Fully connected layer        | (Size: 128)                       |
| Fully connected layer        | (Size: 64)                        |
| Fully connected layer        | (Size: 4)                         |
| Fully connected layer        | (Output; size: 1)                 |

The DNN code is implemented using the Tensorflow framework<sup>53</sup>. The models were trained with a mini-batch size of 50 and the RMSE error is used as the cost function for optimizing the weights of the network.

#### 4. SOAP+GP

GAP fits were generated for each alloy system using a 2-body + SOAP approach. The standard deviation (SD, parameter  $\delta$ ) of the Gaussian process for the 2-body GAP is set to match the SD in energies of the training set. After fitting the 2-body potential, another SOAP GAP is fit with its SD set to match the remaining RMSE of the 2-body GAP relative to the DFT energies in the training set. The fits were performed using `teach_sparse`<sup>54</sup> with the following parameters for the 2-body GAP:

| Parameter     | Value |
|---------------|-------|
| Cutoff        | 6.0 Å |
| Sparse points | 10    |

and for SOAP, parameters were set to:

| Parameter              | Value |
|------------------------|-------|
| Cutoff                 | 4.5 Å |
| Sparse points          | 500   |
| $l_{\max}$             | 8     |
| $n_{\max}$             | 8     |
| $\zeta$                | 2     |
| $\sigma_{\text{atom}}$ | 5     |

As described above,  $\delta$  is set using the standard deviation of the Gaussian Process based on the training set for the 2-body and SOAP fits respectively. The following table lists these values for each of the alloy systems.

| Parameter | $\delta$ (2-body) | $\delta$ (SOAP) |
|-----------|-------------------|-----------------|
| AgCu      | 0.43              | 0.0126          |
| AlFe      | 0.84              | 0.044           |
| AlMg      | 0.43              | 0.0126          |
| AlNi      | 0.396             | 0.0193          |
| AlTi      | 0.78              | 0.03            |
| CoNi      | 0.27              | 0.0337          |
| CuFe      | 0.84              | 0.0446          |
| CuNi      | 0.315             | 0.0207          |
| FeV       | 0.184             | 0.0407          |
| NbNi      | 0.9               | 0.05            |

For all alloy fits, the error hyperparameter  $\sigma$  was set to 1 meV for energies. Force and virial were not used in the fits. Because pure energies have a large effect on the predicted formation enthalpies, we increased the error hyperparameter to  $10^{-4}$  for those training configurations that represented pure elements. This ensured accurate reproduction of the pure energies so that enthalpy errors closely match errors in total energy for configurations.

The parameter  $\epsilon_0$  was calculated for each isolated atom by including a padding of 10 Å around a single atom and using the same pseudopotential as the bulk calculations discussed above. These energies were converged with respect to basis set size and used only the  $\gamma$  k-point.

#### 5. MTP

MTP was introduced in Ref. 11 for single-component system and in Refs. 19,20 was extended to multicomponent systems. MTP partitions the predicted energy into contributions of environments of each atom. Around the central atom of an environment, the neighboring atoms form shells. In these shells, atoms are assigned fictitious weights depending on the distance to the central atom, their types, and the type of the central atom. These weights are free parameters fitted from data. An environment is described by moments of inertia of these shells. All possible contractions of one or more moment tensor to a scalar comprise an infinite sequence of basis functions. This sequence is truncated to yield a finite set of basis functions used in a particular MTP model. The contribution of an environment to the energy is, thus, a linear combination of basis function with coefficients which are also found from data. Refer to Ref. 20 for more details.

In this work for binary systems, we used an MTP with about 300 basis functions. The cutoff for atomic environments was 7 Å. The environments were described by five shells, and the dependence of the weight of a neighbor on the distance to the central atom of the environment was described by eight basis functions. Thus, the total number of parameters in a binary MTP is  $5 \times 8 \times 2^2 + 300 \approx 450$  (the factor  $2^2$  follows from the fact that there can be two types of the central atom and two types of each neighboring atoms). For the 10-component MTP, we

used six shells and 850 basis functions, totaling about  $6 \times 8 \times 30 + 864 \approx 2300$  parameters, where the factor 30 is the number of interacting pairs of atoms.

### B. *k*-point density

Table IV. Shown are the minimum and maximum values of *k*-point density across all structures for each of the alloys for computing the DFT total energies.

| System | Number of <i>k</i> -points /Å <sup>3</sup> |         |
|--------|--|---------|
|        | Maximum                                    | Minimum |
| AgCu   | 550  | 516     |
| AlFe   | 596  | 468     |
| AlMg   | 635  | 478     |
| AlNi   | 589  | 464     |
| AlTi   | 535  | 399     |
| CoNi   | 554  | 433     |
| CuFe   | 568  | 444     |
| CuNi   | 561  | 440     |
| FeV    | 480  | 401     |
| NbNi   | 516  | 472     |

### C. Analysis of dataset and models

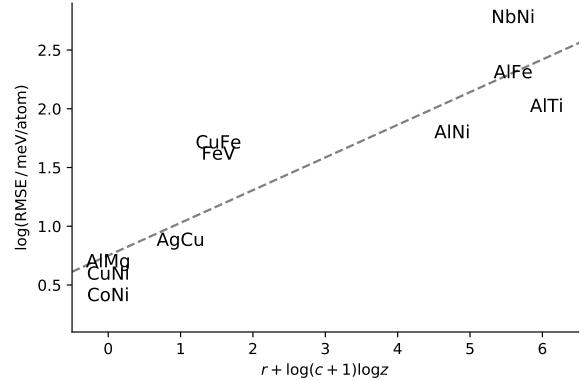


Figure 5. *Alchemical similarity explains prediction errors.* Shown are the logarithmized root mean squared error (RMSE; compare Figure 2) as a function of an analytic expression in the difference in row  $r$  and column  $c$  of the periodic table as well as atomic number  $z$  of the two chemical element species of a binary alloy.  $R^2 = 0.81$ .

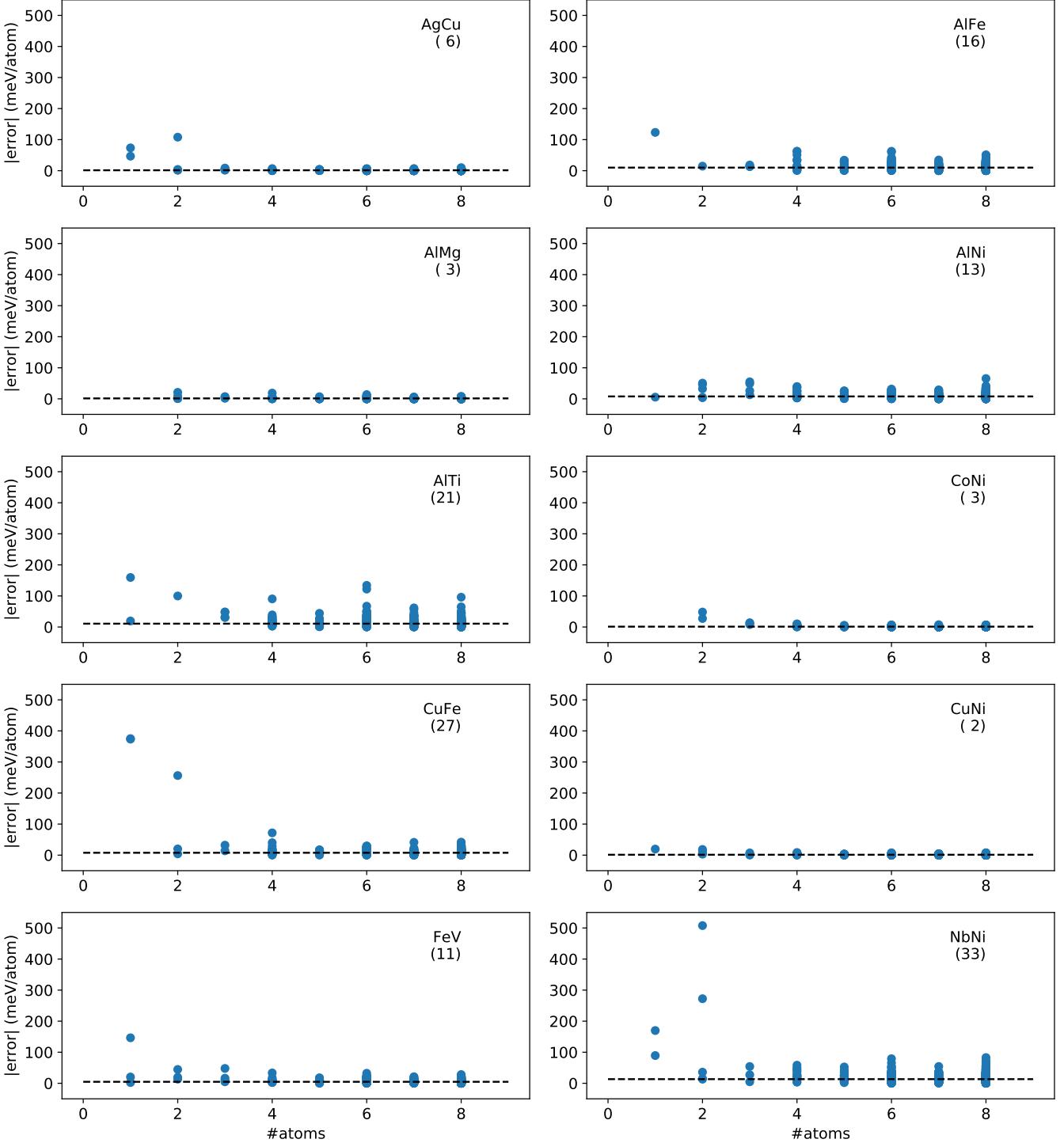


Figure 6. *Influence of unit cell size on errors.* Shown are the absolute errors (meV/atom) as a function of the number of atoms in the unit cell for a validation set of 595 randomly chosen structures. The number in brackets and the dashed line indicate the root mean squared error (RMSE, meV/atom) and the median absolute error (meV/atom) on the same set. If small structures (one or two atoms in the unit cell) are not contained in the training data (that is, are shown in the plot) they tend to have larger errors, increasing overall RMSE as well. If all small structures are contained in the training data, the overall RMSE is low (AlMg, CoNi). Retraining models with small structures included in the training set improved RMSE in all cases, by an amount depending on how many structures were added.

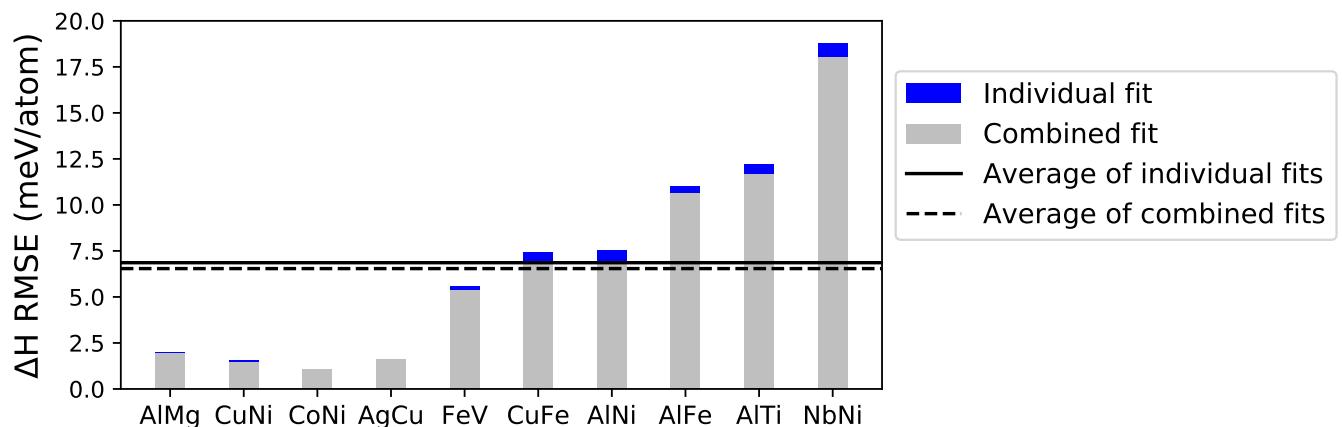


Figure 7. *Improvement of MBTR+DNN model on all alloys.* Shown are the root mean squared error (RMSE) when trained on each alloy separately (blue bars) and on all alloys simultaneously (grey bars).