Bandgap engineering in the configurational space of solid solutions via machine learning: (Mg,Zn)O case study

THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCL

Letter

Bandgap Engineering in the Configurational Space of Solid Solutions via Machine Learning: (Mg,Zn)O Case Study

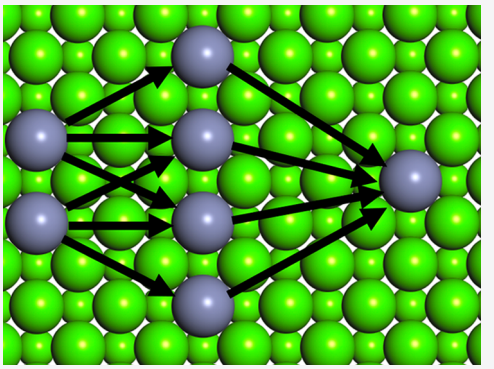
Scott D. Midgley, Said Hamad, Keith T. Butler,* and Ricardo Grau-Crespo*

Cite This: *J. Phys. Chem. Lett.* 2021, 12, 5163−5168

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Computer simulations of alloys' properties often require calculations in a large space of configurations in a supercell of the crystal structure. A common approach is to map density functional theory results into a simplified interaction model using so-called cluster expansions, which are linear on the cluster correlation functions. Alternative descriptors have not been sufficiently explored so far. We show here that a simple descriptor based on the Coulomb matrix eigenspectrum clearly outperforms the cluster expansion for both total energy and bandgap energy predictions in the configurational space of a MgO−ZnO solid solution, a prototypical oxide alloy for bandgap engineering. Bandgap predictions can be further improved by introducing non-linearity via gradient-boosted decision trees or neural networks based on the Coulomb matrix descriptor.



Density functional theory (DFT) is the most widely used electronic structure simulation technique in modern materials theory research. Despite its widespread use, DFT can incur a very high computational cost, making access to a high-performance computer a requisite for many applications, and prompting research into cheaper and more efficient ways to compute electronic properties of materials.

In recent years, machine learning (ML) has seen growing research interest in theoretical materials science because of its potential to reduce computational cost by several orders of magnitude compared with traditional DFT-only approaches.[1−4] The development of atomic-level descriptors such as the Coulomb matrix has led to great progress in the accelerated prediction of molecular and material properties.[5,6] The Coulomb matrix, defined as

$$M_{ij} = \begin{cases} 0.5Z_i^{2.4}, & i = j \\ \dfrac{Z_iZ_j}{|\boldsymbol{R}_i - \boldsymbol{R}_j|}, & i \neq j \end{cases}$$

where Z_i and $\boldsymbol{R}_i$ are the atomic numbers and positions of the atoms in the structure, was first used by Rupp and co-workers to show that a Gaussian regression method was able to accurately predict atomization energies in gas-phase molecules, significantly reducing the computational cost of a standard *ab initio* approach.[7] Typically the matrix is flattened to vector form by using the sorted spectrum of eigenvalues, leading to a convenient vector shape for the descriptor (the Coulomb matrix eigenspectrum or CME), which is invariant to translation, rotations, and permutations of atom indices. The CME descriptor has been generalized to periodic systems and employed for the description of formation energies in solids.[8]

The investigation of the vast configurational space of solid solutions is another area where ML can accelerate predictions. The most established approach to calculate the energies (and sometimes other properties) of solid solution configurations is to create a so-called cluster expansion, where the energy is represented as a linear expansion of cluster correlation functions (CCFs) of increasing order, i.e., points, pairs, trios, quartets, etc.[9] Cluster expansions have been hugely successful in the theoretical understanding of alloys, but they also have limitations, for example, related to relaxation effects and numerical errors.[10,11] Rosenbrock et al. have recently proposed ML potentials as an alternative to cluster expansions for the investigation of alloy phase diagrams.[12] Natarajan and van der Ven employed ML tools including neural networks to generalize the cluster expansion approach by relaxing the condition of linearity on the CCFs.[13] An alternative approach, which we follow in this work, is to use a different descriptor altogether, one that is not constrained by the locality of the CCFs, like the CME mentioned above. This is especially worth exploring for the prediction of non-additive properties, such as bandgaps, where the cluster expansion might not perform as well as for energies.

Received: March 31, 2021
Accepted: May 17, 2021

https://doi.org/10.1021/acs.jpclett.1c01031
J. Phys. Chem. Lett. 2021, 12, 5163−5168

Solid solutions offer the possibility of band structure engineering for many applications. $Mg_{1-x}Zn_xO$ solid solutions, chosen here as a case study, constitute an important family of wide-gap semiconductors with tunable bandgaps from 3.3 to 7.8 eV.[14,15] Thin films made of these solid solutions are of interest in the field of ultraviolet optoelectronic devices.[16−18] Precise bandgap engineering is therefore needed, which can be achieved to a great extent via compositional optimization. We are interested here in the possibility of optimizing the bandgap in the configurational space (at fixed composition) rather than in the compositional space, since it is known that modern crystal-growth techniques, like molecular beam epitaxy, can produce targeted crystal structures, often in defiance of equilibrium thermodynamics. Previous DFT calculations performed in alloy models in a small 16-atom cell have already suggested the existence of large bandgap fluctuations due to differences in the local arrangement of Mg and Zn atoms.[19] However, expanding these DFT-based studies to larger supercells to properly explore the configuration space would have a prohibitively large computational cost.

We present here an investigation of different computational approaches to map the bandgaps of alloy configurations into a simple model that allows fast prediction and screening across a large configurational space. We use the 3:1 MgO−ZnO rocksalt solid solution as a case study, both because it is a well-known system with important applications and because it does not pose extra challenges to DFT like partially filled *d* orbitals or spin polarization. We will compare the performance of CCF vs CME descriptors, as well as linear vs non-linear regression models, in the hope of discovering new routes for more accurate bandgap engineering in solid solutions.

The MgO and ZnO end members have cubic and hexagonal crystal structures, respectively.[20,21] A 64-atom cubic supercell with composition $Zn_8Mg_{24}O_{32}$ was chosen as a case study for the assessment of ML methods for the prediction of mixing energy (E_{mix}) and band gap energy (E_{gap}) in the solid solution. This composition and cell size give 8043 symmetrically inequivalent cation configurations, with configurations considered equivalent if they are related by a symmetry operator of the parent structure.[22] We used the Supercell code to generate the inequivalent configurations.[23] This number of configurations is both large enough to yield statistically meaningful data-driven conclusions and small enough to permit a full DFT treatment for training and validation of the ML models.

Symmetrically inequivalent configurations were subject to full geometry and cell vector optimization using DFT simulations with periodic boundary conditions, as implemented in the VASP code.[24] The generalized gradient approximation (GGA) was used for the exchange-correlation term, with the functional by Perdew, Burke, and Ernzerhof (PBE).[25] The projector augmented wave (PAW) method was used to describe the interactions between atomic cores and valence electrons.[26,27] A plane wave kinetic energy cutoff of 520 eV was used, which is 30% above the recommended value for the set of PAW potentials used, to minimize Pulay stress errors. The end members are modeled with high accuracy with this type of calculations, as we can see by the good agreement between DFT-optimized cell parameters and experimental values in Table 1.

It is well known that GGA-PBE gives a poor description of bandgaps, generally underestimating the experimental values. In order to find out how to correct the PBE values, a small subset of 20 configurations across the full range of bandgaps

Table 1. Relaxed Cell Parameters and Bandgaps of the Solid Solution End Members (MgO and ZnO) from DFT Calculations, in Comparison with Experimental Values

	MgO		ZnO	
	cubic ($Fm\bar{3}m$)		hexagonal ($P6_3mc$)	
crystal system (space group)	calc	exp	calc	exp
a/Å	PBE: 4.24	4.22	PBE: 3.24	3.25
c/Å	–	–	PBE: 5.18	5.21
E_{gap}/eV	PBE: 4.5		PBE: 1.4	
	HSE: 6.2[a]	7.8[b]	HSE: 2.6[a]	3.3[c]

[a]Calculated using the HSE functional at PBE geometry. [b]Ref 14. [c]Ref 15.

was chosen for more accurate calculations using the screened hybrid functional by Heyd, Scuseria, and Ernzerhof (HSE), which incorporates 25% Hartree−Fock exchange energy and is much better than GGA at predicting bandgaps.[28] We demonstrate that for the ZnO/MgO alloy studied here, the PBE bandgaps may be easily corrected via a simple linear transformation to reproduce the HSE bandgaps. The linear relation between the bandgap values calculated with PBE and with HSE can be seen in Figure S1a in the Supporting Information (SI). This strong linear correlation between PBE and HSE bandgaps is not general, and in systems including transition-metal or rare-earth elements, for example, we would expect much weaker correlations. For such systems, the non-linear relationship between PBE and HSE bandgaps can be established using a machine-learned transformation.[29] However, in our case the simple linear relationship will allow us to use PBE band gaps for training the bandgap predicting models, instead of the more expensive but more accurate HSE values. It can also be seen from Table 1 that, while giving better predictions than PBE, HSE still underestimates the experimental bandgaps for pure MgO and ZnO, in both cases by ∼20%. So it is reasonable to expect a similar underestimation by HSE of the solid solution bandgaps.

We used ML methods to learn from DFT-derived E_{mix} and E_{gap} values for a subset of configurations and to predict the values for the rest of the configurations. This procedure permits a significant reduction of the computational cost, brought about by a reduction in the number of DFT calculations required to obtain accurate E_{mix} and E_{gap} values for the entire configurational space. As descriptors of the alloy configurations, we used either the full vector of cluster correlation functions (CCFs) or the Coulomb matrix eigenspectrum (CME). The CCF vectors have 90 components, corresponding to all the symmetrically distinct clusters up to four-body terms, as calculated using the CELL code.[30] More information about the CCF descriptor employed in this work is given in the SI. The 64-component CME vectors were generated using the Python 3 packages Matminer and Pymatgen.[31,32]

Linear regression (LR) and gradient-boosted decision tree (GBDT) methods were performed using Python 3 Scikit-Learn packages.[33] For LR models, we added weak *LASSO* regularization to obtain physically meaningful parameters.[34] Deep-learning neural networks of the feedforward multilayer perceptron (MLP) architecture were written using the Keras[35] package, which is built on the TensorFlow[30] platform. MLP models were subject to extensive architecture testing, though only two architectures, which we will refer to as *shallow* and *deep*, are discussed forthwith. The *shallow* architecture is a

https://doi.org/10.1021/acs.jpclett.1c01031
J. Phys. Chem. Lett. 2021, 12, 5163−5168

three-layer feedforward perceptron with 64-32-1 nodes per layer, whereas the *deep* architecture is a five-layer feedforward perceptron with 256-128-64-32-1 nodes per layer. Data was split into sets based on a percentage of the 8043 datapoints available: training (fractions between 10% and 80% were tried), validation (10%), and testing (10%). This ensured that ML vs DFT energy plots involved data that had not been used for either training or validation, and that the testing dataset size stayed constant when varying the training dataset size. More details about the ML algorithms can be found in the SI.

We briefly discuss the DFT results first, before moving into the regression models. Figure 1 reports the mixing energies

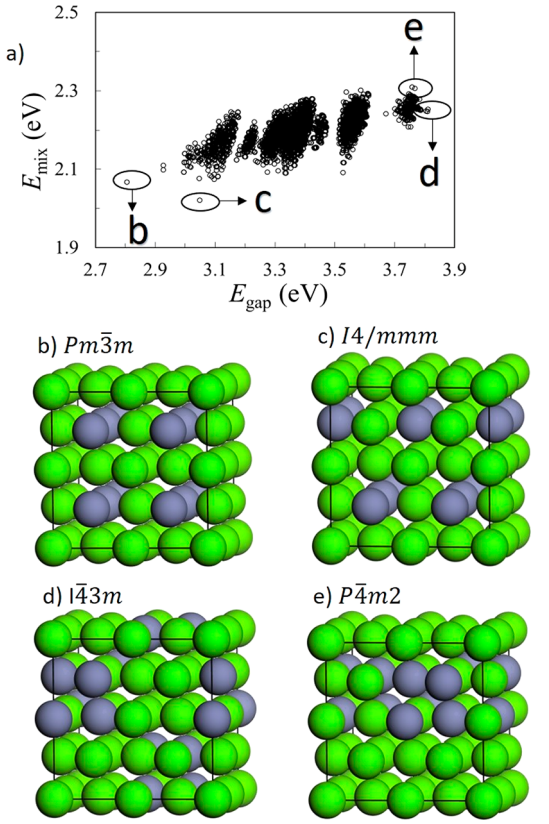


Figure 1. (a) DFT data (mixing energy vs band gap energy) for all 8043 symmetrically different $Zn_8Mg_{24}O_{32}$ configurations. Structures of the configurations with (b) minimum E_{gap}, (c) minimum E_{mix}, (d) maximum E_{gap}, and (e) maximum E_{mix}. Green and gray balls represent Mg and Zn atoms, respectively (O atoms are omitted for clarity).

plotted against the bandgaps as obtained by DFT calculations for the whole dataset of 8043 configurations. The wide range of bandgaps (~1 eV difference between the minimum and maximum PBE values, which can be estimated to correspond to a range width of ~1.5 eV in the experimental scale), together with the small stability difference between configurations (less than 0.3 eV per supercell, which is less than 0.01 eV per formula unit), confirms that this would be a suitable system for *configurational* bandgap optimization, at fixed composition. There is some weak but clear correlation between E_{mix} and E_{gap}, suggesting that thermodynamics might oppose the arrangement of cation distributions in the ways that lead to maximum bandgaps. However, given the small energy differences, we would not expect thermodynamics to prevent the experimental realization of these wide-gap configurations.

The geometries of the configurations with minimum and maximum values of E_{mix} and E_{gap} are also shown in Figure 1. The configuration with the lowest bandgap (Figure 1b) has the same distribution of ions as the ordered fcc alloy Cu_3Au, i.e., has the structure with Strukturbericht designation $L1_2$ and space group $Pm\bar{3}m$. This configuration is characterized by −Zn−O−Zn−O− one-dimensional chains along the three equivalent [100], [010], and [001] directions of the crystal structure. Since ZnO has a much lower bandgap than MgO, it is not surprising that the presence of periodic ZnO-only chains tends to lower the bandgaps. The configuration with the lowest mixing energy, i.e., the configurational ground state for the composition $Mg_{3/4}Zn_{1/4}O$, is the one with Strukturbericht designation $D0_{22}$ and space group $I4/mmm$, as in the ordered alloy Al_3Ti, which agrees with the conclusion from the previous theoretical study by Sanati et al.[36] This configuration also has −Zn−O−Zn−O− one-dimensional chains along two of the crystal axes, but the cations alternate in the third direction, forming −Mg−O−Zn−O− chains (Figure 1c). The configurations with the maximum values of E_{gap} (Figure 1d) and E_{mix} (Figure 1e) both have all Zn dopants forming alternating −Mg−O−Zn−O− chains, with no pure −Zn−O−Zn−O− chains along the crystal axes. However, in the most unstable configuration (maximum E_{mix}), with space group $P\bar{4}m2$, these chains aggregate within two neighboring layers (the cation size disparity between Zn and Mg is likely to cause high crystal strain when concentrated at one side of the cell, which explains the high mixing energy), whereas in the former the distribution of the chains forms a more homogeneous, checkered-like pattern with space group $I\bar{4}3m$.

The main purpose of this work is not, however, to identify configurations with extremal properties, but to devise fast and accurate methods to calculate the properties of any alloy configuration. Figure 2a shows the plot of predicted vs true data for the test set, using models based on the CCF (i.e., the

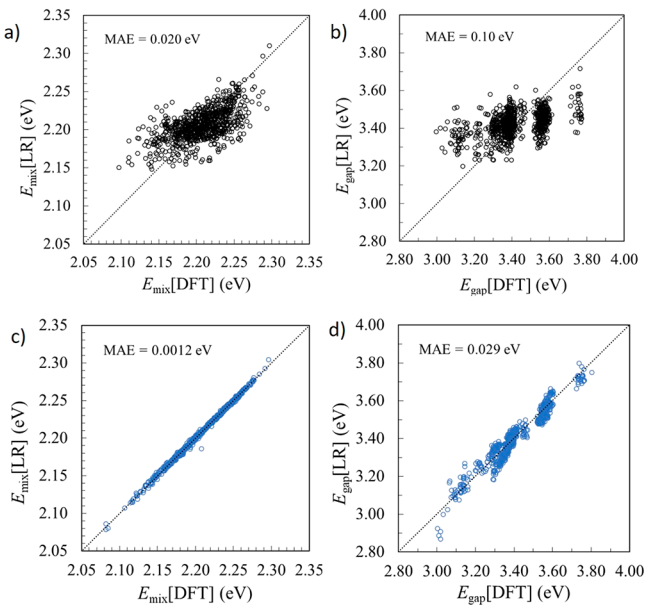


Figure 2. Performance of linear regression models, trained on 80% of the data, when used for the test set using (a) cluster correlation function (CCF) descriptor for E_{mix}, (b) CCF descriptor for E_{gap}, (c) Coulomb matrix eigenspectrum (CME) descriptor for E_{mix}, and (d) CME descriptor for E_{gap}.

https://doi.org/10.1021/acs.jpclett.1c01031
J. Phys. Chem. Lett. 2021, 12, 5163−5168

cluster expansions), when 80% of the data was used for training. The correlation between the cluster expansions and the mixing energies obtained directly from DFT is rather poor. This is somewhat surprising, since cluster expansions generally perform well at describing energy differences between alloy configurations. For MgO−ZnO solid solutions, Yin et al. have previously presented a cluster expansion for the formation energies, which fitted well their DFT energies, with one-point and two-point clusters found to be dominant in the expansion.[37] However, in that case the authors were examining configurations across a range of compositions, and therefore the one-point correlation functions were the dominant term, thus improving the correlation between predicted and target energies. In our case, we are working at a fixed composition (so we leave the one-point cluster correlation functions out of the regression) and the range of energies is very narrow, which is more challenging for the cluster expansion. So even when the mean absolute error for our cluster expansion is small (0.02 eV per supercell, which is less than 1 meV per formula unit), the correlation between the predicted and target data is still weak (R^2 = 0.39).

The plot of predicted vs true data for the cluster expansion of the bandgaps, also based on training with 80% of the data, is shown in Figure 2b. In this case the correlation is even poorer (R^2 = 0.22), which is not surprising, given that bandgaps are not additive and depend on the long-range pattern in the distribution of ions in the solid, which is not necessarily well captured by the local cluster functions. But even when cluster expansions of bandgaps are not as well established or justified as the cluster expansions of energies,[38] the method has been widely used for bandgaps,[39,40] and no reliable alternatives have been developed. Relaxing the linearity condition on the CCFs (as done recently in a different context in ref 13) did not significantly improve the performance of the descriptor: a MLP model trained using the CCF descriptor yielded equally poor correlations (see SI).

In contrast, using the CME as a descriptor leads to excellent correlation between predicted and target data, as shown in Figure 2c,d. The prediction for E_{mix} is particularly outstanding, with a mean absolution error of ∼1 meV per supercell on the test set. Even the bandgap prediction is quite good, although with some more dispersion. The observation that the CME descriptor performs better than the CCFs, which are traditionally used for cluster expansions, is very interesting, since cluster expansions have been the preferred theoretical tool for the investigation of the configurational space of alloys for several decades. Using widely available tools, generating the CME is just as easy and computationally cheap as generating the CCFs, and we demonstrate here that it can lead to more accurate predictions. Of course, the advantage of a model based on the CCFs is that, once the cluster expansion is generated, it can be used to explore the configuration energies in supercells larger than those used in the fitting. This is useful to compute thermodynamic properties with converged cell sizes. However, this advantage does not translate trivially to the prediction of bandgaps or other non-additive quantities. In these cases, as we are constrained to make predictions within the same supercell size from where configurations are sampled for training, the CME descriptor might be a more accurate and equally cheap choice.

Finally, we consider whether non-linear regression models can further improve the CME-based description of the bandgaps, based on the CME descriptor. Figure 3a,b shows

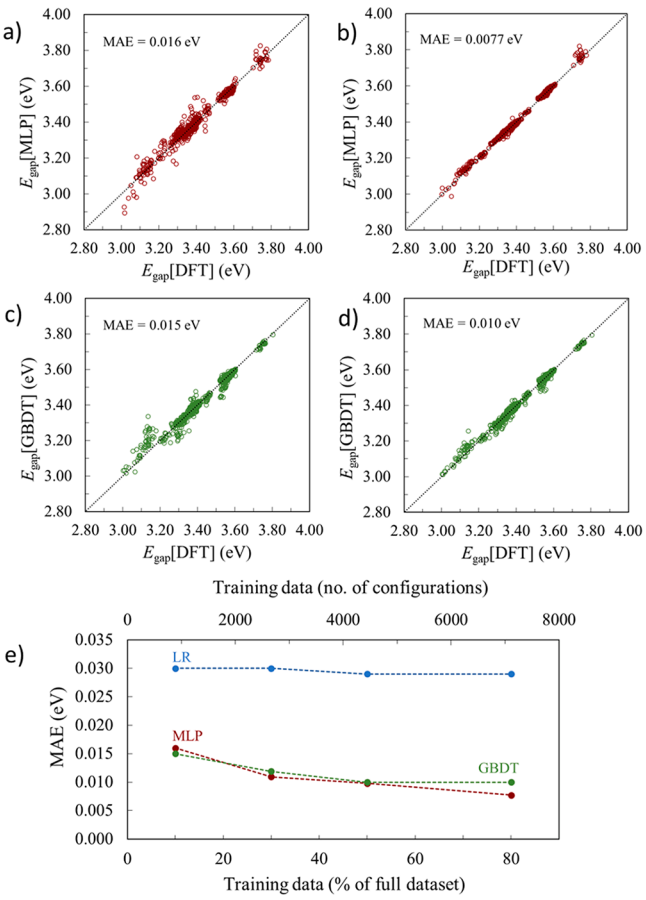


Figure 3. CME machine learning models for E_{gap}: (a) *deep* MLP-10%, (b) *deep* MLP-80%, (c) GBDT-10%, (d) GBDT-80%, and (e) MAE vs % training data (of 8043 total configurations).

the bandgap prediction made by the *deep* MLP model (the *shallow* MLP results are reported in Figure S3 of the SI). Clearly, the MLP improves the prediction with respect to the linear regression model. Even when only 10% of the data is used for training, the predicted bandgaps are much better (with roughly half of the MAE) than those predicted with the linear regression model using 80% of the data for training. Furthermore, MLP models show significant improvement when increasing the dataset size, whereas the linear regression model does not seem to benefit from the use of additional training data. A comparison between the MLP methods in terms of performance for the *shallow* and *deep* architectures is reported in Table S1 of the SI. The *deep* MLP is deeper and wider than the *shallow* MLP and provides slightly improved performance because the increased complexity of this MLP may capture more non-linearities in the CME−E_{gap} relationship. There is a slightly increased risk of overfitting when using a more complex MLP, though we found no evidence of this during training.

GBDT models (Figure 3c,d), trained using optimized hyperparameters reported in the SI, also proved to be very effective in predicting bandgaps, especially for the small- to medium-sized training sets. The performance of the GBDT model saturates after a certain size of training set between 50% and 80% of the data used here, meaning that it is unlikely to benefit as much as MLP from increasing the training dataset size. However, given that the associated mean absolute errors are similar to those of the MLP models, GBDT models

https://doi.org/10.1021/acs.jpclett.1c01031
J. Phys. Chem. Lett. 2021, 12, 5163−5168

constitute an attractive alternative, since the computational cost of training these models is smaller than for the neural networks. A full performance comparison for the three ML methods is given in the SI, Table S2.

In conclusion, we have shown that Coulomb matrix eigenspectrum descriptors outperform the cluster correlation functions typically used for cluster expansions in the prediction of both properties for a MgO−ZnO solid solution. Cluster expansions are more justified for configurational thermodynamics, because energy expansions are trivially extrapolated to the very large supercells required for accurate statistical mechanics. However, for the screening of bandgaps in the configurational space, cluster expansions are not ideal, not only because of the non-additive character of bandgaps which limits the extrapolation to larger supercells but also because the cluster expansions might not capture well the bandgap variations in the first place, as we have shown in this study. We suggest that, for this problem, a better approach is to sample the configurational space in an affordable supercell, perform DFT calculations, and then use modern machine learning tools, based on Coulomb matrix eigenspectrum descriptors and linear or non-linear regression models (depending on the size of the available datasets). Given the wide availability and low computational cost of these machine learning tools, we believe that this approach will become the new standard for the prediction of electronic properties in the configurational space of semiconducting alloys.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.1c01031.

Bandgap correction using the screened hybrid functional HSE; calculation of the cluster correlation functions; performance of MLP neural network using CCF descriptor; comparison of shallow vs deep MLP neural networks using CME descriptor; further details about the machine learning algorithms; metrics summary; and information about codes and data (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Keith T. Butler − *SciML, Scientific Computing Department, Rutherford Appleton Laboratory, Harwell OX11 0QX, United Kingdom;* orcid.org/0000-0001-5432-5597; Email: keith.butler@stfc.ac.uk

Ricardo Grau-Crespo − *Department of Chemistry, University of Reading, Reading RG6 6DX, United Kingdom;* orcid.org/0000-0001-8845-1719; Email: r.grau-crespo@reading.ac.uk

Authors

Scott D. Midgley − *Department of Chemistry, University of Reading, Reading RG6 6DX, United Kingdom*

Said Hamad − *Department of Physical, Chemical and Natural Systems, Universidad Pablo de Olavide, 41013 Seville, Spain;* orcid.org/0000-0003-4148-2344

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.1c01031

Notes

The authors declare no competing financial interest.

Associated codes and data are available online: https://doi.org/10.5281/zenodo.4736810 (data) and https://github.com/scott-midgley/Machine-Learning-for-Solid-Solutions (codes).

■ ACKNOWLEDGMENTS

We thank Dr. Gonzalo Nápoles (Tilburg University) for useful comments. This work made use of ARCHER, the UK's national high-performance computing service, via the UK's HPC Materials Chemistry Consortium, which is funded by EPSRC (EP/R029431), and of the Young supercomputer, via UK Materials and Molecular Modelling Hub, which is partially funded by EPSRC (EP/T022213/1). S.H. acknowledges funding from the Agencia Estatal de Investigación and the Ministerio de Ciencia, Innovación y Universidades, of Spain (PID2019-110430G B-C22), and from the EU FEDER Framework 2014-2020 and Consejería de Conocimiento, Investigación y Universidad of the Andalusian Government (FEDER-UPO-1265695).

■ ABBREVIATIONS

DFT, density functional theory; ML, machine learning; CME, Coulomb matrix eigenspectrum; CCF, cluster correlation function; GGA, generalized gradient approximation; PBE, Perdew, Burke, and Ernzerhof; PAW, projector augmented wave; HSE, Heyd, Scuseria, and Ernzerhof; LR, linear regression; GBDT, gradient-boosted decision tree; MLP, multilayer perceptron; MAE, mean absolute error

■ REFERENCES

(1) Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95* (14), 144110.

(2) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301.

(3) Hong, Y.; Hou, B.; Jiang, H.; Zhang, J. Machine learning and artificial neural network accelerated computational discoveries in materials science. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10* (3), No. e1450.

(4) Agrawal, A.; Choudhary, A. Deep materials informatics: Applications of deep learning in materials science. *MRS Commun.* **2019**, *9* (3), 779−792.

(5) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.

(6) Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* **2010**, *9* (4), 273−276.

(7) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108* (5), 058301.

(8) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **2015**, *115* (16), 1094−1101.

(9) Sanchez, J. M.; Ducastelle, F.; Gratias, D. Generalized cluster description of multicomponent systems. *Phys. A* **1984**, *128* (1), 334−350.

(10) Nguyen, A. H.; Rosenbrock, C. W.; Reese, C. S.; Hart, G. L. W. Robustness of the cluster expansion: Assessing the roles of relaxation and numerical error. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *96* (1), 014107.

(11) Grau-Crespo, R.; Waghmare, U. V. Simulation of Crystals with Chemical Disorder at Lattice Sites. In *Molecular Modeling for the*

Design of Novel Performance Chemicals and Materials; Rai, B., Ed.; CRC Press, 2012; p 303.
(12) Rosenbrock, C. W.; Gubaev, K.; Shapeev, A. V.; Pártay, L. B.; Bernstein, N.; Csányi, G.; Hart, G. L. W. Machine-learned interatomic potentials for alloys and alloy phase diagrams. *npj Computational Materials* **2021**, *7* (1), 24.
(13) Natarajan, A. R.; Van der Ven, A. Machine-learning the configurational energy of multicomponent crystalline solids. *npj Computational Materials* **2018**, *4* (1), 56.
(14) Roessler, D. M.; Walker, W. C. Electronic Spectrum and Ultraviolet Optical Properties of Crystalline MgO. *Phys. Rev.* **1967**, *159* (3), 733−738.
(15) Srikant, V.; Clarke, D. R. On the optical band gap of zinc oxide. *J. Appl. Phys.* **1998**, *83* (10), 5447−5451.
(16) Sharma, A.; Narayan, J.; Muth, J.; Teng, C.; Jin, C.; Kvit, A.; Kolbas, R. M.; Holland, O. Optical and structural properties of epitaxial $Mg_xZn_{1-x}O$ alloys. *Appl. Phys. Lett.* **1999**, *75* (21), 3327−3329.
(17) Choopun, S.; Vispute, R.; Yang, W.; Sharma, R.; Venkatesan, T.; Shen, v. Realization of band gap above 5.0 eV in metastable cubic-phase $Mg_xZn_{1-x}O$ alloy films. *Appl. Phys. Lett.* **2002**, *80* (9), 1529−1531.
(18) Han, S.; Zhang, J.; Zhang, Z.; Zhao, Y.; Wang, L.; Zheng, J.; Yao, B.; Zhao, D.; Shen, D. Mg0.58Zn0.42O Thin Films on MgO Substrates with MgO Buffer Layer. *ACS Appl. Mater. Interfaces* **2010**, *2* (7), 1918−1921.
(19) Onuma, T.; Ono, M.; Ishii, K.; Kaneko, K.; Yamaguchi, T.; Fujita, S.; Honda, T. Impact of local arrangement of Mg and Zn atoms in rocksalt-structured $Mg_xZn_{1-x}O$ alloys on bandgap and deep UV cathodoluminescence peak energies. *Appl. Phys. Lett.* **2018**, *113* (6), 061903.
(20) Sasaki, S.; Fujino, K.; Takeuchi, Y. X-Ray Determination of Electron-Density Distributions in Oxides, MgO, MnO, CoO, and NiO, and Atomic Scattering Factors of their Constituent Atoms. *Proc. Jpn. Acad., Ser. B* **1979**, *55* (2), 43−48.
(21) Albertsson, J.; Abrahams, S. C.; Kvick, Å. Atomic displacement, anharmonic thermal vibration, expansivity and pyroelectric coefficient thermal dependences in ZnO. *Acta Crystallogr., Sect. B: Struct. Sci.* **1989**, *45* (1), 34−40.
(22) Grau-Crespo, R.; Hamad, S.; Catlow, C. R. A.; Leeuw, N. H. d. Symmetry-adapted configurational modelling of fractional site occupancy in solids. *J. Phys.: Condens. Matter* **2007**, *19* (25), 256201.
(23) Okhotnikov, K.; Charpentier, T.; Cadars, S. Supercell program: a combinatorial structure-generation approach for the local-level modeling of atomic substitutions and partial occupancies in crystals. *J. Cheminf.* **2016**, *8* (1), 17.
(24) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54* (16), 11169−11186.
(25) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865−3868.
(26) Blochl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50* (24), 17953−17979.
(27) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59* (3), 1758−1775.
(28) Heyd, J.; Scuseria, G. E. Efficient hybrid density functional calculations in solids: Assessment of the Heyd−Scuseria−Ernzerhof screened Coulomb hybrid functional. *J. Chem. Phys.* **2004**, *121* (3), 1187−1192.
(29) Lentz, L. C.; Kolpak, A. M. Predicting HSE band gaps from PBE charge densities via neural network functionals. *J. Phys.: Condens. Matter* **2020**, *32* (15), 155901.
(30) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. TensorFlow: A System for Large-Scale Machine Learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, Savannah, GA, Nov 2−4, 2016; pp 265−283.
(31) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.
(32) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152* (C), 60.
(33) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine Learning in Python. *J. Machine Learning Res.* **2011**, *12*, 2825−2830.
(34) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **1996**, *58* (1), 267−288.
(35) Chollet, F. *Keras: The python deep learning library*, https://github.com/fchollet/keras.
(36) Sanati, M.; Hart, G. L.; Zunger, A. Ordering tendencies in octahedral MgO-ZnO alloys. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2003**, *68* (15), 155210.
(37) Yin, W.-J.; Dai, L.; Zhang, L.; Yang, R.; Li, L.; Guo, T.; Yan, Y. Stability, transparency, and conductivity of $Mg_xZn_{1-x}O$ and $Cd_xZn_{1-x}O$: Designing optimum transparency conductive oxides. *J. Appl. Phys.* **2014**, *115* (2), 023707.
(38) Xu, X.; Jiang, H. Cluster expansion based configurational averaging approach to bandgaps of semiconductor alloys. *J. Chem. Phys.* **2019**, *150* (3), 034102.
(39) Burton, B.; Demers, S.; Van de Walle, A. First principles phase diagram calculations for the wurtzite-structure quasibinary systems SiC-AlN, SiC-GaN and SiC-InN. *J. Appl. Phys.* **2011**, *110* (2), 023507.
(40) Magri, R.; Froyen, S.; Zunger, A. Electronic structure and density of states of the random $Al_{0.5}Ga_{0.5}As$, $GaAs_{0.5}P_{0.5}$, and $Ga_{0.5}In_{0.5}As$ semiconductor alloys. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1991**, *44* (15), 7947−7964.

https://doi.org/10.1021/acs.jpclett.1c01031
J. Phys. Chem. Lett. 2021, 12, 5163−5168