Machine-learning correction to density-functional crystal structure optimization

Robert Hussein,* Jonathan Schmidt, Tomás Barros, Miguel A.L. Marques, and Silvana Botti

Density functional theory is routinely applied to predict crystal structures. The most common exchange-correlation functionals used to this end are the Perdew–Burke–Ernzerhof (PBE) approximation and its variant PBEsol. We investigate the performance of these functionals for the prediction of lattice parameters and show how to enhance their accuracy using machine learning. Our data set is constituted by experimental crystal structures of the Inorganic Crystal Structure Database matched with PBE-optimized structures stored in the materials project database. We complement these data with PBEsol calculations. We demonstrate that the accuracy and precision of PBE/PBEsol volume predictions can be noticeably improved *a posteriori* by employing simple, explainable machine learning models. These models can improve PBE unit cell volumes to match the accuracy of PBEsol calculations, and reduce the error of the latter with respect to experiment by 35 percent. Further, the error of PBE lattice constants is reduced by a factor of 3–5. A further benefit of our approach is the implicit correction of finite temperature effects without performing phonon calculations.

Impact statement

Knowledge about the crystal structure of solids is essential for describing their elastic and electronic properties. In particular, their accurate prediction is essential to predict the electronic properties of not-yet-synthesized materials. Lattice parameters are most commonly calculated by density functional theory using the Perdew–Burke–Ernzerhof (PBE) approximation and its variant PBEsol as exchange-correlation functional. They are successful in describing materials properties but do, however, not always achieve the desired accuracy in comparison with experiments. We propose a computationally efficient scheme based on interpretable machine learning to optimize crystal structures. We demonstrate that the accuracy of PBE- and PBEsol-structures can be, therewith, enhanced noticeably. In particular, the PBE unit cells, available in materials databases, can be improved to the level of the more accurate PBEsol calculations and the error of the latter with respect to the experiment can be reduced by 35 percent. An additional advantage of our scheme is the implicit inclusion of finite temperature corrections, which makes expensive phonon calculations unnecessary.

Introduction

Computational high-throughput studies form the basis for the discovery of new materials in modern materials science. In solid-state physics, these studies are mostly performed within Kohn–Sham density functional theory (DFT).[1–3] Although DFT formally provides an exact description of the many-body Schrödinger equation, it relies in practice on approximations for the exchange-correlation energy. In solid-state physics, one commonly utilizes the Perdew–Burke-Ernzerhof's (PBE) functional.[4] Although PBE and its variants are successful in predicting structural and electronic properties of solids, they may yield, nevertheless, non-negligible deviations from experiments. Specifically, PBE underestimates atomic bond lengths, thus, overestimating lattice constants[5–7] and volumes. Variants of PBE such as the PBE for solids (PBEsol)[8] were designed to improve upon this problem.[9] However, they still do not achieve the desired accuracy in comparison with experiments.

The correctness of the lattice constants and the corresponding unit cell volumes

Robert Hussein, Institut für Festkörpertheorie und -optik, Friedrich-Schiller-Universität Jena, Jena, Germany; European Theoretical Spectroscopy Facility, Jena, Germany; robert.hussein@uni-jena.de
Jonathan Schmidt, European Theoretical Spectroscopy Facility, Jena, Germany; Institut für Physik, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany
Tomás Barros, European Theoretical Spectroscopy Facility, Jena, Germany; Institut für Physik, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany
Miguel A.L. Marques, European Theoretical Spectroscopy Facility, Jena, Germany; Institut für Physik, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany
Silvana Botti, Institut für Festkörpertheorie und -optik, Friedrich-Schiller-Universität Jena, Jena, Germany; European Theoretical Spectroscopy Facility, Jena, Germany
*Corresponding author
doi:10.1557/s43577-022-00310-9

is indispensable for a reliable prediction of bulk electronic properties[10–12] and when considering experimental realizations of composite materials. For instance, the lattice mismatch between growth substrates and films can be a source of major problems in experiments. Another reason to focus on lattice parameters is the fact that this is the material property for which it is possible to find the largest amount of experimental data, collected in the Inorganic Crystal Structure Database (ICSD).[13]

In this article, we demonstrate that machine learning methods can improve the lattice volume predictions based on PBE/PBEsol without increasing the computational effort. Machine learning enjoyed over the past few years great success in a wide variety of applications[14] ranging from property predictions of bandgaps[15–17] and elastic moduli[18] to the stability analysis of crystals[19,20] and molecular force-field estimations.[21] Recently, the prediction of lattice constants and volumes generated much interest.[22–30] The majority of these studies are, however, limited to a particular crystal structure. In contrast to previous studies that are mainly based on direct calculations, we are building our approach on first-principles calculations, aiming at improving their accuracy in comparison with corresponding experiments. Our approach is not limited to a specific crystal structure or a subset of chemical elements. We will focus here on applying explainable machine learning methods[31,32] to correct errors of PBE/PBEsol calculations of crystal structures of newly predicted materials. Specifically, we will employ model agnostic supervised local explanations (MAPLEs)[33] in combination with a random forest model[34] to combine the high accuracy of tree models for small data sets and the interpretability of MAPLE models.

Before diving into detail, we can observe in **Figure 1** the primitive unit cell volumes V_{pred}, predicted from DFT calculations using PBE and PBEsol functionals, plotted against the experimental unit cell volumes V_{exp}. We remark that the primitive cell volume is the simplest quantity that can be directly compared, independently of the specific details of the crystal structure and chemical composition. Their correlation gives a first impression about the accuracy (systematic error) and precision (variability) of the theoretically estimated unit cell volumes. Calculations with the PBE functional (magenta circles) significantly overestimate the measured volumes by roughly 11%, whereas PBEsol (green squares) provide a much better approximation of them.

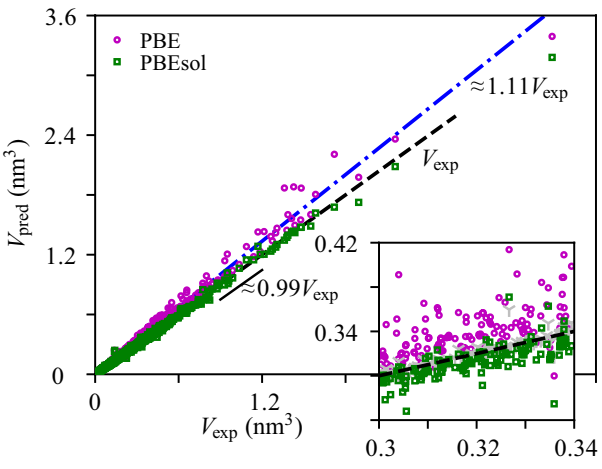

Figure 1. Correlation between measured primitive unit cell volumes V_{exp} and predicted ones V_{pred}. Ideal predictions match the black dashed line. The blue dash-dotted line depicts PBE's linear regression. The inset additionally considers the machine learning prediction pred = PBE+MAPLE marked by gray tripods.

On a closer look, one sees that V_{exp} constitutes a soft lower bound on the PBE volumes in the sense that about 90% of the predicted volumes lay above it. This is a consequence of the tendency of PBE to underbind. This soft bound entails a skewness on the predicted PBE volume distribution, which we revisit later. The inset of Figure 1 shows a close-up view of primitive unit cell volumes of about 0.32 nm^3 to better distinguish the individual data points. We additionally include in the inset the volumes obtained by correcting PBE calculations with machine learning (gray tripods) to anticipate visually the strong error reduction. We will discuss thoroughly the machine learning corrections in the next sections.

The remaining article is organized as follows. In "Predictive models," we present the employed machine learning models. Details on the experimental and theoretical data sets and their matching are discussed in "Data set." We analyze in "Volume predictions" our predictive models and compare their performance with the one of underlying DFT calculations. In "Lattice constants," we discuss the correction of the lattice constants. The last section is devoted to our conclusions.

Predictive models

Tree-ensemble-based models such as random forests[34] and gradient boosting[35,36] are known to be suitable to the description of materials properties for relatively small data sets,[37,38] but they are not restricted to them.[39] A drawback of employing multiple-decision trees is however their general lack of interpretability.[32] Appropriate combination with local linear models,[40–42] as in model agnostic supervised local explanations (MAPLEs),[33] overcomes this deficiency by providing local and example-based explanations. The former addresses causal relations between specific *input features* of an individual prediction (such as lattice constants) and its outcome by identifying their importance.[33,43,44] The latter asks instead for the contribution of specific *training points*.[45–47] Note that local and example-based explanations are helpful to understand the specific predictions of the model, but do not explain its global behavior. For this, we have to resort to simpler models, such as the analytic formulas used in symbolic regression methods.[48,49]

In this work, we employ the MAPLE implementation of Plumb et al.[33] as well as tree models and utility functions from Reference 50. We evaluate the machine learning models by

tenfold Monte Carlo cross-validation. We choose this approach instead of using an independent test set, because our full data set exhibits a large variance in structures/elements while being relatively small in size. In each independent run of the cross-validation scheme, the full data set is randomly split at a ratio 1:9 without replacement into a test and a training set. The hyper parameter optimization has been performed on a separate random splitting. Here, the number of trees forming the tree ensemble turns out to be the most important hyper parameter. The minimal number of samples controlling the splitting is of minor importance. The theoretical crystal structures calculated using DFT at the PBE and PBEsol level serve as input parameters for the training. Specifically, we consider seven structural input parameters, which are namely the primitive unit cell volume, the lattice constants, and the angles between the lattice vectors. We complement them with 44 composition-specific features provided by Matminer.[51]

By training the MAPLE models with the data sets discussed in "Data set," we find that the primitive cell volume prediction is, indeed, to a large extent based on structural quantities (see **Table** SI in the Online Resource). Here, we use the average of the root-node impurity over the decision tree ensemble as an estimator to quantify the relevance of the features. The binary splittings of an individual decision tree are such constructed that they minimize its impurity. In this sense, the first splitting and, therewith, the respective root-node feature has a major impact on the decision tree structure and the model prediction. In particular, the splitting of the training set with respect to the root feature is in more than 50% of the cases directly related to the crystal structure of the compounds, through (e.g., the volume V, the lattice constants a, b, c, and the corresponding angles α, β, γ). Concerning the compositional features, the periodic-table-based features and averaged thermal properties of the elements play by far the largest role exceeding also two of the lattice angles in importance. In order to better assess the contribution of structure- and composition-specific features to the volume optimization, we study in Section I of the Online Resource how the MAPLE models perform when only trained on one of these subsets. Our models are available at Reference 52.

In Section II of the Online Resource, we address different machine learning methods. In particular, we consider the symbolic regression method Operon[49] as well as the Crystal Graph Convolutional Neural Network[53] discussed among others in the benchmark by Dunn et al.[54]

Data set

For our analysis and model training, we consider roughly 2000 PBE-structures from the Materials Project (MP).[55] The corresponding PBEsol calculations are available from Reference 56. The experimental crystal structures are extracted from the ICSD.[13] A mapping of ICSD- and MP-identifiers is provided in **Table** SII of the Online Resource.

We remark that experiments are conducted at finite temperature (2–373 K) and pressure (≤ 1 bar) whereas DFT calculations describe equilibrium structures at zero temperature and pressure. In order to obtain PBE(sol) crystal structures in the same thermodynamic conditions than experimental samples, they should be corrected by expensive phonon calculations for thermal expansions and zero-point effects arising from finite lattice fluctuations.[57–60] For small molecules, the ambient pressure has additionally to be taken into account[61] but may be neglected for solids.[57] By training the predictive models on finite temperature volumes V_{exp} as target variables one has the advantage to implicitly include finite temperature corrections. In principle, the ambient temperature of the measurements could be included as an input parameter for machine learning. However, it turns out that the resulting models are mostly independent of temperature, because the large majority of the experiments are performed at roughly the same temperature (about 293 K, median and mode). The mean value of the temperature distribution is indeed 271 K.

Volume predictions

In this section, we compare volume predictions obtained with various machine learning models. First, we give an overview of these predictions by discussing the central characteristics of their volume residuals. Then, we study their cross-validation error and finally, address the convergence of the model training.

We show in **Figure 2**a violin plots[62] of the volume residuals $V_{exp} - V_{pred}$. One sees clearly that the median of the DFT-PBE calculations (red line in magenta violin) of about $Q_2 \approx -7.4$ Å^3 is, indeed, corrected by simple linear regression (blue violin). Also its interquartile range IQR $\equiv Q_3 - Q_1$ of about 14 Å^3 is roughly reduced by 2 Å^3 with the drawback that already well predicted volumes worsen. The skewness remains. Employing MAPLE cures the skewness and reduces the spreading further up to a third of the initial value (see gray violin). Intriguingly, its volume forecast is comparable

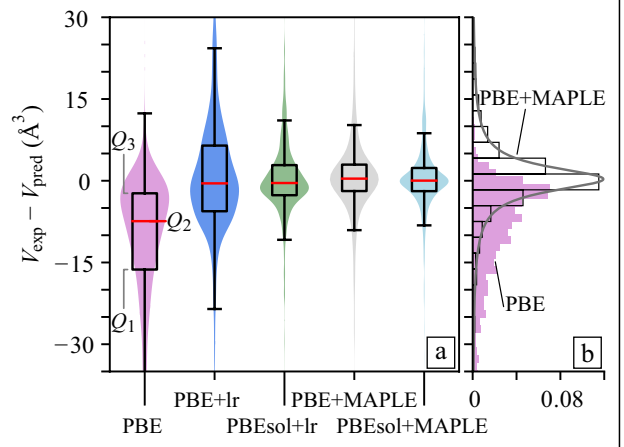

Figure 2. Violin plot (a) and probability densities (b) of the volume residuals $V_{exp} - V_{pred}$ for the indicated test sets. The suffix "+lr" indicates linear regression and Q_k is the kth quartile.

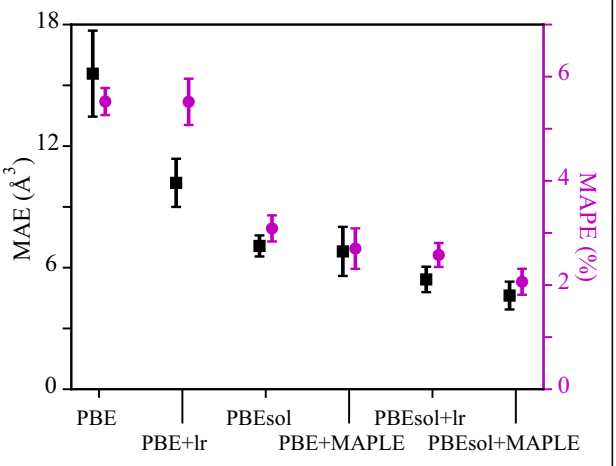

Figure 3. Cross-validation errors for different models with standard deviations as error bars.

to the simple linear regression prediction starting from PBEsol volumes. Beyond the median and the interquartile range, the violin shapes in **Figure** 2a estimate the entire probability densities of the volume residuals. For the purpose of assessing their estimation quality, we compare the DFT-PBE estimate with the corresponding normalized histogram depicted in **Figure** 2b. The estimate captures well the curve progression but is less pronounced around its mode located at −1.3 Å^3. Additionally, we show in panel (b) the normalized histogram of the MAPLE prediction that corrects PBE volumes. As prefigured, it is considerably narrower and can be well approximated by a slightly biased Lorentzian (gray solid line) with a linewidth of roughly its interquartile range.

For a more quantitative comparison of the different models, we focus in the following on the cross-validation error. The cross-validation error of a specific model is obtained by evaluating for each test set the error of its prediction with respect to the measured value, and taking the arithmetic mean of these errors. Additionally, we determine the standard deviation of the individual errors. As error metrics we chose the mean absolute error $\mathrm{MAE} = \sum_{k=1}^{n} | y_{k,exp} - y_k | /n$ and the mean absolute percentage error $\mathrm{MAPE} = 100 \sum_{k=1}^{n} | y_{k,exp} - y_k | / | n y_{k,exp} |$, where $y_{k,exp}$ (y_k) indicates the measured (predicted) property of the k-th sample. Because all experimental volumes are finite, MAPE is well defined. In **Figure 3**, we show the cross-validation errors of the predicted primitive unit-cell volumes for different models. Their numerical mean values and standard deviations are listed in **Table I**. As expected, DFT-PBE itself leads overall to the worst cross-validation errors whereas PBE corrected with linear regression (+lr) improves the MAE leaving the MAPE unchanged. The MAPLE model based on PBE volumes reduces the PBE-MAE by about 50% and is slightly better than DFT-PBEsol volumes. However, PBEsol corrected with linear regression is once again better. Most importantly, the MAPLE model based on PBEsol shows the best MAE improving by roughly 35% upon DFT calculations alone with this functional. The IQRs in **Table I** show the same tendency as the MAE and MAPE, supporting the conclusion regarding the possible improvements achievable with the MAPLE models. Additionally, we report therein the MAPLE models using the measured temperature as an input feature. They perform, however, very similarly to the models that do not include such feature, as expected in view of the fact that most experiments were conducted at about the same temperature.

To assess the learning progress of the MAPLE models, we study in **Figure 4** the dependence of their MAE on the training-set size n_{train}. The MAE's are again obtained by ten-fold cross-validation. As expected, they decrease polynomially with the training-set size n_{train}.[63,64] In particular, we find with $\mathrm{MAE} \propto n_{train}^{-0.28}$ for PBE+MAPLE and $\mathrm{MAE} \propto n_{train}^{-0.24}$ for PBEsol+MAPLE a similar learning behavior for both predictive models. Including the total data available in the ICSD, the cross-validation error could be potentially reduced by 50 percent. The relatively fast decay of the cross-validation errors with respect to the training-set size makes these correction procedures already applicable for small training sets.

Table I. Cross-validation errors of volume predictions and standard deviations for different DFT functionals and correction models.

Model	MAE (Å^3)		MAPE (%)		IQR (Å^3)	
	Mean	Std	Mean	Std	Mean	Std
PBE	15.6	2.1	5.5	0.3	14.0	0.7
PBE+lr	10.2	1.2	5.5	0.4	11.9	1.1
PBEsol	7.1	0.5	3.1	0.3	6.9	0.7
PBE+MAPLE	6.8	1.2	2.7	0.4	5.1	0.8
PBE+MAPLE[a]	6.6	0.9	2.6	0.3	4.9	0.4
PBEsol+lr	5.4	0.6	2.6	0.2	5.5	0.5
PBEsol+MAPLE[a]	4.8	0.7	2.1	0.3	4.2	0.6
PBEsol+MAPLE	4.6	0.7	2.1	0.2	4.2	0.5

[a] Indicates the corresponding models including the temperature of the measurement as an additional feature.

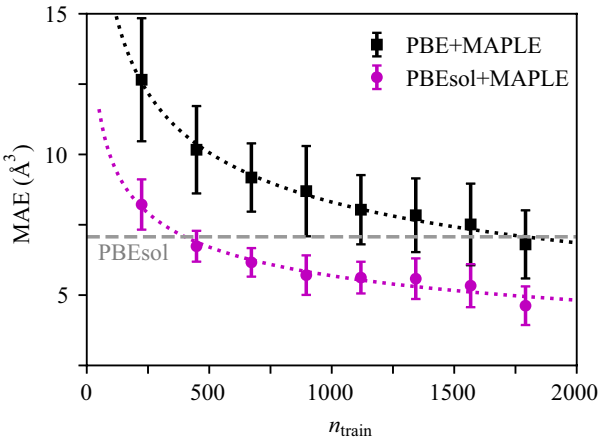

Figure 4. Cross-validation errors for different models as a function of the training-set size n_{train}. The error bars indicate their standard deviation and the dotted lines correspond to their polynomial fit discussed in the main text. As reference, we indicate PBEsol's MAE for the test set with the horizontal dashed line.

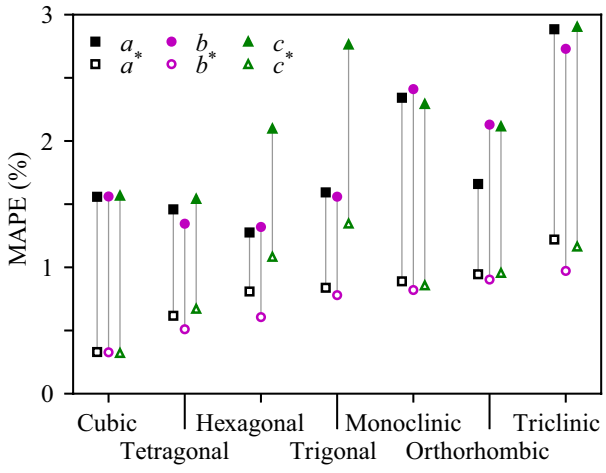

Figure 5. Errors of the DFT-PBE lattice constants a, b, c and the rescaled ones a^*, b^*, c^* according to Equation 1 with respect to experimental values from the ICSD for different crystal structures. The gray vertical lines serve as a guidance for the eye for tracking the individual improvements.

We would expect it to generalize to other materials properties such as bulk moduli or formation energies for which very few experimental data are available and DFT results are a worse starting point.

Lattice constants

Thus far, we have discussed volume corrections. To a certain extent, we can, therewith, also improve the lattice constants as we show in this section. To this end, we recall how the volume is calculated. The unit cell volume $V(\mathbf{a},\mathbf{b},\mathbf{c})$ is obtained from the triple product of the three lattice vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and can be written as product $V = abc \mid \text{polsin}(\alpha,\beta,\gamma) \mid$ of a factor abc only depending on their lengths and a dimensionless factor $\mid \text{polsin}(\alpha,\beta,\gamma) \mid$ only depending on their interior angles.* The latter is for cubic crystal systems well predicted by PBE and PBEsol with a MAPE on the order of 0.02% whereas lower symmetric systems do not exceed a MAPE of 0.6 percent. Exploiting the simplification that all lattice constants coincide for cubic crystals, we can extract the lattice constant correction by the prescription

$$a \to a^* \equiv a\sqrt[3]{\frac{V_{\text{PBE(sol)+MAPLE}}}{V_{\text{PBE(sol)}}}} \tag{1}$$

for DFT calculations using PBE(sol). Therewith, the MAPE of the lattice vectors is roughly reduced by a factor of 5 (see **Figure 5**). If we use the same prescription as an approximate way to correct the lattice constants of non-cubic systems, we observe a consistent reduction of the MAPE of a factor of 3–5.

In particular, we observe that when the three lattice parameters display different errors, the largest errors are those that get reduced more effectively, suppressing overall the MAPE of PBE lattice constants to less than 1 percent.

Conclusions

We have investigated machine-learning-based unit cell volume corrections for DFT calculations. Model agnostic supervised local explanations improve both PBE's and PBEsol's volume prediction of the primitive unit cell. By applying MAPLE on PBE, one achieves overall improvements on the level of PBEsol calculations, hence, trivially reducing PBE's volume deviations from experiments by about 50 percent. This is of great convenience because all large solid-state databases rely on PBE calculations. We provide our implementation at Reference 52. Furthermore, PBEsol+MAPLE outperforms PBEsol with a roughly 1.5 times smaller mean absolute error. The most relevant features contributing to the MAPLE models are, indeed, given by the lattice parameters calculated with DFT, whereas composition-specific features are significantly less important. A further benefit of our approach is the implicit correction of finite temperature effects rendering time-consuming phonon calculations unnecessary. Because the considered experiments are mostly performed at the same (room) temperature, our trained MAPLE models are, however, not expected to generalize well to other temperatures. We plan to address this point in future by training on data sets with a larger temperature variation.

* The polar sine satisfies $\text{polsin}(\alpha,\beta,\gamma) \equiv \det([\mathbf{a},\mathbf{b},\mathbf{c}])/abc = \left(1 + 2\cos\alpha\cos\beta\cos\gamma - \cos^2\alpha - \cos^2\beta - \cos^2\gamma\right)^{1/2}$.[65]

Acknowledgments

This work was supported by the Volkswagen Stiftung (Momentum) through the project dandelion.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Data availability

Available in the supporting information.

Conflict of interest

The authors declare that there is no conflict of interest.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1557/s43577-022-00310-9.

References

1. P. Hohenberg, W. Kohn, *Phys. Rev.* **136**, B864 (1964). https://doi.org/10.1103/PhysRev.136.B864
2. W. Kohn, L.J. Sham, *Phys. Rev.* **140**, A1133 (1965). https://doi.org/10.1103/PhysRev.140.A1133
3. R.O. Jones, O. Gunnarsson, *Rev. Mod. Phys.* **61**, 689 (1989). https://doi.org/10.1103/RevModPhys.61.689
4. J.P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996). https://doi.org/10.1103/PhysRevLett.77.3865
5. F. Tran, R. Laskowski, P. Blaha, K. Schwarz, *Phys. Rev. B* **75**, 115131 (2007). https://doi.org/10.1103/PhysRevB.75.115131
6. G.X. Zhang, A.M. Reilly, A. Tkatchenko, M. Scheffler, *New J. Phys.* **20**, 063020 (2018). https://doi.org/10.1088/1367-2630/aac7f0
7. P. Kovács, F. Tran, P. Blaha, G.K.H. Madsen, *J. Chem. Phys.* **150**, 164119 (2019). https://doi.org/10.1063/1.5092748
8. J.P. Perdew, A. Ruzsinszky, G.I. Csonka, O.A. Vydrov, G.E. Scuseria, L.A. Constantin, X. Zhou, K. Burke, *Phys. Rev. Lett.* **100**, 136406 (2008). https://doi.org/10.1103/PhysRevLett.100.136406
9. P. Haas, F. Tran, P. Blaha, *Phys. Rev. B* **79**, 085104 (2009). https://doi.org/10.1103/PhysRevB.79.085104
10. F.D. Murnaghan, *Proc. Natl. Acad. Sci. U.S.A.* **30**, 244 (1944). https://doi.org/10.1073/pnas.30.9.244
11. F. Birch, *Phys. Rev.* **71**, 809 (1947). https://doi.org/10.1103/PhysRev.71.809
12. E. Ziambaras, E. Schröder, *Phys. Rev. B* **68**, 064112 (2003). https://doi.org/10.1103/PhysRevB.68.064112
13. G. Bergerhoff, I.D. Brown, *Crystallographic Databases* (International Union of Crystallography, Chester, 1987)
14. J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, *NPJ Comput. Mater.* **5**, 83 (2019). https://doi.org/10.1038/s41524-019-0221-0
15. T. Gu, W. Lu, X. Bao, N. Chen, *Solid State Sci.* **8**, 129 (2006). https://doi.org/10.1016/j.solidstatesciences.2005.10.011
16. G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis, T. Lookman, *Sci. Rep.* **6**, 19375 (2016). https://doi.org/10.1038/srep19375
17. Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, *J. Phys. Chem. Lett.* **9**, 1668 (2018). https://doi.org/10.1021/acs.jpclett.8b00124
18. M. de Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta, A. Gamst, *Sci. Rep.* **6**, 34256 (2016). https://doi.org/10.1038/srep34256
19. W. Ye, C. Chen, Z. Wang, I.H. Chu, S.P. Ong, *Nat. Commun.* **9**, 3800 (2018). https://doi.org/10.1038/s41467-018-06322-x
20. J. Schmidt, L. Chen, S. Botti, M.A.L. Marques, *J. Chem. Phys.* **148**, 241728 (2018). https://doi.org/10.1063/1.5020223
21. A. Glielmo, C. Zeni, A. De Vita, *Phys. Rev. B* **97**, 184307 (2018). https://doi.org/10.1103/PhysRevB.97.184307
22. S.G. Javed, A. Khan, A. Majid, A.M. Mirza, J. Bashir, *Comput. Mater. Sci.* **39**, 627 (2007). https://doi.org/10.1016/j.commatsci.2006.08.015
23. K. Takahashi, L. Takahashi, J.D. Baran, Y. Tanaka, *J. Chem. Phys.* **146**, 204104 (2017). https://doi.org/10.1063/1.4984047
24. A. Majid, A. Khan, G. Javed, A.M. Mirza, *Comput. Mater. Sci.* **50**, 363 (2010). https://doi.org/10.1016/j.commatsci.2010.08.028
25. Y. Zhang, X. Xu, Chem. *Phys. Lett.* **760**, 137993 (2020). https://doi.org/10.1016/j.cplett.2020.137993
26. M. Nait Amar, M.A. Ghriga, M.E.A. Ben Seghier, H. Ouaer, *J. Phys. Chem. B* **124**, 6037 (2020). https://doi.org/10.1021/acs.jpcb.0c04259
27. I.O. Alade, I.A. Olumegbon, A. Bagudu, *J. Appl. Phys.* **127**, 015303 (2020). https://doi.org/10.1063/1.5130664
28. Y. Li, W. Yang, R. Dong, J. Hu, MLatticeABC: Generic lattice constant prediction of crystal materials using machine learning. arXiv:2010.16099 (2020)
29. Y. Zhang, X. Xu, *ChemistrySelect* **5**, 9999 (2020). https://doi.org/10.1002/slct.202002532
30. Y. Zhang, X. Xu, *Int. J. Appl. Ceram. Technol.* **18**, 661 (2021). https://doi.org/10.1111/ijac.13709
31. Z.C. Lipton, *Queue* **16**, 31 (2018). https://doi.org/10.1145/3236386.3241340
32. M. Du, N. Liu, X. Hu, *Commun. ACM* **63**, 68 (2019). https://doi.org/10.1145/3359786
33. G. Plumb, D. Molitor, A.S. Talwalkar, *Proc. Adv. Neural. Inform. Process. Syst.* **31** (Curran Associates, Red Hook, 2018), p. 2515
34. L. Breiman, *Mach. Learn.* **45**, 5 (2001). https://doi.org/10.1023/A:1010933404324
35. J.H. Friedman, J.J. Meulman, *Stat. Med.* **22**, 1365 (2003). https://doi.org/10.1002/sim.1501
36. S. Theodoridis, *Machine Learning* (Academic Press, London, 2015). https://doi.org/10.1016/C2013-0-19102-7
37. O. Isayev, D. Fourches, E.N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, *Chem. Mater.* **27**, 735 (2015). https://doi.org/10.1021/cm503507h
38. A. Furmanchuk, A. Agrawal, A. Choudhary, *RSC Adv.* **6**, 95246 (2016). https://doi.org/10.1039/C6RA19284J
39. V. Stanev, C. Oses, A.G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, *NPJ Comput. Mater.* **4**, 29 (2018). https://doi.org/10.1038/s41524-018-0085-8
40. W.S. Cleveland, S.J. Devlin, *J. Am. Stat. Assoc.* **83**, 596 (1988). https://doi.org/10.1080/01621459.1988.10478639
41. D. Ruppert, M.P. Wand, *Ann. Stat.* **22**, 1346 (1994). https://doi.org/10.1214/aos/1176325632
42. J. Barrientos-Marin, F. Ferraty, P. Vieu, *J. Nonparametr. Stat.* **22**, 617 (2010). https://doi.org/10.1080/10485250903089930
43. D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, K.R. Müller, *J. Mach. Learn. Res.* **11**, 1803 (2010)
44. S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.I. Lee, *Nat. Mach. Intell.* **2**, 56 (2020). https://doi.org/10.1038/s42256-019-0138-9
45. R.D. Cook, S. Weisberg, *Technometrics* **22**, 495 (1980). https://doi.org/10.1080/00401706.1980.10486199
46. B. Kim, R. Khanna, O.O. Koyejo, *Proc. Adv. Neural. Inform. Process. Syst.* **29** (Curran Associates, Red Hook, 2016), p. 2280
47. J. Bien, R. Tibshirani, *Ann. Appl. Stat.* **5**, 2403 (2011). https://doi.org/10.1214/11-AOAS495
48. R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L.M. Ghiringhelli, *Phys. Rev. Mater.* **2**, 08382 (2018). https://doi.org/10.1103/PhysRevMaterials.2.083802
49. B. Burlacu, G. Kronberger, M. Kommenda, in *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion* (Association for Computing Machinery, New York, 2020), p. 1562. https://doi.org/10.1145/3377929.3398099
50. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **12**, 85 (2011)
51. L. Ward, A. Dunn, A. Faghaninia, N.E.R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K.A. Persson, G.J. Snyder, I. Foster, A. Jain, *Comput. Mater. Sci.* **152**, 60 (2018). https://doi.org/10.1016/j.commatsci.2018.05.018

52. The implementation of the predictive models is provided at https://github.com/hyllios/utils/tree/main/models/unit_cell_volume
53. T. Xie, J.C. Grossman, *Phys. Rev. Lett.* **120**, 145301 (2018). https://doi.org/10.1103/PhysRevLett.120.145301
54. A. Dunn, Q. Wang, A. Ganose, D. Dopp, A. Jain, *NPJ Comput. Mater.* **6**, 138 (2020). https://doi.org/10.1038/s41524-020-00406-3
55. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, *APL Mater.* **1**, 011002 (2013). https://doi.org/10.1063/1.4812323
56. J. Schmidt, H.C. Wang, T.F.T. Cerqueira, S. Botti, M.A.L. Marques, *Mater. Cloud Arch.* **4**, 5 (2021). https://doi.org/10.24435/materialscloud:r5-gx
57. B. Grabowski, T. Hickel, J. Neugebauer, *Phys. Rev. B* **76**, 024309 (2007). https://doi.org/10.1103/PhysRevB.76.024309
58. K. Lejaeghere, V. Van Speybroeck, G. Van Oost, S. Cottenier, *Crit. Rev. Solid State Mater. Sci.* **39**, 1 (2014). https://doi.org/10.1080/10408436.2013.772503
59. P.B. Allen, *Phys. Rev. B* **92**, 064106 (2015). https://doi.org/10.1103/PhysRevB.92.064106
60. E.T. Ritz, S.J. Li, N.A. Benedek, *J. Appl. Phys.* **126**, 171106 (2019). https://doi.org/10.1063/1.5125779
61. J. Hoja, A.M. Reilly, A. Tkatchenko, *WIREs Comput. Mol. Sci.* **7**, e1294 (2017). https://doi.org/10.1002/wcms.1294
62. J.L. Hintze, R.D. Nelson, *Am. Stat.* **52**, 181 (1998). https://doi.org/10.1080/00031305.1998.10480559
63. J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M.A.L. Marques, *Chem. Mater.* **29**, 5090 (2017). https://doi.org/10.1021/acs.chemmater.7b00156
64. Y. Zhang, C. Ling, *NPJ Comput. Mater.* **4**, 25 (2018). https://doi.org/10.1038/s41524-018-0081-z
65. F. Eriksson, *Geom. Dedic.* **7**, 71 (1978). https://doi.org/10.1007/BF00181352