How to represent crystal structures for machine learning: Towards fast prediction of electronic properties

K. T. Schütt,[1,*] H. Glawe,[2,*] F. Brockherde,[1,2] A. Sanna,[2] K. R. Müller,[1,3,†] and E. K. U. Gross[2,†]

[1]*Machine Learning Group, Technische Universität Berlin, Marchstrasse 23, 10587 Berlin, Germany*

[2]*Max-Planck-Institut für Mikrostrukturphysik, Weinberg 2, 06120 Halle, Germany*

[3]*Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Republic of Korea*

(Received 3 July 2013; revised manuscript received 10 February 2014; published 21 May 2014)

High-throughput density functional calculations of solids are highly time-consuming. As an alternative, we propose a machine learning approach for the fast prediction of solid-state properties. To achieve this, local spin-density approximation calculations are used as a training set. We focus on predicting the value of the density of electronic states at the Fermi energy. We find that conventional representations of the input data, such as the Coulomb matrix, are not suitable for the training of learning machines in the case of periodic solids. We propose a novel crystal structure representation for which learning and competitive prediction accuracies become possible within an unrestricted class of *spd* systems of arbitrary unit-cell size.

DOI: 10.1103/PhysRevB.89.205118 PACS number(s): 71.15.−m, 71.20.−b, 81.30.−t, 89.20.Ff

I. INTRODUCTION

In recent years *ab initio* high-throughput computational methods (HTM) have proven to be a powerful and successful tool to predict new materials and to optimize desired materials properties. Phase diagrams of multicomponent crystals [1–3] and alloys [4] have been successfully predicted. High-impact technological applications have been achieved by improving the performance of lithium-based batteries [5–7], by tailoring the nonlinear optical response in organic molecules [8] for optical signal processing, by designing desired current-voltage characteristics [9] for photovoltaic materials, by optimizing the electrode transparency and conductivity [10] for solar cell technology, and by screening metals for the highest amalgamation enthalpy [11] to efficiently remove Hg pollutants in coal gasification.

However, the computational cost of electronic structure calculations poses a serious bottleneck for HTM. Thinking of quaternary, quinternary, etc., compounds, the space of possible materials becomes so large and the complexity of the unit cells so high that, even within efficient Kohn-Sham density functional theory (KS-DFT), a systematic high-throughput exploration grows beyond reach for present-day computing facilities. As a way out, one would like to have a more direct way to access the physical property of interest without actually solving the KS-DFT equations. Machine learning (ML) techniques offer an attractive possibility of this type. ML-based calculations are very fast, typically requiring only fractions of a second to predict a specific property of a given material, after the ML model has been trained on a representative training set of materials.

ML methods rely on two main ingredients, the learning algorithm itself and the representation of the input data. There are many different ways of representing a given material or compound. While, from the physicist's point of view, the information is simply given by the charges and the positions of the nuclei, for ML algorithms the specific mathematical form in which this information is given to the machine is crucial. Roughly speaking, ML algorithms assume a nonlinear map between input data (representing the materials or compounds in our case) and the material-specific property to be predicted. Whether or not a machine can approximate the unknown nonlinear map between input and property well and efficiently mainly depends on a good representation [12–14]. Recently, ML has contributed accurate models for predicting molecular properties [15,16], transition states [17], potentials [18], and self-consistent solutions for DFT [19]. All these applications deal with finite systems (atoms, molecules, clusters). For this type of system, one particular way of representing the material, namely, the so-called Coulomb matrix, has been very successful.

In electronic-structure problems, the single most important property is the value of the density of states (DOS) at the Fermi energy. Susceptibilities, transport coefficients, the Seebeck coefficient, and the critical temperature of superconductors are all closely related to the DOS at the Fermi energy. Therefore, we have chosen this quantity to be predicted by ML.

In this work, we shall report a fundamental step forward in the application of machine learning to predict the DOS at the Fermi energy. The two main questions this work aims to address are as follows: (a) How can we describe an infinite periodic system in a way that supports the learning process well? (b) How large should the data basis for ML training be, i.e., the *training set* of calculations? Answering these questions will provide us with exactly the sought-after method of direct and fast prediction and with the knowledge of whether such prediction is indeed possible given the finite amount of training data compatible with present-day computing power.

II. LEARNING AND REPRESENTATION

We employ so-called kernel-based learning methods [20,21] that are based on a mapping to a high-dimensional *feature space* such that an accurate prediction can be achieved

*K. T. Schütt and H. Glawe contributed equally to this work.

†Corresponding authors: These authors jointly directed the project. hardy@mpi-halle.mpg.de, klaus-robert.mueller@tu-berlin.de

with a linear model in this space. The so-called *kernel trick* allows us to perform this mapping implicitly using a kernel function, e.g., the Gaussian kernel $k(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2)$ or the Laplacian kernel $k(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|/\sigma)$. Kernels can be viewed as a similarity measure between data; in our case they should measure the proximity between materials for a certain property. The property to be predicted is computed as a linear combination of kernel functions of the material of interest and the training materials. Therefore, constructing a structure representation in which crystals have a small distance when their properties are similar is beneficial for the learning process (see below for details).

In order to predict the DOS, we employ *kernel ridge regression* (KRR), which is a kernelized variant of least-squares regression with ℓ_2 regularization. Additionally, the predictive variance can be estimated, which can serve as a measure of how well a material of interest is represented in the training set. We use nested cross-validation for the model selection process [22,23]; that is, the parameter selection and performance evaluation are performed on separate held-out subsets of the data that are independent from the set of training materials. This ensures we find optimal parameters for the kernel and the model regularization in terms of generalization while avoiding overfitting.

In the solid-state community crystals are conventionally described by the combination of the *Bravais matrix*, containing the primitive translation vectors, and the *basis*, setting the position and type of atoms in the unit cell. This type of description is not unique and thus not a suitable representation for the learning process since it depends on an arbitrary choice of the coordinate system in which the Bravais matrix is given. Namely, there exists an infinite number of equivalent representations that would be perceived as distinct crystals by the machine. In principle, recognizing equivalent representations could also be tackled by machine learning directly as done for molecules in Refs. [16,24,25]. However, a significant computational cost in terms of size of the training set had to be paid. Due to the aforementioned larger ambiguity in the case of crystals, an even higher cost is expected.

For the case of molecules the *Coulomb matrix* has proven to be a well-performing representation [15,16]. This is given by

$$C_{ij}^{\text{mol}} = \begin{cases} 0.5 Z_i^{2.4} & \text{for } i = j, \\ \frac{Z_i Z_j}{\|\mathbf{r}_i - \mathbf{r}_j\|} & \text{for } i \neq j, \end{cases}$$

with nuclear charges Z_i and positions $\mathbf{r}_i$ of the atoms. This description is invariant under rotation and translation, but unfortunately, it cannot be applied directly to infinite periodic crystals.

A simple extension to crystals is to combine a Coulomb matrix of one single unit cell with the Bravais matrix (B+CM). However, this representation suffers from the *degeneracy problem* mentioned above. The Coulomb matrix representation assumes a similarity relation between atoms with close nuclear charges. However, this is most often not the case for the chemical properties.

In order to include more physical knowledge about crystals, we propose a crystal representation inspired by radial distribution functions as used in the physics of x-ray powder diffraction [26] and text mining from computer science [27,28]. The

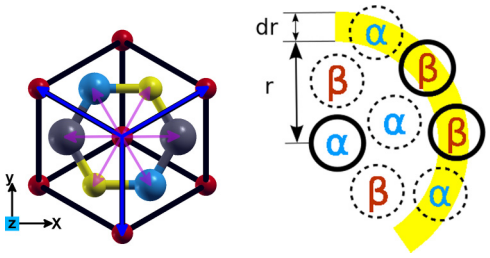

FIG. 1. (Color online) Alternative crystal representations. (left) A crystal unit cell indicating the Bravais vectors (blue) and base (pink). (right) Illustration of one shell of the discrete partial radial distribution function $g_{\alpha\beta}(r)$ with width dr.

partial radial distribution function (PRDF) representation considers the distribution of pairwise distances $d_{\alpha\beta}$ between two atom types, α and β. This can be seen as the density of atoms of type β in a shell of radius r and width dr centered around an atom of type α (see Fig. 1). Averaged over all atoms of a type, the discrete (PRDF) representation is given by

$$g_{\alpha\beta}(r) = \frac{1}{N_\alpha V_r} \sum_{i=1}^{N_\alpha} \sum_{j=1}^{N_\beta} \theta(d_{\alpha_i\beta_j} - r)\theta(r + dr - d_{\alpha_i\beta_j}),$$

where N_α and N_β are the numbers of atoms of types α and β, respectively, while V_r is the volume of the shell. We only need to consider the atoms in one primitive cell as shell centers for calculation. The distribution is globally valid due to the periodicity of the crystal and the normalization with respect to the considered crystal volume. In this work, the type criterion for "counting" an atom is its nuclear charge; however, other more general criteria could, in principle, also be used, such as the number of valence electrons or the electron configuration.

As input for the learning algorithm, we employ the feature matrix X with entries $x_{\alpha\beta,n} = g_{\alpha\beta}(r_n)$, i.e., the PRDF representation of all possible pairs of elements as well as shells up to an empirically chosen cutoff radius. The distance of two crystals is then defined as the distance induced by the Frobenius norm between those matrices and may be plugged into one of the previously described kernels. In this manner, we have defined a global descriptor as well as a similarity measure for crystals which is invariant under translation, rotation, and the choice of the unit cell.

The DOS_F we use to train and validate the learning are computed [29] for crystals from the inorganic crystal structure database (ICSD) [35] with the experimental lattice parameters reported therein. The chosen subset contains only nonduplicated materials with a maximum of six atoms per primitive cell. We subdivide the set into *sp* (1716 crystals) and *spd* (5548 crystals).

For the DOS_F prediction, we first consider the *sp* and *spd* material sets separately. The mean absolute errors of the predictions of all presented crystal representations are collected in Table I. Furthermore, we list the mean predictor that always predicts the average DOS_F value of the training set as a simple baseline. Both representations yield models that are significantly better than the mean predictor. Figure 2 illustrates how the error decreases steadily with an increasing number of materials used for training. However, the PRDF

TABLE I. Mean absolute errors and standard errors of DOS predictions in 10^{-2} states/eV/Å^3. Errors in bold indicate the best performance.

Predictor	Features	*sp* systems	*spd* systems
Mean predictor		1.50 ± 0.02	1.82 ± 0.03
KRR (linear)	B+CM	1.45 ± 0.04	1.68 ± 0.01
KRR (Gaussian)	B+CM	1.19 ± 0.03	1.62 ± 0.01
KRR (Laplacian)	B+CM	1.20 ± 0.04	1.63 ± 0.02
KRR (linear)	PRDF	0.87 ± 0.02	1.68 ± 0.03
KRR (Gaussian)	PRDF	0.74 ± 0.03	0.95 ± 0.02
KRR (Laplacian)	PRDF	$\mathbf{0.68} \pm 0.03$	$\mathbf{0.86} \pm 0.01$

features consistently outperform the B+CM description. The further analysis will therefore focus on PRDF with the slightly better performing Laplacian kernel.

The higher complexity of the *spd* systems can clearly be observed in the learning curves, which show how much better the prediction problem can be solved as a function of the available data. The mean error is much lower in *sp* materials. Furthermore, the learning curves are steeper; that is, increasing the training set size within the restricted materials class improves the prediction accuracy rapidly. One origin of this higher complexity lies in the growing dimensionality of the input space: given N_{el} possible chemical elements in all material compositions, $\dim(\mathrm{X}) \propto N_{\mathrm{el}}^2$. Furthermore, by including materials with *d* electrons, the physics becomes richer. For both reasons, much more training data are required to achieve an improvement comparable to that of *sp* systems.

The prediction of $\mathrm{DOS_F}$ for *spd* systems is shown in Fig. 3 as a density plot of computed versus predicted values. It is evident that the density is accumulated along the diagonal of the plot, demonstrating that the machine is giving meaningful predictions. However, the average error is smaller than 6% of the DOS value range. Thus, this result represents a proof of principle that a complex output of the Kohn-Sham equations can be predicted directly by means of machine learning, despite the considerable variance of errors.

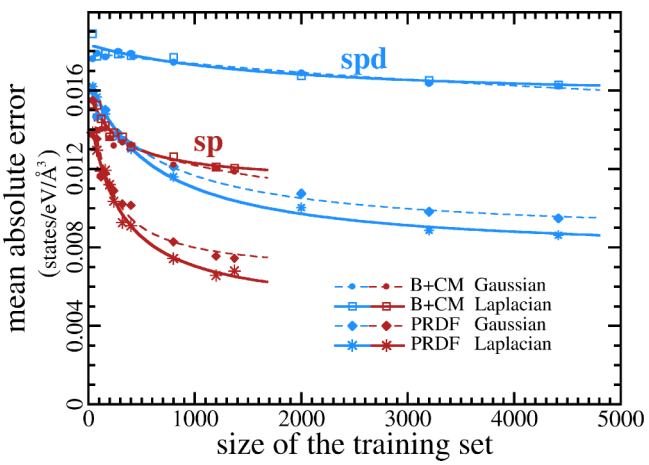

FIG. 2. (Color online) Learning process as a function of the number of materials used for training for all three feature representations (conventional B+CM and PRDF) and for the two data sets.

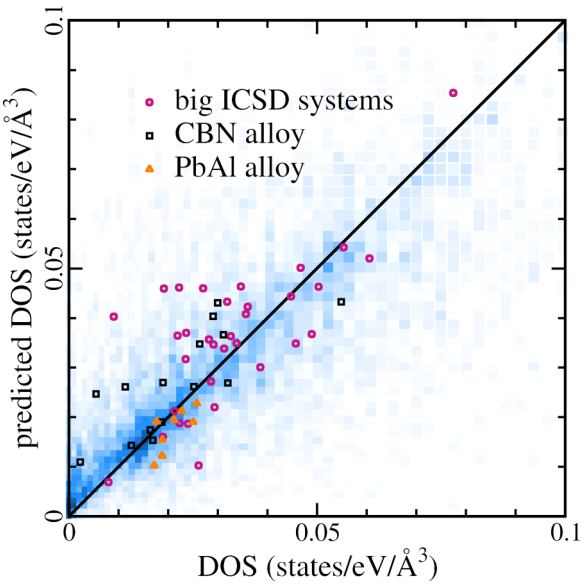

FIG. 3. (Color online) Comparison between predicted and calculated $\mathrm{DOS_F}$ for *spd* systems. The background density distribution refers to the cross-validation systems. Dots are additional systems (see legend) of far larger size than those used for training.

III. PREDICTIONS FOR LARGER SYSTEMS

From Fig. 2, it is clear that, in order to increase the prediction accuracy, the size of the training set should be extended, possibly at the limits of present computing facilities. Instead of a brute-force approach, the problem may become less costly by using an active learning scheme, e.g., by extending the set where the *predicted variance* is higher. We still expect that in order to obtain highly accurate results the computation costs will be large. Can the proposed approach still be useful at the present accuracy level?

To answer this question we first point out that ML is at least 2 or 3 orders of magnitude faster than any direct computational approach; this means that a fast ML scan may always be of use as a preliminary step before consuming precious resources on detailed calculations. Second, a remarkable feature of the PRDF representation is that it is not fixed to a certain number of atoms in the unit cell of the training materials. This means that once the machine has been trained, it can be used to predict the properties of any other system. This is virtually independent of its size as long as it is well represented by the training set.

As a proof for this ability, we consider additional test systems, divided into three sets: The first set (pink circles in Fig. 3) contains only systems taken from the ICSD database, chosen among those with between 30 and 80 atoms per unit cell that are well represented by the training set. Therefore, only ICSD materials with a relatively low predictive variance were chosen for calculation. The second set (orange triangles in Fig. 3) contains a purely metallic alloy (within the unit cell) of lead and aluminum. All the systems in this set are crystals with 125 atoms per unit cell, differing by the Al/Pb concentration. The third set (black squares in Fig. 3) is a solid solution of three atomic types in a diamond lattice, carbon,

boron, and nitrogen, at different compositions and with a total of 45 atoms per periodic unit cell.

Unlike the training systems, each of these involve a large computational cost and would not be feasible without access to a computation facility. While the PbAl alloys are quite well predicted, some of the DOS_F for the large ICSD systems (mostly oxides) are overestimated, as well as some of the DOS_F of the CBN solid solution. Again, a clear diagonal accumulation is achieved. Nevertheless, for these large systems that the learning machine has never been trained on, the average quality of the prediction of large systems is comparable to that of the much smaller, cross-validated systems.

IV. CONCLUSION

We have investigated a machine learning approach for fast solid-state predictions. A set of local spin-density approximation calculations has been used to train a DOS_F predictor. We expect that our method can be extended to directly predict other complex materials properties as well. It certainly can be combined with other, more accurate, electronic structure techniques such as GW. The accuracy of predictions depends strongly on how crystals are represented. We found that Coulomb matrices, while being successful for predicting properties in molecules [24,25], are not suitable to describe crystal structures well enough. Instead, we have proposed a representation inspired by partial radial distribution functions which is invariant with respect to translation, rotation, and the choice of the unit cell. Our results clearly demonstrate that a fast prediction of electronic properties in solids with ML algorithms is indeed possible. Although currently the accuracy leaves room for improvement, we consider the predictions useful for a first screening of a huge number of materials for properties within a desired value range. In a second step, high-accuracy electronic structure calculations are then performed on the promising candidates only. What makes the approach extremely appealing is that the PRDF representation allows us to learn on small systems with low computational cost and then extrapolate to crystals with arbitrary numbers of atoms per unit cell for which conventional DFT calculations would be prohibitive.

ACKNOWLEDGMENTS

F.B. and K.R.M. gratefully acknowledge helpful discussions with M. Scheffler, C. Draxl, S. Levchenko, and L. Ghiringhelli, who pointed out to us that for electron densities and band gaps the local topology and connectivity of the atoms are appropriate descriptors and not the Coulomb matrix. Furthermore, we acknowledge valuable comments from A. Tkatchenko, K. Hansen, and A. von Lilienfeld. K.R.M., K.S., and F.B. thank the Einstein Foundation for generously funding the ETERNAL project.

[1] S. Curtarolo, A. N. Kolmogorov, and F. H. Cocks, Calphad **29**, 155 (2005).
[2] A. R. Oganov and C. W. Glass, J. Chem. Phys. **124**, 244704 (2006).
[3] C. J. Pickard and R. J. Needs, J. Phys. Condens. Matter **23**, 053201 (2011).
[4] D. Morgan, G. Ceder, and S. Curtarolo, Meas. Sci. Technol. **16**, 296 (2005).
[5] K. Kang, Y. S. Meng, J. Bréger, C. P. Grey, and G. Ceder, Science **311**, 977 (2006).
[6] H. Chen, G. Hautier, A. Jain, C. Moore, B. Kang, R. Doe, L. Wu, Y. Zhu, Y. Tang, and G. Ceder, Chem. Mater. **24**, 2009 (2012).
[7] G. Hautier, A. Jain, T. Mueller, C. Moore, S. P. Ong, and G. Ceder, Chem. Mater. **25**, 2064 (2013).
[8] S. Keinan, M. J. Therien, D. N. Beratan, and W. Yang, J. Phys. Chem. A **112**, 12203 (2008).
[9] R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sanchez-Carrera, L. Vogt, and A. Aspuru-Guzik, Energy Environ. Sci. **4**, 4849 (2011).
[10] H. Peng, A. Zakutayev, S. Lany, T. R. Paudel, M. d'Avezac, P. F. Ndione, J. D. Perkins, D. S. Ginley, A. R. Nagaraja, N. H. Perry, T. O. Mason, and A. Zunger, Adv. Funct. Mater. **23**, 5267 (2013).
[11] A. Jain, S.-A. Seyed-Reihani, C. C. Fischer, D. J. Couling, G. Ceder, and W. H. Green, Chem. Eng. Sci. **65**, 3025 (2010).
[12] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[13] O. A. von Lilienfeld, M. Rupp, and A. Knoll, arXiv:1307.2918.
[14] M. L. Braun, J. M. Buhmann, and K.-R. Müller, J. Mach. Learn. Res. **9**, 1875 (2008).
[15] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[16] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, O. A. von Lilienfeld, and K.-R. Müller, in *Advances in Neural Information Processing Systems*, Vol. 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc., Red Hook, NY, 2012), pp. 440–448.
[17] Z. Pozun, K. Hansen, D. Sheppard, M. Rupp, K.-R. Müller, and G. Henkelman, J. Chem. Phys. **136**, 174101 (2012).
[18] J. Behler, J. Chem. Phys. **134**, 074106 (2011).
[19] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, Phys. Rev. Lett. **108**, 253002 (2012).
[20] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, IEEE Trans. Neural Networks **12**, 181 (2001).
[21] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond* (MIT Press, Cambridge, MA, 2002).
[22] T. Hastie, R. Tibshirani, and J. J. H. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2001).

[23] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Muller, J. Chem. Theory Comput. **9**, 3404 (2013).
[24] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, New J. Phys. **15**, 095003 (2013).
[25] G. Montavon, M. L. Braun, T. Krueger, and K.-R. Müller, IEEE Signal Process. Mag. **30**, 62 (2013).
[26] S. J. Billinge and M. Thorpe, *Local Structure from Diffraction* (Springer, Berlin, 1998).
[27] G. Forman, J. Mach. Learn. Res. **3**, 1289 (2003).
[28] T. Joachims, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features* (Springer, Berlin, 1998).
[29] All calculations are performed within KS spin density functional theory [30,31], with local spin-density approximation exchange-correlation functional [32]. Core states are accounted for in the pseudopotential approximation as implemented in the ESPRESSO package [33]. k points are sampled with a Monkhorst-Pack grid [34] with a density of about 1510 Å^3. Magnetic ordering is assumed to be ferromagnetic.
[30] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[31] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[32] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[33] P. Giannozzi *et al.*, J. Phys. Condens. Matter **21**, 395502 (2009).
[34] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).
[35] Inorganic Crystal Structure Database, Version 2010-2, Fachinformationszentrum Karlsruhe, http://www.fiz-karlsruhe.de/icsd_home.html