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, Marchstr. 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*

(Dated: August 26, 2013)

High-throughput density-functional calculations of solids are extremely time consuming. As an alternative, we here propose a machine learning approach for the fast prediction of solid-state properties. To achieve this, LSDA calculations are used as training set. We focus on predicting metallic vs. insulating behavior, and 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. Due to magnetic phenomena learning on d systems is found more difficult than in pure sp systems.

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 having trained the ML model 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 physicists 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]. Recently, ML has contributed accurate models for predicting molecular properties [13, 14], transition states [15], reaction surfaces [16], potentials [17] and self-consistent solutions for DFT [18]. All these applications deal with finite systems (atoms, molecules, clusters). For this type of systems, 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 Siebeck coefficient, 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 answer are: (a) How can we describe an infinite periodic system in a way that supports the learning process well? (b) How large should the basis for ML training be, i.e., the *training set* of calculations? Answering these questions will provide us exactly with 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's computing power.

We employ so-called kernel-based learning methods [19, 20] that are based on a mapping to a high-

dimensional *feature space* such that an accurate prediction can be achieved with a linear model in this space. The so-called *kernel trick* allows to perform this mapping implicitly using a kernel function, e.g., the Gaussian kernel $k(\mathbf{x},\mathbf{y}) = \exp\left(-\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2\right)$. Kernels can be viewed as a similarity measure between data, in our case they should measure 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 small distance when their properties are similar is essential for the learning process (see below for details).

For the insulator vs. metal classification, we use a *support vector machine (SVM)* that finds a separating hyperplane in feature space while maximizing the space between the two classes [21]. 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.

We use nested cross-validation for the model selection process [22], i.e., 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 to 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 the 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 Ref. [14, 23, 24]. However, a significant computational cost in terms of size of the training set had to be paid, particularly in the case of crystals.

For the case of molecules the *Coulomb matrix* has proven to be a well-performing representation [13, 14]. 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 would be to combine a Coulomb matrix of one single unit cell with the Bravais matrix. We call this representation *Bravais + Coulomb matrix* (B+CM).

In order to avoid the above discussed *degeneracy problem* associated with the Bravais description, a direct generalization to crystals can be formulated by fixing one atom in the crystal and taking the nearest k atoms to build up the Coulomb matrix from those. We then sort the matrix by the nuclear charges, i.e., $C_{ii} < C_{jj}$ for $i < j$. This leads to a representation that takes the periodicity of the crystal directly into account. We call this second candidate representation *Crystal Coulomb matrix* (CCM). The Coulomb matrix representation assumes a similarity relation between atoms with close nuclear charges. However, this is not necessarily the case for materials.

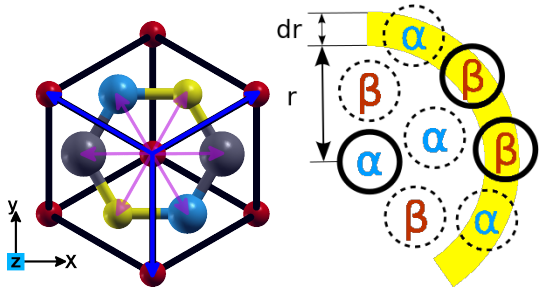

FIG. 1. Alternative crystal representations. Left: a crystal unit cell with indicated the Bravais vectors (blue) and base (pink). Right: Illustration of one shell of the discrete partial radial distribution function $g_{\alpha\beta}(r)$ with width dr.

In order to include more physical knowledge about crystals, we propose the partial radial distribution function *(PRDF) representation* inspired by radial distribution functions as used in the physics of x-ray powder diffraction [25] and text mining from computer science [26, 27]. It considers the distribution of pair-wise distances $d_{\alpha\beta}$ between two atom types α and β, respectively. 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 type 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 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 cut-off 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 the previously described Gaussian kernel. In this manner, we have defined a novel global descriptor as well as a similarity measure for crystals which is invariant under translation, rotation and the choice of the unit cell.

The $\mathrm{DOS_F}$ we use to train and validate the learning are computed [28] on crystals from the inorganic crystal structure database (ICSD) [29] with the experimental lattice parameters reported therein. The chosen subset contains only non-duplicated materials with a maximum of 8 atoms per pr/imitive cell. We subdivide the set into *sp* (1736 crystals of which 1151 are insulators) and *spd* (5134 crystals of which 1979 are insulators).

Since the $\mathrm{DOS_F}$ is only clearly defined for metals, its prediction is closely linked to the classification of materials into metals and insulators. Thus as a first step, we trained an SVM classifier on the whole dataset. By shifting the classification threshold, sensitivity vs. specifity, i.e., the trade-off between correctly detected insulators and metals incorrectly classified as insulators can be adjusted. E.g., our classifier is able to detect 85.0% of the insulators while only mistaking 7.3% of the metals as insulators on the whole data set.

TABLE I. Mean absolute errors in DOS predictions in $10^{-2}\,\mathrm{states/eV/Å^3}$

Predictor	**sp systems**	**spd systems**
Mean predictor	1.49	1.99
KRR / B+CM	1.02	1.65
KRR / CCM	1.11	1.64
KRR / PRDF	0.78	1.19

For the $\mathrm{DOS_F}$ prediction, we only consider metals in the sp and spd material sets. 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 $\mathrm{DOS_F}$ value of the training set as a simple baseline. Fig. 2 illustrates how the error decreases steadily with increasing number of materials used for training. All three representations yield models that are significantly better than the mean predictor. However, the PRDF features consistently outperform the CCM and the B+CM description. The further analysis will therefore focus on PRDF.

We note that the B+CM description does not appear to be worse than the CCM. This could be a result of the intrinsic conventional crystallographic choices within the ICSD. There are conventions that partially restrict alternative representations, e.g., hexagonal crystals are usually represented with the hexagonal axis along the z Cartesian direction, so the degeneracy problem affects only the xy plane.

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, i.e., 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 more rich. Due to both reasons, much more training data is required to achieve an improvement comparable to that of sp systems.

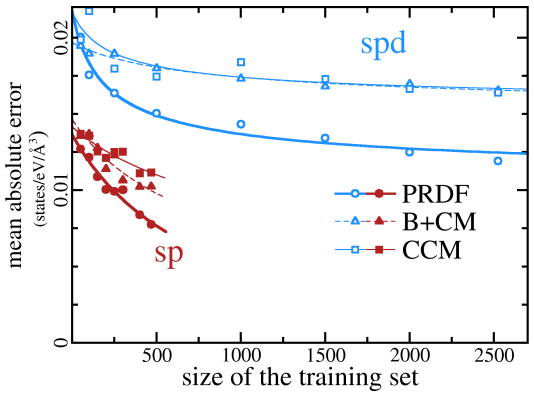

FIG. 2. Learning process as a function of the number of materials used for training for all three feature representations (conventional CCM and PRDF), and for the two datasets.

The prediction of $\mathrm{DOS_F}$ is shown in Fig. 3, as a density plot of computed versus predicted values. In both sp and spd systems the density is clearly accumulated along the diagonal of the plot, demonstrating that the machine is giving meaningful predictions. In the case of sp systems, the upper limit of the absolute deviation from the diagonal seems to be independent from the value of $\mathrm{DOS_F}$ by itself, implying a comparatively low *relative* prediction error for high $\mathrm{DOS_F}$.

As a matter of fact, predictions on sp materials exhibit a considerably lower upper limit of the absolute deviation compared to spd systems. In the spd set, we observe some severe mispredictions for high $\mathrm{DOS_F}$, where the machine shows a tendency to underestimate the DOS. We investigated this aspect by a detailed analysis of all examples with a $\mathrm{DOS_F} \geq 0.15\,\mathrm{states/eV/Å^3}$ [30], comparing examples with a relative prediction error $\delta\mathrm{DOS_F} < 25\%$ (called 'well predicted' in the following) to the remaining ones (called 'badly predicted' in the following). First of all, 18% of the badly predicted materials contain pairs of

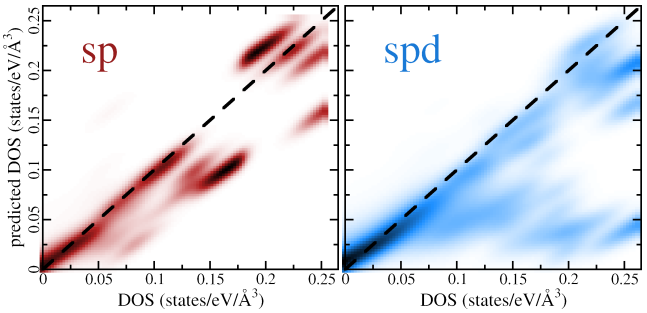

FIG. 3. Comparison between predicted and calculated DOS_F for sp (left) and spd systems (right).

chemical elements not found in any other material, which is important due to the PRDF representation. Taking into account the 5 training examples of highest influence onto each prediction [31], we observe that in every material with $\delta DOS_F < 25\%$ contains at least one material with the same chemical composition at a close, but nevertheless different pressure (78%), or a material whose chemical composition is a superset (50%). Only 30% of the materials with $\delta DOS_F > 25\%$ have such contributors.

Magnetic phase transitions are a further important aspect. At the crossing of a transition we can have crystals of similar composition and lattice structure while having a large difference in DOS_F (as the magnetic transition opens a gap in the d electronic states at the Fermi energy causing a drop in the value of DOS_F). This high sensitivity of the DOS to the lattice properties is clearly difficult to learn and leads to evident mispredictions. Since there are only few training materials with a high DOS_F the misprediction is usually in the direction of an underestimation.

In summary, we have investigated a machine learning approach for fast solid-state predictions. A set of LSDA spin-DFT calculations has been used to train a DOS_F predictor and a metal/insulator classifier. Prediction quality strongly depends on how crystals are represented. We found that Coulomb matrices, while being successful for predicting properties in molecules [23, 24], are not suitable to describe crystal structures well enough. Instead, we have proposed an invariant representation inspired by partial radial distribution functions. While the learning curves suggest significant improvement for the more restricted class of sp systems by increasing the number of training materials, learning in d systems, although still clearly visible, is much slower due to the high dimensionality of the chemical compound subspace and the complexity of magnetic phenomena. Our results clearly demonstrate that a fast prediction of electronic properties in solids with ML algorithms is indeed possible. The suggested representation of periodic solids is rather generic. We expect that our method can be extended beyond the DOS_F to directly predict other and more complex materials properties.

F. Brockherde and K.-R. Müller gratefully acknowledge helpful discussions with Matthias Scheffler, Claudia Draxl, Sergey Levchenko and Luca Ghiringhelli, who pointed out to us that for electron densities and band gaps the local topology and connectivity of the atoms is an appropriate descriptor and not the Coulomb matrix. Their explanation was based on tight-binding theory and work by D. Pettifor. Furthermore we acknowledge valuable comments of Alexandre Tkatchenko, Katja Hansen and Anatole von Lilienfeld. KRM, KS and FB thank the Einstein Foundation for generously funding the ETERNAL project.

* K.T. Schütt and H. Glawe contributed equally to this work.
† Corresponding authors; these authors jointly directed the project.

[1] S. Curtarolo, A. N. Kolmogorov, and F. H. Cocks, Calphad **29**, 155 (2005).
[2] A. R. Oganov and C. W. Glass, The Journal of Chemical Physics **124**, 244704 (2006).
[3] C. J. Pickard and R. J. Needs, Journal of Physics: Condensed Matter **23**, 053201 (2011).
[4] D. Morgan, G. Ceder, and S. Curtarolo, Measurement Science and Technology **16**, 296 (2005).
[5] K. Kang, Y. S. Meng, J. Brger, C. P. Grey, and G. Ceder, Science **311**, 977 (2006), http://www.sciencemag.org/content/311/5763/977.full.pdf.
[6] H. Chen, G. Hautier, A. Jain, C. Moore, B. Kang, R. Doe, L. Wu, Y. Zhu, Y. Tang, and G. Ceder, Chemistry of Materials **24**, 2009 (2012), http://pubs.acs.org/doi/pdf/10.1021/cm203243x.
[7] G. Hautier, A. Jain, T. Mueller, C. Moore, S. P. Ong, and G. Ceder, Chemistry of Materials **25**, 2064 (2013), http://pubs.acs.org/doi/pdf/10.1021/cm400199j.
[8] S. Keinan, M. J. Therien, D. N. Beratan, and W. Yang, The Journal of Physical Chemistry A **112**, 12203 (2008), http://pubs.acs.org/doi/pdf/10.1021/jp806351d.
[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. a. d'Avezac, P. F. Ndione, J. D. Perkins, D. S. Ginley, A. R. Nagaraja, N. H. Per ry, T. O. Mason, and A. Zunger, Advanced Functional Materials , n/a (2013).
[11] A. Jain, S.-A. Seyed-Reihani, C. C. Fischer, D. J. Couling, G. Ceder, and W. H. Green, Chemical Engineering Science **65**, 3025 (2010).
[12] M. L. Braun, J. M. Buhmann, and K.-R. Müller, The Journal of Machine Learning Research **9**, 1875 (2008).
[13] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Physical Review Letters **108**, 058301 (2012).
[14] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. von Lilienfeld, and K.-R. Müller, in *Advances in Neural Information Processing Systems 25* (2012) pp. 449–457.
[15] Z. A. Pardos and N. T. Heffernan, Journal of Machine Learning Research W & CP (2010).
[16] Z. Pozoun, K. Hansen, D. Sheppard, M. Rupp, K.-R.

Müller, and G. Henkelman, Journal of Chemical Physics **136**, 174101 (2012).
[17] J. Behler, The Journal of chemical physics **134**, 074106 (2011).
[18] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, Physical Review Letters **108**, 253002 (2012).
[19] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, Neural Networks, IEEE Transactions on **12**, 181 (2001).
[20] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization and beyond* (the MIT Press, 2002).
[21] C. Cortes and V. Vapnik, Machine learning **20**, 273 (1995).
[22] T. Hastie, R. Tibshirani, and J. J. H. Friedman, *The elements of statistical learning* (Springer New York, 2001).
[23] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, New Journal of Physics (2013), to appear. arXiv:1305.7074.
[24] G. Montavon, M. L. Braun, T. Krueger, and K.-R. Müller, IEEE Signal Processing Magazine **30**, 62 (2013).
[25] S. J. Billinge and M. Thorpe, *Local structure from diffraction* (Springer, 1998).
[26] G. Forman, The Journal of machine learning research **3**, 1289 (2003).
[27] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features* (Springer, 1998).
[28] All calculations are performed within KS spin density functional thory [**?** **?**], with LSDA xc [**?**]. Core states are accounted in the pseudo-potential approximation as implemented in the ESPRESSO package [**?** **?**] k-points are sampled with a Monkhorst-Pack grid [**?**] with a density of about nn points*Omega. Magnetic ordering is assumed to be ferro-magnetic.
[29] ICSD, Inorganic Crystal Structure Database, Fachinformationszentrum Karlsruhe: Karlsruhe, Germany, (2011).11.
[30] The high $\mathrm{DOS_F}$ values themselves arise from materials with overestimated lattice parameters in ICSD, which is unrelated to our intention to predict LSDA results.
[31] The magnitude of the coefficients exhibits an exponential decay.