

Predicting Thermal Properties of Crystals Using Machine Learning

Sherif Abdulkader Tawfik,* Olexandr Isayev,* Michelle J. S. Spencer,* and David A. Winkler*

Calculating vibrational properties of crystals using quantum mechanical (QM) methods is a challenging problem in computational material science. This problem is solved using complementary machine learning methods that rapidly and reliably recapitulate entropy, specific heat, effective polycrystalline dielectric function, and a non-vibrational property (band gap) for materials calculated by accurate but lengthy QM methods. The materials are described mathematically using property-labeled materials fragment descriptors. The machine learning models predict the QM properties with root mean square errors of 0.31 meV per atom per K for entropy, 0.18 meV per atom per K for specific heat, 4.41 for the trace of the dielectric tensor, and 0.5 eV for band gap. These models are sufficiently accurate to allow rapid screening of large numbers of crystal structures to accelerate material discovery.

Dr. S. A. Tawfik, Prof. M. J. S. Spencer School of Science **RMIT University** GPO Box 2476, Melbourne, Victoria 3001, Australia E-mail: sherif.abbas@rmit.edu.au; michelle.spencer@rmit.edu.au Prof. O. Isavev Laboratory for Molecular Modeling Division of Chemical Biology and Medicinal Chemistry UNC Eshelman School of Pharmacy University of North Carolina Chapel Hill, NC 27599, USA E-mail: olexandr@olexandrisayev.com Prof. D. A. Winkler School of Pharmacy University of Nottingham Nottingham, NG7 2RD, UK E-mail: d.winkler@latrobe.edu.au Prof. D. A. Winkler Monash Institute of Pharmaceutical Sciences Monash University 381 Royal Parade, Parkville, Victoria 3052, Australia Prof. D. A. Winkler Latrobe Institute for Molecular Science La Trobe University Kingsbury Drive, Bundoora, Victoria 3086, Australia Prof. D. A. Winkler **CSIRO** Manufacturing Bayview Avenue, Clayton, Victoria 3168, Australia

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adts.201900208

DOI: 10.1002/adts.201900208

1. Introduction

Machine learning (ML) is a powerful tool for high-throughput prediction of properties of materials.^[1] The entire machine learning work-flow, starting from the mathematical descriptors^[2] that can accurately and succinctly capture the properties of materials, up to the choice of specific machine learning algorithms, has been undergoing important development recently. High throughput, automated synthesis, and testing technologies now generate large materials databases, providing increasingly rich datasets for training data-driven ML models that can be used to design and optimize new bespoke

materials. These developments provide a novel approach for the calculation of the properties of many important materials for applications in photonics,^[3] medicine,^[4] superconductivity,^[5] energy,^[6] and programmable materials^[7] and many other areas.

Recently, quantum chemical methods such as densityfunctional theory (DFT) have accelerated the pace of materials discovery.^[8] A number of very large databases of quantum mechanical (QM)-computed values of properties of materials are now available, for example, AFLOW^[9] (\approx 3 million materials) and Materials Project^[10] (≈120 000 materials). The Inorganic Crystal Structure Database (ICSD)^[11] is a subset of these databases that contains materials with experimental structural data. The rest of the materials in these online databases were generated by in silico methods that find thermodynamically stable structures. While these databases have a large range of calculated electronic and structural properties for each material, they only have a few systems in which the vibrational properties are calculated. Calculation of vibrational properties of large numbers of materials is challenging because the computational expense is much greater than that for calculation of ground state their properties. Therefore, new materials discovery will be greatly aided by the development of a rapid method for calculating the dynamical properties of materials. Here, we show how ML can be used for this purpose.

Legrain et al.^[2b] reported the prediction of two vibrational properties, the vibrational free energy and the entropy, by ML methods, using a relatively small data set of 292 materials. The structures were obtained from the ICSD section of the aflow.org repositories, with additional phonon calculations being performed. Legrain et al. compared the prediction accuracy of ML models trained using a very simple set of descriptors based on the atomic molar composition of the material, with more sophisticated descriptors. Surprisingly, they found that the simple composition descriptors outperformed the other descriptors in prediction accuracy. Heat capacity can also be predicted with reasonable accuracy using the empirical Neumann–Kopp rules that rely on the thermal properties of the constituent elements.^[12] Kauwe et al.^[13] compared the performance of several ML models with the Neumann–Kopp rules using descriptors that include the chemical formula, temperature, and experimental thermochemical properties, and reported that the ML models were only slightly more accurate. Given that the data set utilized by Kauwe et al. is small and their descriptors were not derived from the atomic structure of the materials, we surmised that more accurate and scalable ML models should result when descriptors derived from the atomic structure are employed.

CIENCE NEWS

www.advancedsciencenews.com

Recently Petretto et al.^[14] reported a larger data set of calculated vibrational properties for 1521 materials. This data set is available online and contains the results of density-functional perturbation theory calculations of a range of vibrational scalar and tensor quantities: the entropy, the Helmholtz free energy, the phonon energy, the heat capacity, the dielectric tensor, the phonon frequencies, and the phonon dispersion. Such a rich data set is very useful for assessing how well ML models can predict vibrational properties of diverse materials. Importantly, the calculated vibrational entropy in this dataset compares well with the experimental values at room temperature, with a calculated mean absolute relative error (MARE) in predicting the experimental data of <3%. Therefore, the dataset provides a reliable reference with which we can build ML models.

Here, we use mathematical descriptors based on the propertylabeled materials fragments (PLMF) method to accurately represent the molecular properties of 1521 crystalline semiconductors. We show that a small set of PLMF descriptors and an appropriate ML method can yield accurate predictions of properties of materials in both the training set used to generate the models, and in an independent test set not used in model building. These results represent an important step toward the high throughput (HT) screening of phase stability at finite temperatures, with important applications for the computational discovery of new functional compounds such as hydrogen storage materials.^[6b]

2. Computational Details

2.1. Data Set

We predicted three important vibrational quantities: entropy, *S*; specific heat, *C*_V; and the dielectric function, ϵ . The entropy and the specific heat quantities are scalar, while the dielectric function is a tensor. These quantities were calculated from the ground state phonon frequencies and modes according to Equations (1) and (3), in ref. 14, at 298 K. The quantity ϵ is a second-rank tensor (a 3 × 3 matrix). Unlike the scalars, ϵ is dependent upon the geometric orientation of the crystal. That is, the values of the matrix elements of ϵ will change when the crystal is rotated. Given that the off-diagonal elements of the dielectric tensor are very small in the data set, we attempted to model an invariant of the matrix, the effective polycrystalline dielectric function, ϵ_{eff} , defined according to the following equation:^[15]

$$\varepsilon_{\rm eff} = \frac{3\varepsilon_x \varepsilon_y \varepsilon_z}{\varepsilon_x \varepsilon_y + \varepsilon_x \varepsilon_z + \varepsilon_y \varepsilon_z} \tag{1}$$

where ε_i is the *i* component of the dielectric tensor, i = x, y, z. There have been very few reports of ML models of tensor properties to date, to the best of our knowledge.^[16]

We removed structures with imaginary phonon frequencies that would be dynamically unstable from the original data set, leaving 1245 materials. For ϵ , we removed all records with excessive $\epsilon_{\rm eff}$ (we set the threshold ${\rm Tr}[\epsilon]$ <100), leaving a final data set of 1228 materials. Due to the large dynamic range of these properties, we modeled the logarithm of these quantities as our target properties, as is traditional in ML studies, and then transformed the predictions back using the antilog function to obtain predictions of the properties themselves.

We also used this data set to predict another important scalar quantity: the electronic band gap. This property has been predicted accurately by ML models previously for a larger data set.^[2a] The accurate prediction of the band gap for the present data set further illustrates the capabilities and generality of our ML approach.

2.2. PLMF Descriptors

The most important factor for a successful machine learning model is the choice of descriptors. In this work, we adopt the descriptors of Isayev et al.,^[2a] previously shown to be useful for modeling crystalline materials. The materials structure is represented as a graph, with vertices annotated according to the properties of the atoms they represent and nodes connecting topological neighbors generated by Voronoi tessellations. The adjacency matrix of this graph determines the global topology for a given system, including interatomic bonds and contacts within a molecule. The final descriptor vector for the ML models is obtained by partitioning a full graph into smaller subgraphs, or fragments. Each fragment starts from a node (an atom and its properties) and captures a path of specified length in the graph through a collection of bonded atoms. Each of the fragments is characterized by general atomic measured and derived properties, such as the Mendeleev properties, effective nuclear charge, and atomic polarizability.[2a]

We used sparse feature selection to choose the smallest number of descriptors that produce a strong model because optimally sparse models are best at predicting the properties of new data. A common approach to eliminating features is to describe their relative importance to a model, then eliminate weak features or combinations of features, and re-evaluate to see if the model more accurately predicts properties of materials in a crossvalidation procedure or test set. We used the parsimonious least absolute shrinkage and selection operator (LASSO) algorithm to select descriptors.^[17] This sets less relevant descriptors to zero, leaving only those with significant contribution to models. To obtain the optimal number of descriptors, we varied the sparsity parameter α to identify a sparse set of descriptors that yielded the highest R^2 and lowest root mean square error values in the LASSO models. We then used these descriptors to predict properties using several ML modeling methods.

The most important quality of a model is how well it predicts the properties of new materials that the model has not encountered; that is, how accurately can the model generalize to new outcomes based on its learning. This can be assessed by randomly partitioning the data set into a training set, used to build the ML model, and a test set used only to assess the ability of the model to predict properties for new materials. The prediction accuracy is best judged by loss functions or measures of dispersion, such as the root mean square error (RMSE) and MARE, although we also calculate the coefficient of determination between predicted and measured values (R^2).^[18]

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}$$
(2)

MARE =
$$\frac{1}{n} \sum_{i=1}^{N} \frac{|Y_i - \hat{Y}_i|}{|Y_i|} \times 100$$
 (3)

$$R^{2} = 1 - \frac{\sum_{i=1}^{N} (Y_{i} - \hat{Y}_{i})^{2}}{\sum_{i=1}^{N} (Y_{i} - \bar{Y})^{2}}$$
(4)

where Y_i is the known test set outcomes, \hat{Y}_i is the predicted test set outcomes, and \bar{Y} is the average of the known test set outcomes. The significance of R^2 is that it expresses the proportion of the variance in \hat{Y}_i that can be predicted from the descriptor vector, and is another important measure of the ML model quality. However, as its values are dependent on the size of the data set,^[18] we reported all three quantities together, R^2 , MARE, and RMSE, to gauge the model prediction accuracy. For band gap prediction, we do not report MARE because some of the values obtained are very small.

2.3. ML Workflow

We developed regression models based on six different ML algorithms to predict the three vibrational properties and band gap. This allowed us to compare the performance of different nonlinear ML methods. We used a feed-forward back-propagation neural network (NN),^[19] a support vector machine (SVM),^[20] a relevance vector machine (RVM),^[21] a random forest (RF), the XGBOOST algorithm,^[22] and the Huber regression algorithm.^[23] The RVM is a sparse Bayesian variation of the SVM. The fully connected feed-forward NN has three layers: an input layer with linear transfer function, a hidden layer with 15 nodes containing a sigmoid transfer function, and an output layer with one output node containing a linear transfer function. In the RF, we set the maximum number of trees to 400. The SVM uses the parameters C = 1 and $\epsilon = 0.1$, and both the SVM and RVM use a radial basis function. The XGBOOST used a linear objective function. Figure 1 summarizes the workflow for ML model generation. For each ML method, we trained the model on 80% of the data set and used the remaining 20% as the test set.

2.4. Data Availability

The row data, trained machine learning models, as well as a python code for applying the row data to the models to obtain the error estimates are available in the following link: https://github.com/sheriftawfikabbas/machinelearningphonons. More information is available in Supporting Information.



Figure 1. Schematic diagram of the workflow of the machine learning procedure applied in the present study.

3. Results and Discussion

Table 1 shows the results of the ML predictions of the properties of the materials in the test set. The predictivities of the six models for *S* and *C*_V are very high, as shown by values of *R*² close to 1 and small RMSE and MARE values. In particular, the *R*² values for predicting the entropy are all above 0.80, and the RVM achieves the highest accuracy with *R*² = 0.90, RMSE = 0.32 meV per atom per K, and MARE = 13.8%. For *C*_V, all of the models achieve an *R*² above 0.80, and the most accurate models are XGBOOST, NN, and RF with *R*² = 0.88–0.89, RMSE ≤0.20 meV per atom per K, and MARE = 11–12%. However, for ϵ_{eff} , the prediction accuracy is slightly lower: all *R*² values lie in the range 0.58–0.72. The most accurate model is the NN, which achieves *R*² = 0.72, RMSE = 4.41, and MARE = 17.9%.

The performance of the ML models was scored primarily by the RMSE values rather than the R^2 values.^[18] The MARE and RMSE values for the test set in Table 1 show that the ML methods generated models with similar predictive performance. This indicates that the descriptor quality, not the particular ML algorithm, is the most important factor in model performance. **Figure 2** displays the correlations between the DFT-calculated and the RVM-predicted quantities. We exemplified the test set predictions using the RVM models because of the tendency of XGBOOST and SVM to overfit models (very low RMSE and very high R^2 for the training set with XBOOST models). For C_V , the XGBOOST has a very similar RMSE (0.18 meV per atom per K) to that of the RF model (0.19 meV per atom per K). The full list of correlation plots is in Supporting Information.

There are a number of materials whose properties are challenging to predict. In particular, we found 34 structures whose properties could not be accurately predicted by the 6 ML models; that is, their RE (relative error) was >10% for all 6 ML models. (The material IDs of the structures, their chemical formulae, and space groups are displayed in Table S1, Supporting Information). The majority of these materials are oxides, complex phosphides, nitrides, and carbides. Table S1, Supporting Information, shows that 13 of them are oxides, including magnesite, MgCO₃ and its reaction product, MgO, and 24 of them are compounds that constitute alkali and alkali earth metal elements. All 34 materials are semiconductors with band gaps between 0.32-7.83 eV.

ADVANCED SCIENCE NEWS _

www.advancedsciencenews.com

Table 1. The MARE, RMSE, R^2 , for the specific heat (C_V) in units of meV per atom per K, the entropy (S) in units of meV per atom per K, and the trace of the effective polycrystalline dielectric function (ε_{eff}).

ML model	Dataset	C _V			S			$\epsilon_{\rm eff}$		
		MARE [%]	RMSE	R ²	MARE [%]	RMSE	R ²	MARE [%]	RMSE	R ²
RF	Test	10.9	0.19	0.88	18.2	0.39	0.85	17.1	4.85	0.66
	Train	4.6	0.09	0.97	6.5	0.16	0.97	7.1	2.46	0.93
SVM	Test	14.4	0.24	0.81	15.7	0.36	0.87	18.4	4.70	0.68
	Train	8.3	0.11	0.95	9.1	0.17	0.97	9.1	3.40	0.86
RVM	Test	12.4	0.22	0.84	13.8	0.32	0.90	17.6	4.57	0.70
	Train	2.9	0.04	0.99	3.4	0.08	0.99	7.6	1.58	0.97
Huber	Test	13.1	0.23	0.83	13.2	0.34	0.89	17.9	5.35	0.58
	Train	8.1	0.16	0.90	7.5	0.23	0.94	11.6	4.33	0.77
XGBOOST	Test	11.1	0.18	0.89	14.9	0.33	0.89	16.8	4.87	0.65
	Train	0.0	0.00	1.00	0.0	0.00	1.00	0.0	0.01	1.00
NN	Test	12.4	0.20	0.88	14.5	0.31	0.90	17.9	4.41	0.72
	Train	0.8	0.06	0.99	1.1	0.09	0.99	1.2	0.89	0.99

All quantities are obtained at 298 K.



Figure 2. The correlation between properties predicted for the test set by RVM-models and those calculated by DFT. a) C_V (meV per atom per K), b) S (meV per atom per K), c) ε_{eff} (eV), and d) E_g (eV).

3.1. Outliers

For most of the ML models, there are a few materials with very large test set errors for their predicted properties. The RF and SVM models have large errors for simple crystals; RF has a relative error of 88% when predicting the entropy of diamond (mp-66), 53% for elemental boron (mp-160), and MARE in the range 45% for AlN and 79% for GaN. SVM has a very large RE of 530% when predicting the entropy of diamond, 220% for elemental boron, and 103% for solid nitrogen (mp-999498). The poor performance of the widely-used SVM method may be due to its tendency to overfit models so the sparse Bayesian equivalent, the RVM, is usually preferred.^[24] The RVM model prediction for diamond has an RE value of 0.5%, 0.07% for elemental boron, and less than 2% for GaN and AlN, reinforcing this point on SVM overfitting.

The three other ML models had relatively large prediction errors for specific materials. The NN had a 21% MARE error in predicting the entropy of the complex nitride Li_2CN_2 , RVM had an RE of 28% error in predicting the entropy of the carbide Be_2C , and XGBOOST had an RE of 19% for predicting the entropy of the complex carbide Al_4C_3 .

The power of ML methods for predicting vibrational properties is most obvious for large and complex structures. All six models accurately predict the entropy for the three structures, K₂NaAlF₆, NaNbO₃, and Be₃P₂, each of which has 40 atoms in the unit cell (the largest structures in the data set).

The ability of machine learning models to predict material properties using the PLMF has previously been demonstrated by Isayev et al.,^[2a] where the authors used the models to obtain highly accurate predictions of an electronic quantity, the band gap, with $R^2 = 0.9$ (obtained from fivefold cross validated predictions for seven regression models) based on a data set of 26674 materials.

We were interested in whether our data set, although smaller, would yield a similar prediction accuracy. As **Table 2** shows, we obtained the same band gap prediction accuracy as Isayev et al. ($R^2 = 0.89$, RMSE = 0.52), for the RF, XBOOST, and NN models.

We compared our ML results for prediction of vibrational entropy with those of Legrain et al.,^[2b] and for heat capacity with those of Kauwe et al., but the following methodological differences should be noted. Legrain et al. applied a 5- to 14-fold cross validation technique and reported prediction errors based on an average of cross validation sets that are \approx 7–20% of the data set. www.advancedsciencenews.com

Table 2. The RMSE and R^2 for predicting the band gap using the six ML models.

ML model	Data set	R ²	RMSE [eV]
RF	Test	0.89	0.52
	Train	0.98	0.24
SVM	Test	0.85	0.59
	Train	0.91	0.50
RVM	Test	0.86	0.58
	Train	0.98	0.22
Huber	Test	0.85	0.59
	Train	0.91	0.50
XGBOOST	Test	0.89	0.50
	Train	1.00	0.00
NN	Test	0.89	0.52
	Train	1.00	0.04

Cross validation tends to give an overly optimistic estimates of the prediction accuracies than test sets because the validation data spend much of the time in the training set so are not really independent. Here, we employ a test that is 20% of the data set, a more stringent and pessimistic measure of the ability of models to generalize to new data. Our vibrational entropy also covered a range that was ten times larger than that modeled by Legrain et al. and our data sets' size and diversity was much higher. Legrain et al.'s reported that the average cross-validated RMSE and R^2 values were 0.052 meV per atom per K and 0.90, respectively, compared to 0.31 meV per atom per K and 0.90 for our study using a more stringent test of prediction accuracy and for a much larger range of entropy values.

A comparison of our ML results for $C_{\rm V}$ prediction to those of Kauwe et al.^[13] showed that our models had lower prediction errors. These authors also used a test set that is 20% of the dataset and the data set sizes were similar, but they quoted MARE values for the entire data set. They reported MARE values within the range of 8–14%. To make a valid comparison, we also calculated the error values for prediction of the entire data set (Table S2, Supporting Information). Our MARE values lie in the range 0.5–2%, much lower than that reported by Kauwe et al.

4. Conclusion

We addressed the challenging problem of calculating vibrational properties for a very large number of crystalline materials by using ML to leverage a much smaller number of computationally expensive DFT results used to train them. Descriptors based on the property-labeled materials fragments approach could successfully model our initial data set of 1521 structures. The models produced a root mean square error for prediction of the entropy of 0.31 meV per atom per K, for specific heat of 0.18 meV per atom per K, and for the trace of the dielectric tensor of 4.41.

The present work offers a massive saving in computational resources compared to calculating these properties using DFT methods alone, while providing similar accuracy. With such a reasonable prediction accuracy, our approach can accelerate the pace of material discovery by identifying new materials with interesting vibrational properties for superconducting, ferroelectric, and upconverting materials, for example.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

crystal properties, density-functional theory, dielectric constant, entropy, machine learning

Received: October 18, 2019 Revised: November 7, 2019 Published online:

- S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, Nat. Mater. 2013, 12, 191.
- [2] a) O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, *Nat. Commun.* 2017, *8*, 15679; b) F. Legrain, J. Carrete, A. van Roekeghem, S. Curtarolo, N. Mingo, *Chem. Mater.* 2017, *29*, 6220; c) T. Xie, J. C. Grossman, *Phys. Rev. Lett.* 2018, *120*, 145301.
- [3] J. Zhou, B. Huang, Z. Yan, J.-C. G. Bünzli, Light: Sci. Appl. 2019, 8.
- [4] a) A. L. Hook, M. R. Alexander, D. A. Winkler, in *Tissue Engineering* (Eds: C. A. V. Blitterswijk, J. D. Boer), Academic Press, Oxford **2014**, pp. 253–281; b) P. Mikulskis, M. R. Alexander, D. A. Winkler, *Adv. Intell. Syst.* **2019**, https://doi.org/10.1002/aisy.201900045; c) V. C. Epa, J. Yang, Y. Mei, A. L. Hook, R. Langer, D. G. Anderson, M. C. Davies, M. R. Alexander, D. A. Winkler, *J. Mater. Chem.* **2012**, *22*, 20902.
- [5] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, *Chem. Mater.* 2015, *27*, 735.
- [6] a) L. M. R. Daniel, P. Tabor, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, A. Aspuru-Guzik, *Nat. Rev. Mater.* 2018, *3*, 5; b) A. W. Thornton, C. M. Simon, J. Kim, O. Kwon, K. S. Deeg, K. Konstas, S. J. Pas, M. R. Hill, D. A. Winkler, M. Haranczyk, B. Smit, *Chem. Mater.* 2017, *29*, 2844.
- [7] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Nat. Commun. 2016, 7, 11241.
- [8] Materials Genome Initiative, https://www.nist.gov/system/files/ documents/2017/05/09/MGI-StrategicPlan-2014.pdf (accessed: December 2019).
- [9] S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, *Comput. Mater. Sci.* 2012, 58, 218.
- [10] 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.* 2013, 1, 011002.
- [11] G. Bergerhoff, R. Hundt, R. Sievers, I. D. Brown, J. Chem. Inf. Model. 1983, 23, 66.
- [12] J. Leitner, P. Voňka, D. Sedmidubský, P. Svoboda, Thermochim. Acta 2010, 497, 7.
- [13] S. K. Kauwe, J. Graser, A. Vazquez, T. D. Sparks, Integr. Mater. Manuf. Innovation 2018, 7, 43.

ADVANCED SCIENCE NEWS

www.advancedsciencenews.com

- [14] G. Petretto, S. Dwaraknath, H. P. C. Miranda, D. Winston, M. Giantomassi, M. J. van Setten, X. Gonze, K. A. Persson, G. Hautier, G. M. Rignanese, *Sci. Data* **2018**, *5*, 180065.
- [15] I. Petousis, W. Chen, G. Hautier, T. Graf, T. D. Schladt, K. A. Persson, F. B. Prinz, Phys. Rev. B 2016, 93.
- [16] A. Grisafi, D. M. Wilkins, G. Csanyi, M. Ceriotti, Phys. Rev. Lett. 2018, 120, 036002.
- [17] R. Tibshirani, J. R. Stat. Soc.: Ser. B 1996, 58, 267.
- [18] D. L. Alexander, A. Tropsha, D. A. Winkler, J. Chem. Inf. Model. 2015, 55, 1316.
- [19] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Berlin **2006**.
- [20] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York **1995**.
- [21] M. E. Tipping, J. Mach. Learn. Res. 2001, 1, 211.
- [22] T. Chen, C. Guestrin, in Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, ACM, New York 2016, pp. 785– 794.
- [23] P. Huber, Robust Statistics, Wiley, New York 1981.
- [24] F. R. Burden, D. A. Winkler, J. Chem. Inf. Model. 2015, 55, 1529.