MRS Advances © 2018 Materials Research Society
DOI: 10.1557/adv.2018.260

Efficient Discovery of Optimal *N*-Layered TMDC Hetero-Structures

Lindsay Bassman[1,2*], Pankaj Rajak[1,4*], Rajiv K. Kalia[1,2,3,4], Aiichiro Nakano[1,2,3,4,5], Fei Sha[3,5], Muratahan Aykol[6], Patrick Huck[6], Kristin Persson[6], Jifeng Sun[7], David J. Singh[7], Priya Vashishta[1,2,3,4]

*[1]Collaboratory for Advanced Computing and Simulations, [2]Department of Physics, [3]Department of Computer Science, [4]Department of Chemical Engineering and Material Science, [5]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089,USA; [6]Lawrence Berkeley National Lab, 1 Cyclotron Rd, Berkeley, CA 94720, USA; [7]Department of Physics and Astronomy, University of Missouri, Columbia, MO 65211, USA, *These authors contributed equally to this work*

ABSTRACT

Vertical hetero-structures made from stacked monolayers of transition metal dichalcogenides (TMDC) are promising candidates for next-generation optoelectronic and thermoelectric devices. Identification of optimal layered materials for these applications requires the calculation of several physical properties, including electronic band structure and thermal transport coefficients. However, exhaustive screening of the material structure space using ab initio calculations is currently outside the bounds of existing computational resources. Furthermore, the functional form of how the physical properties relate to the structure is unknown, making gradient-based optimization unsuitable. Here, we present a model based on the Bayesian optimization technique to optimize layered TMDC hetero-structures, performing a minimal number of structure calculations. We use the electronic band gap and thermoelectric figure of merit as representative physical properties for optimization. The electronic band structure calculations were performed within the Materials Project framework, while thermoelectric properties were computed with BoltzTraP. With high probability, the Bayesian optimization process is able to discover the optimal hetero-structure after evaluation of only ~20% of all possible 3-layered structures. In addition, we have used a Gaussian regression model to predict not only the band gap but also the valence band maximum and conduction band minimum energies as a function of the momentum.

INTRODUCTION

Multi-layered hetero-structures made from vertically stacked, two-dimensional, transition metal dichalcogenides (TMDC) have unique optoelectronic and thermoelectric characteristics, due to the relatively weak interlayer interactions and the concomitant two-dimensional confinement [1-3]. Unlike large band gap materials like hBN (insulator) [4] and zero-band gap graphene (semi-metal) [5], these hetero-structures have band gaps comparable to conventional semi-conductors like Si and GaAs. Furthermore, the band gap and other properties (*e.g.* thermal conductivity) of these materials can be tuned to desired values by changing the composition of each layer as well as the total number of layers. This makes these multi-layered materials promising candidates for next-generation electronic devices (*e.g.* field effect transistors), where band gap and thermal properties can be used as a screening parameter for specific applications [6, 7].

As the number of layers, N, of the hetero-structure increases, both the combinatorial number of layer configurations and the computational time required for *ab initio* calculations of each hetero-structure's material properties increases as $O(n^3)$, where n is the number of atoms. Hence, performing exhaustive density functional theory (DFT) calculations (which are the root of the $O(n^3)$ computational complexity) becomes infeasible for large N. Recently, machine-learning methods have shown phenomenal success in the material science domain for high-throughput screening and property prediction [8, 9]. For example, statistical models built using a small fraction of a family of structures can be used to accurately predict a wide range of material properties like band gap [10, 11], dielectric breakdown strength [12, 13] and melting point [11]. Machine-learning methods like support vector regression [10, 11], neural network [14] and kernel ridge regression [15] are shown to accurately predict the band gap in double perovskites [16] and in polymers [17]. For many applications, however, we only need to find materials with physical properties beyond a certain threshold value. Building a regression model and using that model to predict material properties of each structure until the desired structure is found, is not an efficient process as it requires a substantial amount of expensive computation to build the model. A promising machine-learning technique to solve this problem, known as Bayesian optimization [18-20], optimizes a black box function with minimal function evaluations.

In this work, we have developed a model based on Bayesian optimization to find the TMDC hetero-structure with a desired property by performing a minimal number of structure evaluations for 3-layer hetero-structures. In our case we use the maximum band gap and thermoelectric figure of merit as examples of such desired properties. We also developed a regression model using Gaussian processes [16, 21, 22] to predict the band gap and band structure for 3-layer hetero-structures.

METHOD

Material Property Calculations

Each layer in a N-layer hetero-structure consists of two types of atoms: A and B. A can be either molybdenum (Mo) or tungsten (W), while B can be either sulfur (S), selenium (Se) or tellurium (Te). Thus, there exist 6^N possible configurations for N-layer hetero-structures. There are, then, 216 3-layer hetero-structures, for all of which we computed material properties. Structure files for all 3-layer hetero-structures were generated automatically and uploaded to the Materials Project (MP) database [23] using the pymatgen [24] library. The electronic band structures are calculated using density functional theory (DFT) with the projector augmented wave [25] method implemented in the Vienna Ab Initio Simulation Package (VASP) [26, 27]. The exchange and correlation energies are approximated with the Perdew-Burke-Ernzerhof version of the generalized gradient approximation [28]. All 3-layered hetero-structures are represented by a 9-atom unit cell, with periodic boundary conditions in all directions. The 9 atoms consist of three sets of 3 atoms, each set representing one layer. First, both the unit cell and the atoms within are allowed to relax. Next, a self-consistent field iteration is performed to obtain the electron wave functions. Finally, the electronic band structure is calculated from the resulting wave functions. Once the electronic structures are computed within the MP database, the data is downloaded, again using pymatgen, and BoltzTraP [29] is run on the results to compute thermoelectric properties of the hetero-structures.

Feature Vector

Two important atomic properties that determine both electronic band structure and thermoelectric properties of a material are the electronegativity and first ionization potential. Thus, to uniquely represent each structure, every atom type in each layer is represented by its electronegativity and first ionization potential. Since each layer consists of two elements, a 3-layered structure will be represented by a 12-dimensional vector, where the first four components of the vector are the electronegativity and first ionization potential of the atoms species in the first layer, the next four components are from the second layer, and remaining four are from the third layer. To build a Gaussian process regression model for band gap and band structure prediction, a squared exponential kernel is used: $K(X, X') = \exp[-(\sum_1^{12} \|x_i - x_i'\|^2)/\sigma_i^2]$. Here, X and X' each run over all 3-layered hetero-structures and are 12-dimensional vectors with components x_i and x'_i, respectively. σ_i are tunable hyper-parameters, associated with each component of the feacture vector, estimated from the training data using maximum likelihood estimate.

RESULTS

Band Gap Prediction

We built a Gaussian regression model to predict the band gap in 3-layered hetero-structures. To determine the appropriate training data set size, regression models are built using different percentages of the total data set (216 structures) as the training data set, ranging from 40% to 70%. Figures 1(a) and 1(b) show the band gap prediction for the test data set (all remaining structures not in the training data set) for two models built using 40% and 60% of the data in the training data set, respectively.

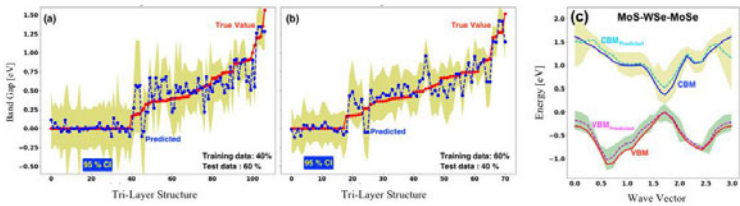

Figure 1: Prediced band gap on test data for two models built using 40% (a) and 60% (b) training data sets. Here, the red, solid curve is the true value of the band gap (computed from DFT calculations), while the blue, dashed curve is the predicted value of the band gap (predicted using the Gaussian regression model). The yellow regions specify a 95% CI of the predicted value. (c) Predicted CBM (dashed, cyan) and VBM (dashed, meganta) along with their true values in solid blue (CBM) and red (VBM) of one model tri-layer structure in the test data set. Here, the yellow and green shaded areas represent 95% CI of the predicted value and the wave vector is given in arbitrary units running along a line of high symmetry.

The model built using 40% has a larger confidence interval (CI) and the predicted band gap for many structures lies outside the 95% CI. In contrast, the model made from 60% is more robust, with smaller CI. To further test the robustness of the model, all 3-layer structures are randomly split into training and test data sets, with either 40% or 60% training data, 100 times. After building the model, the mean square error (MSE) for each model is used to calculate the prediction accuracy. The average prediction error of these 100 runs for models built using 40% and 60% training data set is 0.167 and 0.144, respectively. While the MSE only slightly

decreases, the width of the confidence interval (CI) becomes much smaller with the larger training data set. We also observe that increasing the training data set beyond 60% shows little improvement in results. Hence, a training data set comprised of 60% of the total number of structures is sufficient to build a model both with low MSE and CI.

Band Structure Prediction

A Gaussian regression model is also built to predict the shape of the valence-band maximum (VBM) and conduction-band minimum (CBM) as a function of the momentum along lines of high symmetry in the first Brillouin zone. Unlike the previous case where the target variable Y (band gap) is a scalar, here the target variables are two vectors, one for the VBM ($\overrightarrow{Y_1}$) and the other for the CBM ($\overrightarrow{Y_2}$). These vectors consist of 30 discretized points, where each point corresponds to a particular wave vector and the component of $\overrightarrow{Y_1}(\overrightarrow{Y_2})$ is the energy value of VBM(CBM) at that point. A regression model is build using ~60% of the data in the training data set at each of these 30 points and the predicted output from these 30 models are used to construct the shape of VBM and CBM. Figure 1(c) shows the predicted VBM and CBM for one of the structures in test data.

Bayesian Optimization

For many applications, we only need to find the structure with the maximum band gap or thermo-electric figure of merit, T_2. Since each calculation is time consuming, Bayesian optimization can be used for efficient discovery of the material with desired properties (*i.e.* with minimal structure calculations). While we attempt to find the maximum values of each property, Bayesian optimization is also capable of finding the structure with a property closest to a desired value. In the Bayesian optimization process, first, a Gaussian process regression for the band gap or T_2 value is built by randomly selecting 10 structures from all possible 3-layered structures. Then, the next structure to be computed is chosen based on the trade-off between exploration (to diversify the search) and exploitation (to follow the trend found by the current estimates). Since the true functional form of the objective function is unknown, the procedure optimizes a surrogate function called acquisition function [18, 20]. Among the available acquisition functions, such as probability of improvement, upper confidence bounds, and expected improvement, we used expected improvement (EI). The value of EI for the structures which are not in the training data is calculated, and the structure with the maximum EI is used as a guess for the optimal structure with respect to the desired property. This completes one iteration of Bayesian optimization, and we perform a total of 30 iterations.

In 3-layer structures with a total of 216 structures, only 5 structures have band gap above 1.35 eV, and the remaining structures have bandgap below 1.30eV. The band-gap values of these 5 structures are shown in Table 1.

Structure	Band Gap [eV]	Frequency (%)
WS_2-WS_2-WS_2	1.56	46.4
WSe_2-WSe_2-WSe_2	1.51	16.2
$MoSe_2$-$MoSe_2$-$MoSe_2$	1.41	25.6
MoS_2-MoS_2-MoS_2	1.39	10.4
MoS_2-MoS_2-WS_2	1.35	1.2
	Total	99.8

Table 1: (Column 2) Band gap of the top five structures in 3-layered hetero-structure. (Column 3) Frequency of each of these structures in 500 runs of Bayesian optimization, where each Bayesian optimization run consists of 30 iterations.

The table also shows the frequency of each of these structure in 500 runs of Bayesian optimization, where in each run 10 initial structures are randomly choosen and the model is ran for a total of 30 iterations. It can be seen from Table 1 that our model is able to determine one

of the best five structures 99.8% of the runs within 30 iterations. Figure 2 shows the histogram of the number of runs the model takes to predict the optimal structure. Here, 0 steps corresponds to the cases where the initial 10 data points contained one of the five best structures. In the remaining cases, we see that at most 30 iterations are required to determine one of the five best structures.

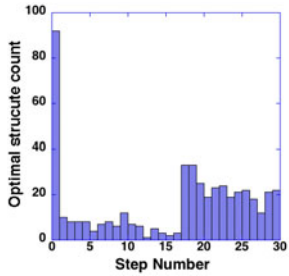

Figure 2: Histogram of the total number of iteration steps required to determine the optimal structure.

We have also performed Bayesian optimization of the thermoelectric figure of merit, and found that the top three structures are WTe_2-$MoTe_2$-WTe_2, $MoSe_2$-WSe_2-WSe_2 and WSe_2-$MoSe_2$-WSe_2. Again Bayesian optimization is able to predict one of these three structures within 30 iterations, taking only 5 structures as initial data points.

CONCLUSIONS

We have demonstrated that Bayesian optimization significantly reduces the computation time required to search for the material with a desired physical property, using the band gap and thermo-eclectic figure of merit as examples. With only few initial sample structures as a training data set, we were able to find the 3-layer hetero-structures with maximum band gap (WS_2-WS_2-WS_2) and maximum thermo-electric figure of merit (WTe_2-$MoTe_2$-WTe_2) within 30 structure evaluations.

ACKNOWLEDGMENTS

This work was supported as part of the Computational Materials Sciences Program funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Award Number DE-SC00014607. All simulations were performed at the Center for High Performance Computing of the University of Southern California.

References

1. A. Gupta, T. Sakthivel and S. Seal, Prog. Mater. Sci. **73**, 44-126 (2015).
2. Y. Venkata Subbaiah, K. Saji and A. Tiwari, Adv. Funct. Mater. **26** (13), 2046-2069 (2016).
3. Y. Zhang, Y.-W. Tan, H. L. Stormer and P. Kim, Nature **438** (7065), 201-204 (2005).
4. F. Deepak, C. Vinod, K. Mukhopadhyay, A. Govindaraj and C. Rao, Chem. Phys. Lett. **353** (5), 345-352 (2002).
5. P. R. Wallace, Phys. Rev. **71** (9), 622 (1947).
6. A. Jain, Y. Shin and K. A. Persson, Nat. Rev. Mater. **1**, 15004 (2016).

7. 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** (12), 4849-4861 (2011).
8. K. Rajan, Mater. Today **8** (10), 38-45 (2005).
9. R. LeSar, Statistical Analysis and Data Mining: The ASA Data Science Journal **1** (6), 372-374 (2009).
10. J. Lee, A. Seko, K. Shitara, K. Nakayama and I. Tanaka, Phys. Rev. B **93** (11), 115104 (2016).
11. T. Gu, W. Lu, X. Bao and N. Chen, Solid State Sci. **8** (2), 129-136 (2006).
12. C. Kim, G. Pilania and R. Ramprasad, J. Phys. Chem. C **120** (27), 14575-14580 (2016).
13. C. Kim, G. Pilania and R. Ramprasad, Chem. Mater. **28** (5), 1304-1311 (2016).
14. Z. Zhaochun, P. Ruiwu and C. Nianyi, Mater. Sci. Eng. B **54** (3), 149-152 (1998).
15. T. D. Huan, A. Mannodi-Kanakkithodi and R. Ramprasad, Phys. Rev. B **92** (1), 014106 (2015).
16. A. I. Forrester, A. Sóbester and A. J. Keane, presented at the Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 2007 (unpublished).
17. A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman and R. Ramprasad, Sci. Rep. **6**, 20952 (2016).
18. E. Brochu, V. M. Cora and N. De Freitas, arXiv preprint arXiv:1012.2599 (2010).
19. B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. de Freitas, Proc. IEEE **104** (1), 148-175 (2016).
20. J. Snoek, H. Larochelle and R. P. Adams, presented at the Advances in Neural Information Processing Systems, 2012 (unpublished).
21. C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. (MIT press Cambridge, 2006).
22. M. C. Kennedy and A. O'Hagan, Biometrika **87** (1), 1-13 (2000).
23. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner and G. Ceder, APL Mater. **1** (1), 011002 (2013).
24. S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, Comput. Mater. Sci. **68**, 314-319 (2013).
25. P. E. Blöchl, Phys. Rev. B **50** (24), 17953 (1994).
26. G. Kresse and J. Furthmüller, Phys. Rev. B **54** (16), 11169 (1996).
27. G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6** (1), 15-50 (1996).
28. J. P. Perdew, K. Burke and M. Ernzerhof, Phys. Rev. Lett. **77** (18), 3865 (1996).
29. G. K. Madsen and D. J. Singh, Comput. Phys. Commun. **175** (1), 67-71 (2006).