Predicting defect formation energies in semiconductors using machine learning

Savyasanchi Aggarwal[a], **Martin H. Petersen**[b], **Seán R. Kavanagh**[c], **Alex G. Squires**[d], **Kedar Hippalgaonkar**[a], **David O. Scanlon**[d]

[a] *Institute of Materials Research and Engineering, Agency for Science, Technology and Research* aggarwalsavya@gmail.com, kedar@ntu.edu.sg

[b] *Technical University of Denmark, Department of Energy Conversion and Storage, Lyngby, 2800, Denmark* mahpe@dtu.com

[c] *Harvard University Center for the Environment, Cambridge, Massachusetts 02138, United States* skavanagh@seas.harvard.edu

[d] *School of Chemistry, University of Birmingham, Edgbaston, Birmingham, B15 2TT, U.K.* a.squires@bham.ac.uk, d.o.scanlon@bham.ac.uk

* Presenting author

1. Introduction

Point defects are localised imperfections that break the symmetry of a crystalline material, sometimes leading to drastic changes in physical and structural properties. Although there has been considerable effort to characterize and manipulate defects for commercial applications, the huge computational cost of defect modelling has been a considerable bottleneck. This study explores the use of machine learning methods as a low-cost alternative, focusing on predicting formation energies of point defects using a property-driven approach. In doing so, we introduce a novel, high-fidelity database built from results calculated purely with hybrid-level density functional theory (DFT) and extensive ground-state structural validation.

2. Database

Thus far, DFT has been the primary method to obtain energetic and structural information about point defects, and offers an excellent point of comparison to experimental results. However, a high level of theory is necessary to accurately compute the electronic structure. This increases the computational cost exponentially, limiting analysis for complex systems [1]. Furthermore, recent results have shown that using DFT with standard single-relaxation gradient-descent optimisation is unreliable in sampling complex energy landscapes, necessitating advanced structural searching methods that further drive up the computational cost [2, 3]. These constraints render DFT impractical in viable high-throughput workflows, an essential criterion for large-scale screening of the vast chemical space of dopants.

Instead, the use of machine learning (ML) in a property-driven approach can bypass the often complex stages of structural relaxations, mapping multi-dimensional relationships between the input site of the defect in a host material and its formation energy. This will cut the computational expense significantly and gives insights into the physical features governing defect formation. However, to build a ML model that is sufficiently accurate and universal, a highly diverse and accurate database is first required. Thus, this study involves the construction of a point defect database consisting of over 1200+ intrinsic and extrinsic defect configurations ranging from binary to quarternary host systems, as well as a wide range of different charge states. All datapoints were calculated with hybrid-DFT and targeted bond distortion analysis using the doped [4] and ShakeNBreak [5] codes, and compared with experimental results where possible. Some statistics on this dataset are given in Table 1, and the diversity of the database is shown in Figure 1. This database represents an excellent and reliable starting point to understand the relationships governing defect formation.

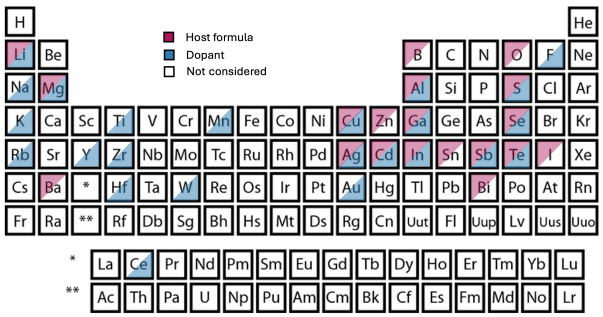

Fig. 1: Elemental diversity in the database. Red represents elements present in host materials, whereas blue represents its presence as a dopant.

3. Feature space

The main target for the model is $E_{combined}$, a sum of all terms associated with only the structural relaxation of the defect supercell. This is the sum of difference in energy between the pristine and defect supercell ($E_{X,q} - E_H$) and the charge correction term

Table 1: Unique values for features in the dataset. *Unique defects* refers to individual charge states for each defect. Some defects have the same formation energies to other, unrelated defects, hence the smaller count of f.e. than datapoints.

Feature	Unique values
Host formula	23
Charge State	13
Unique defects	1231
Defect Formation Energy	944
Space Group	14
Bandgap (eV)	23
Supercell Size	10

$E_{corr}(q)$. The other terms normally calculated for the defect formation energy, as shown in Figure 2, can either be inferred from the information of the bulk material (qE_F), or rely on other adjacent calculations not entirely related to the structural relaxation ($\Sigma n_i \mu_i$).

Excluding these terms does increase the y-range ($E_{combined}$) significantly, affecting model accuracy, but limits information leakage between the training and test sets and improves bias. As for the input feature space, features were defined to encode as much atomic, electronic and structural information about the defect atom and bulk outgoing/incoming sites. Explicit labeling of the atomic numbers or names was avoided to avoid biasing the model towards any particular elements, instead encouraging learning of physical trends. All input features were extracted prior to the structural relaxation of the defect supercell with methods based on the DOPED [4], pymatgen [6, 7] and pydefect [8] open-source packages. A few different structural representations were also tested, with best performance coming from the SOAP [9] representation (n = 2, l = 2).

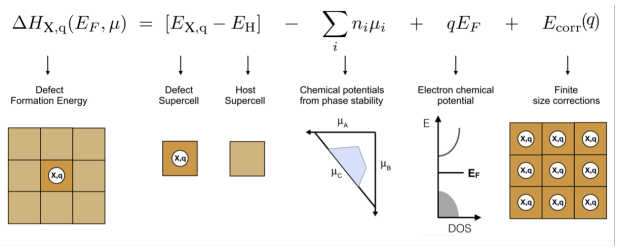

Fig. 2: DFT-calculated terms that constitute the defect formation energies for point defects. Training of the model was performed only on the $E_{X,q} - E_H$ and E_{corr} terms. Adapted with permission from Goyal et al. [10]

4. Performance

Upon considering prediction universality of the model, the training, test and validation splits of the dataset are fundamentally important. This study focused on maximising performance for four main types of $E_{combined}$ prediction splits considering this; (1) new charge states for a previously trained defect, (2) new defects of a previously studied host material, (3) previously unseen extrinsic defects (dopant), (4) previously unseen host materials. Obtaining robust accuracy for any of these alleviates significant computational expense.

Figure 3 shows the performance of XGBoost [11] model in prediction type (1), where the test set contains charge states of defects that are excluded from the training set. This is performed without any encoding of structural information, only containing 'metadata' of the bulk and defect atoms and their atomic features. This model achieves an average MAE of 0.306 eV for new charge states. Considering a range of ~22 eV of (lowest) defect formation energies $E_{formation}$, and a lack of structural information, this is a promising result. As expected, performance drops upon increasing blindness of the test/validation sets from the training sets (going to prediction types 2,3, and 4), and is increasingly dependent on dataset splits. However, the standard error observed for these (without explicit structural encoding) is still relatively low compared to the typical range of $E_{formation}$, and there is much potential for use as a screening study if not to replace the calculations outright.

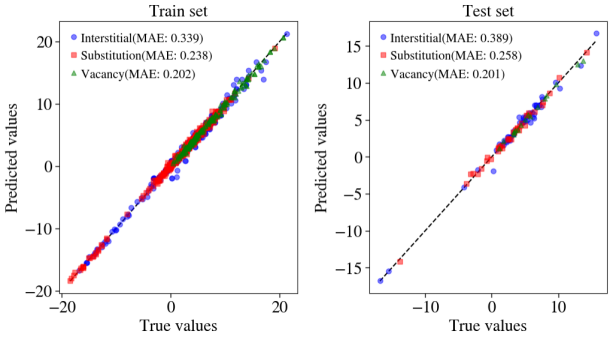

Fig. 3: Parity plot of current model performance using only "metadata". The energy range shown is that of $E_{combined}$, not $E_{formation}$.

Naturally, the model's performance will grow better with (ongoing) additions to the input dataset, as well as a more physically meaningful feature space.

4.1 Related work

While there have been a number of previous studies aiming investigating this problem, most are limited by the reliability of energies in their input datasets. To create large datasets, these studies forgo the necessary level of theory in their calculations, minimising computational expense. With the high-fidelity data implemented for this study, this issue is not encountered. Further, our methodology extends upon the excellent efforts of studies by Witman et al. [12] and Kumagai et al. [8] to ensure an accurate investigation.

Acknowledgments

S.A. gratefully acknowledges Ke Li, Seán Kavanagh, Sabrine Hachmioune, Cibrán López-Álvarez and Joe Willis for their contributions to the dataset, and Andy Chen for the helpful discussions. This project was supported by the UCL-A*STAR Collaborative Programme via the Centre for Doctoral Training in Molecular Modelling and Materials Science (M3S CDT) at UCL. M.H.P acknowledges support from the Det Frie Forskningsråd under Project "Data-driven quest for TWh scalable Na-ion battery (TeraBatt)" (Ref. Number 2035-00232B)., S.R.K thanks the Harvard University Center for the Environment (HUCE) for funding a fellowship. Through our membership of the UK's HEC Materials Chemistry Consortium, which is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) (Nos. EP/L000202, EP/R029431, EP/T022213), this work also used ARCHER2 UK National Supercomputing Services. We are also grateful to the UK Materials and Molecular Modelling Hub for computational resources, which is partially funded by EPSRC (Nos. EP/T022213/1, EP/W032260/1, and EP/P020194/1).

References

[1] Sunghyun Kim, Samantha N Hood, Ji-Sang Park, Lucy D Whalley, and Aron Walsh. Quick-start guide for first-principles modelling of point defects in crystalline materials. *Journal of Physics: Energy*, 2(3):036001, July 2020. Publisher: IOP Publishing.

[2] Irea Mosquera-Lois and Seán R. Kavanagh. In search of hidden defects. *Matter*, 4(8):2602–2605, August 2021. Publisher: Elsevier.

[3] Irea Mosquera-Lois, Seán R. Kavanagh, Aron Walsh, and David O. Scanlon. Identifying the ground state structures of point defects in solids. *npj Computational Materials*, 9(1):1–11, February 2023. Publisher: Nature Publishing Group.

[4] Seán R. Kavanagh, Alexander G. Squires, Adair Nicolson, Irea Mosquera-Lois, Alex M. Ganose, Bonan Zhu, Katarina Brlec, Aron Walsh, and David O. Scanlon. doped: Python toolkit for robust and repeatable charged defect supercell calculations. *Journal of Open Source Software*, 9(96):6433, April 2024.

[5] Irea Mosquera-Lois, Seán R. Kavanagh, Aron Walsh, and David O. Scanlon. ShakeNBreak: Navigating the defect configurational landscape. *Journal of Open Source Software*, 7(80):4817, December 2022.

[6] Jimmy-Xuan Shen and Joel Varley. pymatgen-analysis-defects: A python package for analyzing point defects in crystalline materials. *Journal of Open Source Software*, 9(93):5941, 2024.

[7] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

[8] Yu Kumagai, Naoki Tsunoda, Akira Takahashi, and Fumiyasu Oba. Insights into oxygen vacancies from high-throughput first-principles calculations. *Physical Review Materials*, 5(12):123803, December 2021. Publisher: American Physical Society.

[9] Lauri Himanen, Marc O. J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. DScribe: Library of Descriptors for Machine Learning in Materials Science. *Computer Physics Communications*, 247:106949, February 2020. arXiv:1904.08875 [cond-mat].

[10] Anuj Goyal, Prashun Gorai, Haowei Peng, Stephan Lany, and Vladan Stevanović. A computational framework for automation of point defect calculations. *Computational Materials Science*, 130:1–9, 2017.

[11] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[12] Matthew D Witman, Anuj Goyal, Tadashi Ogitsu, Anthony H McDaniel, and Stephan Lany. Defect graph neural networks for materials discovery in high-temperature clean-energy applications. *Nature Computational Science*, 3(8):675–686, 2023.