Data-driven assessment of thermodynamic stability and search for competing phases: Application to the 2D Material Defect dataset

Roman A. Eremin[a], **Aliaksei V. Krautsou**[a], **Innokentiy S. Humonen**[a], **Artem D. Dembitskiy**[a], **Semen A. Budennyy**[a,b]

[a] *AIRI, Moscow, Russia* eremin@airi.net

[b] *Sber AI, Moscow, Russia*

1. Introduction

In theoretical materials science, combining traditional modeling methods like density functional theory (DFT) with data-driven approaches has become one of the main trends of the last decade. Machine learning models are now widely used for structure-to-property predictions, modeling interatomic interactions, and generating new molecules and crystal structures. Among these, graph neural networks (GNNs) stand out as a particularly popular choice in many such applications [1]. The possibility of DFT/GNN combinations is caused by data accumulation led to creation a number of general purpose (the *Materials Project*, *AflowLib*, *etc.*) databases. For specific problems, e.g., modeling catalytic processes [2], doping effects on phase stability [3], *etc.*, there are often no ready-made data collections, and it is necessary to create new ones to build custom data-driven solutions. Therefore, the development of DFT/GNN approaches and their implementation in a data-efficient manner are of particular interest.

2. Substantial section

2.1 Related works and data

Structural defects and chemical disorder can influence or even determine the properties of many functional materials that are important for their practical applications. From a modeling perspective, the main challenge is computational prediction of the properties, overcoming high combinatorial complexity in disordered structures [4]. Therefore, the need to develop new datasets and heuristics for reasonable and smart selection of training data and search spaces, is beyond doubt. One of the recent examples of such works is the 2D Material Defect (2DMD) dataset [5]. This collection includes DFT properties of six base monolayers – MoS_2, WSe_2, hBN, GaSe, InSe, and black phosphorus – with vacancies and substitutions.

When modeling chemical modifications, competing phases of the same or different structural types are important for assessing thermodynamic stability within the standard convex hull approach [6]. Such phases, despite having a chemical composition different from the target compound, can influence the reference (convex hull) energies throughout the search space, and therefore their determination may require a separate study. In this study, we examine the 2DMD dataset from this perspective.

2.2 Additional data

For easy access and re-evaluation of thermodynamic properties of 2DMD, the previously developed Python tool – *2DMD at a Glance* [7] – is used. Among the 2DMD base monolayers, the subsets with the MoS_2 – WSe_2 and GaSe – InSe compositions (hereafter referred to as MeX_2 and MeX sets, respectively) can be combined because of the same structural types. The corresponding 2DMD entries represent binary, ternary, and quaternary compounds that can be shown in a 3D simplex – a multidimensional polytope suitable for mapping chemical compositions – as depicted in Fig. 1 for MeX_2. Due to the minor modification of the chemical composition at both the low (LDC) and high defect contents (HDC) introduced in the original work [5], the 2DMD entries combined in the above manner represent narrow ranges of compositions in the simplex.

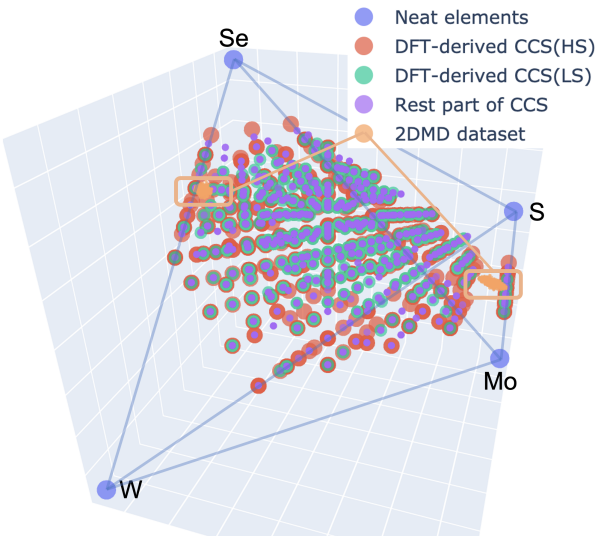

Fig. 1: 3D simplex of quaternary Mo–W–S–Se systems considered in this study and loaded from the 2DMD dataset [5]

In this study, we address thermodynamic properties within the inner part of the simplex (see Fig. 1), i.e. potential competing phases for the 2DMD structure with the same structural type, but in wider chemical compositions. Since in practice this means accounting for up to 100% of the defect content, we create a set of competing phases and develop a hybrid DFT/GNN approach to efficiently estimate its thermodynamic properties.

To keep the combinatorial complexity reasonable for the entire screening space, 2x2x1 supercells are set for the chosen monolayers. The only limitation in generating complete composition/configuration spaces (CCSs) is the absence of more than half of the vacant positions in the both metallic and non-metallic sublattices. Thus, we exclude from further consideration entries for which DFT relaxation may lead to significantly different structural types and chemical compositions compared to the selected base monolayers. Despite the relatively small supercells and the above-mentioned vacancy content limitation, all possible chemical modifications of MeX_2 amounted to a total of 11159 structures, if only symmetrically nonequivalent ones are considered. For the MeX set, this results in 797500 nonequivalent structures.

The former introduced set is convenient for developing and testing the approach to predicting the thermodynamic properties of competing phases. The MeX set, in turn, is of interest for further application of the developed approach in data-efficient manner, due to the obvious impossibility of carrying out DFT evaluation of its entries completely. The assessment of thermodynamic properties in the developed hybrid approach is based on a combination of DFT modeling and the use of structure-to-property predictions (the Allegro [8] and NequIP [9] GNNs). Universal interatomic potentials are also considered as an independent branch of this research and demonstrate similar quality of predictions.

3. Results and discussion

Using the high-symmetry (HS) samples of the introduced CCSs as training sets, we obtain the RMSE scores collected in Table 1 for the MeX_2 low-symmetry (LS) hold out test structures. The same architectures trained on the augmented 2DMD targets – formation energies with respect to the neat constituents – are considered additionally.

Table 1: Test scores within the developed DFT/GNN approach applied to the MeX_2 set part comprising structures with 10+ atoms.

Training dataset (structures)	Pre-training	Allegro RMSE, eV/atom	NequIP RMSE eV/atom
2DMD	–	**0.064**	0.067
(ca. 13K)	Aflow	0.084	0.115
2DMD (LDC)	–	0.069	0.093
(ca. 12K)	Aflow	0.097	0.201
2DMD (HDC)	–	0.069	**0.055**
(1000)	Aflow	0.088	0.158
This work	–	0.075	0.074
(492)	Aflow	0.079	0.073

The lowest errors correspond to the models trained on the 2DMD subsets – 1000 HDC and 12K+ LDC structures – clearly demonstrating its practical applicability. The developed approach shows comparable results, but requires at least two times less data, providing opportunities for subsequent improvements of the quality of predictions. Indeed, a comparison of the obtained scores and the target variations demonstrates the need for such an improvement regardless of the scheme used. First, it may allow a rapid re-evaluation of thermodynamic properties in terms of defect stability and drawing conclusions about the magnitude of the influence of competing phases. Secondly, the obtained most stable structures and energetically favorable arrangements of defects can be considered as an initial guess for further in-depth search for stable structures of defects with remarkable properties. The defect formation energies from the 2DMD data set and the energy above the 4D convex hull of this work are compared in Fig. 2.

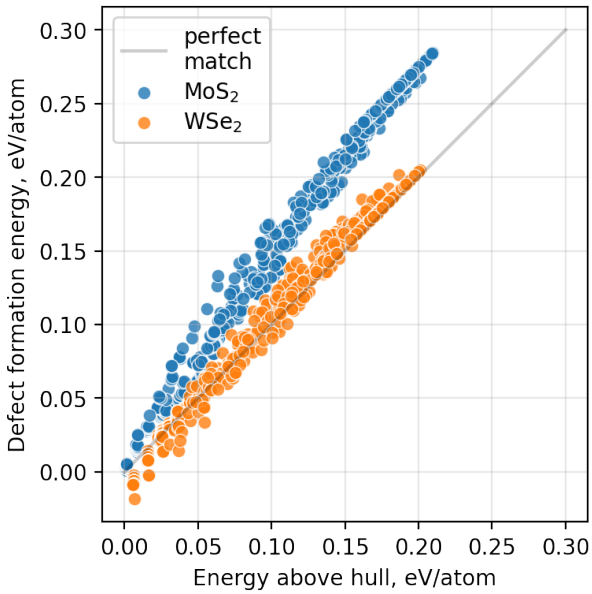

Fig. 2: Comparison of the defect formation energies from the 2DMD dataset and corresponding energies above the convex hull obtained in this work

As can be clearly seen in Fig. 2 (the MoS_2 set), omitting competing phases can result in overestimated (up to 75 meV/atom) values of defect formation energies. Among the MoS_2, WSe_2, InSe, and GaSe base materials studied, such an effect is observed for three out of four sets already at the stage of generating training data. Moreover, the negative formation energies (see the WSe_2 set in Fig. 2) can correspond to the positive energies above the convex hull, which directly points at the presence of a competing phase defining the convex hull.

The authors believe that the developed approach can become a tool for finding competing phases and analyzing their influence in cases where the data obtained using DFT is insufficient for the purposes of reliable training of machine learning models.

Acknowledgments

The authors are grateful to Professor Novoselov's team at the NUS Institute for Functional Intelligent Materials for fruitful discussions on the 2DMD dataset considered in this study.

References

[1] Victor Fung, Jiaxin Zhang, Eric Juarez, and Bobby G Sumpter. Benchmarking graph neural networks for materials chemistry. *npj Computational Materials*, 7:84, 2021.

[2] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021.

[3] Roman A Eremin, Innokentiy S Humonen, Alexey A Kazakov, Vladimir D Lazarev, Anatoly P Pushkarev, and Semen A Budennyy. Graph neural networks for predicting structural stability of Cd- and Zn-doped γ-$CsPbI_3$. *Computational Materials Science*, 232:112672, 2024.

[4] Xiaoze Yuan, Yuwei Zhou, Qing Peng, Yong Yang, Yongwang Li, and Xiaodong Wen. Active learning to overcome exponential-wall problem for effective structure prediction of chemical-disordered materials. *npj Computational Materials*, 9:12, 2023.

[5] P Huang, R Lukin, M Faleev, N Kazeev, A Rashid Al-Maeeni, DV Andreeva, A Ustyuzhanin, A Tormasov, AH Castro Neto, and KS Novoselov. Unveiling the complex structure-property correlation of defects in 2d materials based on high throughput datasets. *npj 2D Materials and Applications*, 7:6, 2023.

[6] Christopher J Bartel. Review of computational approaches to predict the thermodynamic stability of inorganic solids. *Journal of Materials Science*, 57(23):10475–10498, 2022.

[7] https://github.com/AIRI-Institute/2DMD_at_a_Glance.

[8] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14:579, 2023.

[9] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13:2453, 2022.