ISSN 0027-1349, Moscow University Physics Bulletin, 2024, Vol. 79, Suppl. 2, pp. S866–S871. © Allerton Press, Inc., 2024.

Prediction of Defect Structure in MoS_2 by Given Properties

H. E. Karlinski[1]* and M. V. Lazarev[1]**

[1]*HSE University, Moscow, 101000 Russia*

Received October 1, 2024; revised October 10, 2024; accepted October 20, 2024

Abstract—The generation of crystals with tailored properties is a significant challenge in both scientific research and practical applications. Due to the vast configuration space of crystalline structures, finding precise solutions to such problems is computationally intensive. In this study, we propose a method for generating defect configurations in MoS_2 crystals aimed at producing crystals with specific characteristics, focusing on formation energy and HOMO-LUMO energy levels as key examples. The approach leverages symbolic regression techniques, trained on datasets of two-dimensional materials with defects, to predict crystal properties. We introduce methods for identifying defect configurations with both minimal and specific formation energies, as well as for optimizing HOMO-LUMO energy levels. The main advantages of this approach are its efficiency and accuracy in generating valid and optimized crystal structures.

Keywords: 2D materials, MoS_2, machine learning, symbolic regression, evolutionary algorithm

DOI: 10.3103/S0027134924702321

1. INTRODUCTION

Crystals occupy a central position in many modern technologies, playing a crucial role in the development of innovative solutions across various industrial sectors. These materials are the foundation for numerous devices, ranging from microelectronics to quantum computing and energy applications. Precise control over their properties at the atomic level has become a key goal in materials science, enabling the enhancement of existing technologies while opening up new areas of application. Of particular interest is the relatively new class of two-dimensional materials, known for their unique physical properties [1].

The range of applications for two-dimensional materials is vast, and research in this field is expanding rapidly. For instance, materials capable of generating single photons [2] are crucial information carriers for use in quantum computing and quantum cryptography. These materials are increasingly employed in the production of components such as transistors [3], diodes [4], and lasers [5], which are becoming more miniaturized and efficient. In the energy sector, crystals play a vital role in creating solar cells with high energy conversion efficiency [6], and databases containing calculated properties of these materials have been established to facilitate further research [7]. Recently, even major companies like Google have joined the search for new materials, investing heavily in the study of novel materials [8]. Research and development of new crystalline materials continue to drive the creation of more efficient and powerful devices, opening up new opportunities in both scientific research and industry. Methods for precise tuning of material properties are advancing [9], including through the introduction of defects in the material, which can impart novel characteristics [10]. The ability to manipulate atoms at the atomic level allows for the creation of defects with specific properties [11], offering vast potential for the development of new devices based on quantum effects. Some 2D materials have already found commercial applications [12].

However, a significant challenge remains the high cost and long duration of research and implementation of new materials in the industry. On average, the time from invention to application spans decades [13].

Modern methods of computer modeling of crystals based on their atomic structure have become powerful tools for the development of new materials [14]. Key approaches include density functional theory (DFT) [15], which can model quantum effects, and molecular dynamics methods [16], which operate on larger scales but within the framework of classical physics. Nevertheless, methods based on physical model calculations are expensive. Neural networks

*E-mail: gekarlinskiy@edu.hse.ru
**E-mail: mvlazarev@hse.ru

offer the potential to reduce costs and accelerate the prediction of material properties, making the process more efficient.

There are several neural network methods for predicting crystal properties that utilize various material representations [17], such as graph-based representations [18]. These models can be used as filters and hypothesis testing tools. However, a more complex challenge is the generation of crystals with predefined properties. Generative models have been developed to tackle this task [19, 20], but they often prove to be either highly complex to implement or generate unstable crystal structures. Currently, there are no models capable of generating crystal defects that can impart new properties to the material.

This work presents an example of an algorithm for generating defects in the two-dimensional material MoS_2. The algorithm is based on the results of a symbolic regression method [21], which identifies functional dependencies of crystal properties on the positions of defects. The algorithm generates configurations by taking the desired characteristics and a set of selected defects as input, and then it outputs the configuration with properties closest to the desired specifications.

2. DATA AND METHODS

One of the best neural network models for evaluating material properties currently is MEGNet [22]. In MEGNet, a crystal is represented as a graph, where each edge and node is associated with a vector, along with a global vector representing the overall state of the entire crystal graph. A neural network is then constructed to transform these vectors to predict the crystal's properties. However, we have demonstrated that the symbolic regression method operates two orders of magnitude faster, which is critical for the proposed crystal structure generation algorithm.

The foundation of the algorithm presented in Section 3 is based on the results of the SEGVAE algorithm [21]. SEGVAE was applied to uncover the dependencies of material properties on the positions of defect pairs. SEGVAE learns formulas for pairwise interactions of defects (an example is shown in Fig. 1), from which a general formula for any number of defects in the material can be constructed. These formulas can be used for rapid prediction of the properties of the generated material (formulas (1), (2))

$$E_{\text{formation}} = \sum_{i \in N} E_i + \sum_{i,j \in N, i<j} V_{ij}(\text{dist}(i,j)), \quad (1)$$

$$E_{\text{hl}} = \min_{i,j \in N, i<j} |H_{ij}(\text{dist}(i,j)) - L_{ij}(\text{dist}(i,j))|. \quad (2)$$

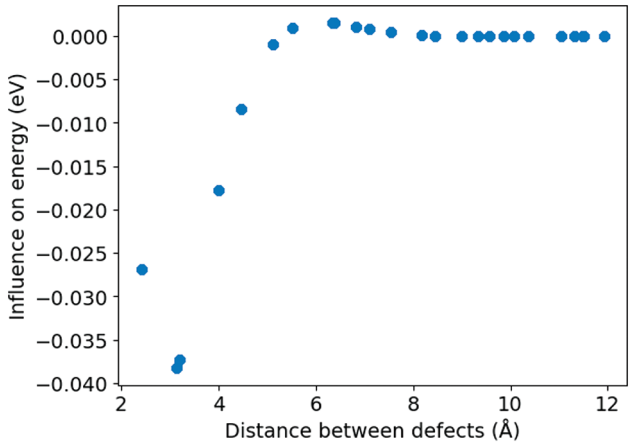

Fig. 1. Example of data for finding the formula for pairwise interaction (between W and vacancy Mo).

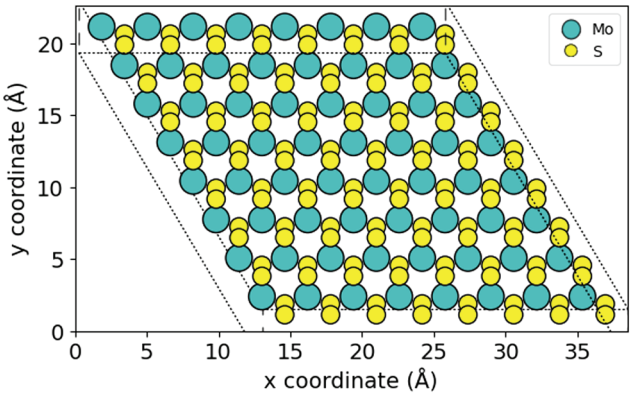

Fig. 2. Example of a MoS_2 8×8 lattice without defects.

The formula for formation energy $E_{\text{formation}}$ (1): E_i is the contribution of defect i, V_{ij} is the formula for the contribution of interaction between defects of defect types i and j distance dependent, $\text{dist}(i,j)$ is the distance between defects i and j.

Formula for HOMO-LUMO E_{hl} (2): H_{ij} is the HOMO function for a pair of defects i and j, L_{ij} similar to the LUMO function, $\text{dist}(i,j)$ is the distance between defects.

The data for training the algorithm were obtained from the dataset [23]. A set of structures based on MoS_2 with a cell size of 8×8 MoS_2 molecules was considered. An example of a defect-free "pristine" crystal is shown in Fig. 2.

The dataset contains the following set of defects:

- vacancy Mo (molybdenum);
- vacancy S (sulfur);

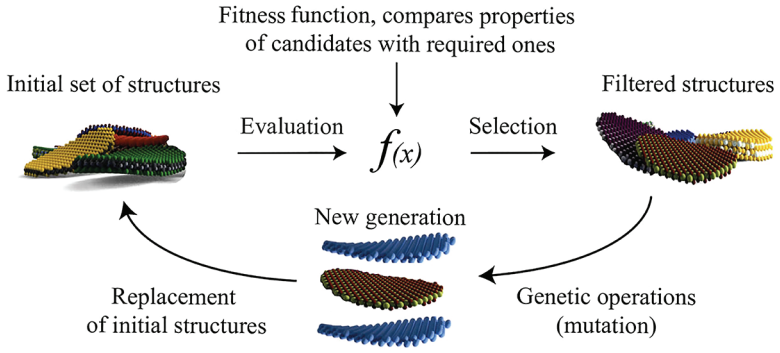

Fig. 3. General picture of genetic algorithms.

- substitution Mo (molybdenum) on W (tungsten);
- substitution S (sulfur) on Se (selenium).

One of the advantages of our approach is that we can evaluate the properties of a crystal of any size and with any number of defects, unlike neural network approaches. We limited the cell size to 16×16 to isolate the defects under consideration from neighboring defects due to the periodic boundary conditions of the crystal lattice. This is equivalent to assuming that outside the region of interest, there are no defects, as the distance over which defect interactions significantly affect the material's properties is limited to 10 Å. To present the generation results, we will focus on the central 8×8 zone of the crystal for aesthetic reasons.

For the task of defect generation, we will consider the crystal lattice as three matrices (the first sulfur layer, the molybdenum layer, and the second sulfur layer), where each layer can contain two unique types of defects (one type of vacancy and one type of substitution).

In this work, we aim to find configurations with specific properties from a predefined set of defects (the number of each defect in the final lattice is known in advance). Defects associated with sulfur atoms are not differentiated between the first and second layers. Although such separation could improve prediction accuracy, it was shown in [24] that the formulas derived from symbolic regression do not underperform MEGNet in terms of accuracy on the MoS_2 dataset when working with formation energy.

3. ALGORITHM

It is important to examine how the methods for predicting crystal properties behave, focusing primarily on the formulas obtained through symbolic regression using SEGVAE. Notably, the pairwise interaction formulas are predominantly monotonic functions, with a few exceptions. This leads to the following hypothesis: it is possible to achieve a reasonably optimal arrangement of defect configurations through small iterative adjustments. To accomplish this, we will employ an evolutionary algorithm. In a nutshell an evolutionary process can be presented as it depicted in Fig. 3.

The input to the algorithm (see Fig. 4) consists of the number of each type of defect and the number of evolution epochs. The algorithm then generates a random arrangement of the input defects and proceeds with the evolution process:

- Each arrangement (set) undergoes mutation: a random defect is selected and moved to a number of possible positions within a specified radius (the exact behavior is controlled by hyperparameters). For each new position, a new set is generated.
- The resulting sets are then added to the list for the next epoch: the top N_{best} sets with the closest target variable, N_{rand} random sets from the current epoch, and N_{mutate} sets with the closest target variable are added again, but with additional mutation.

This iterative evolutionary process refines the defect configurations to optimize the desired properties of the crystal.

After the evolution process, the set closest to the target variable is selected, and greedy optimization is applied: each defect is greedily moved to the position that optimizes the target variable.

The implementation is based on batches, which act as rule executors for selecting sets for the next epoch: *top_batch*, *random_batch*, and *mutate_batch* (additional batches can be easily added). All of these are encapsulated within the

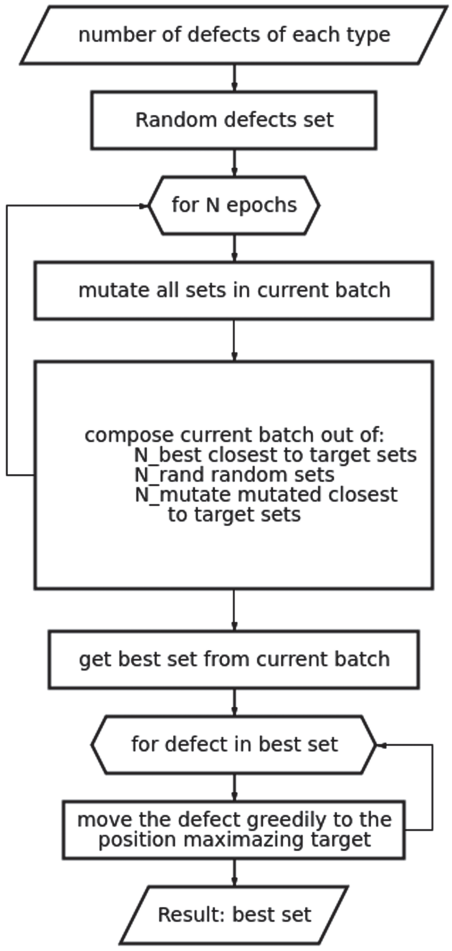

Fig. 4. Evolutionary algorithm flowchart.

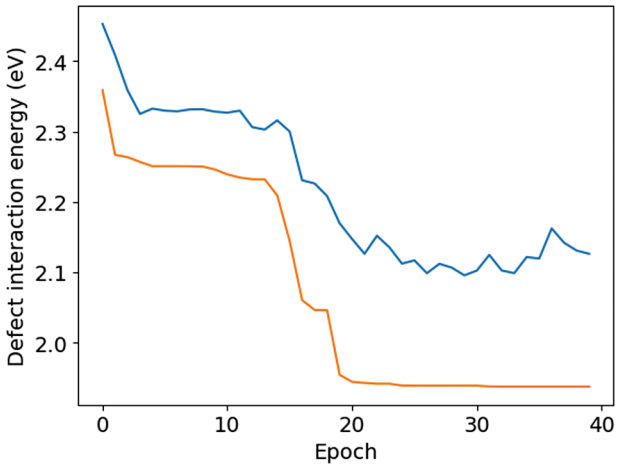

Fig. 5. Formation energy in generations.

batch_operator structure, which manages operations such as retrieving the current sets stored inside and processing a new batch of sets (all mutated sets are passed as input, and each batch retains a portion of them for further processing). Sets themselves are structures that contain a list of defects and a function to calculate the target value. This modular approach allows for flexibility and scalability in managing the evolutionary process and enables efficient selection of defect configurations that move closer to the desired crystal properties.

The algorithm converges to a local minimum and no longer changes the found arrangement after an average of 20 epochs with parameters $N_{\text{best}} = 15$, $N_{\text{rand}} = 30$, $N_{\text{mutate}} = 20$ (see Fig. 5).

Up until this point, we have considered the case where the specific defects to be arranged on the lattice are known. Now, let us address the problem of finding the formation energy for an arbitrary set of defects. We will start by generating an initial set of defects and then, after fixing it, we will run our optimization algorithm. Let us revisit the formula for formation energy (1).

To optimize the solution for an arbitrary set of defects in the arrangement, we will filter out all sets where the sum of E_i exceeds the target value, as well as all sets where the sum of E_i and $\max(V_{ij})$ is less than the target value. Since E_i is several orders of magnitude larger than the values of the functions V_{ij}, this optimization will be efficient.

By applying this filtering, we ensure that only the most promising configurations are processed, significantly speeding up the search for optimal defect arrangements that yield the desired formation energy.

4. RESULTS

Using batch size parameters of 10 for the top batch, 15 for the random batch, and 9 for the mutate batch, the generation of a single structure with 8 defects takes an average of 1.52 s, running on an AMD Ryzen 5 5500U. By starting with different initial random configurations, various final configurations can be obtained (see Fig. 6). The performance results are largely driven by the functional dependence of the properties on the positions of the defects. Another advantage of this approach is its simplicity, which does not require complex software dependencies, making it easy to use.

The accuracy of the results and comparison with the baseline are presented in Table 1. The baseline is a greedy arrangement, identical to the one executed after the evolutionary optimization. Target formation energy was set to 2.4 eV and HOMO-LUMO gap to 0.41 eV.

The evolutionary algorithm achieves greater accuracy in finding the arrangement with the minimal

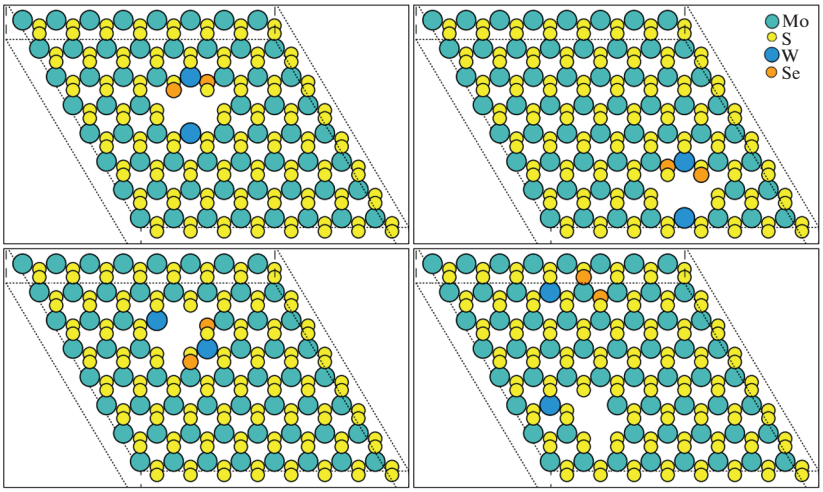

Fig. 6. Variations of generated defect formations minimizing the formation energy.

Table 1. Structure generation task with given structure properties

Algorithm	Target energy [eV]	Error min [meV]	Error mean [meV]	Std [meV]	No. epochs
Our formation energy	2.4	4.16e−5	0.349	0.899	20
Baseline formation energy	2.4	3.50e−4	1.84	2.78	–
Our HOMO-LUMO gap	∼0.41	1.31	1.31	1.28e−7	20
Baseline HOMO-LUMO gap	∼0.41	1.31	1.31	1.32e−7	–

Table 2. Minimum formation energy task

Minimal energy	Energy min [eV]	Energy mean [eV]	Etd [eV]	No. epochs
Our formation energy	1.9379	1.95	0.029	20
Baseline formation energy	1.939	1.998	0.057	–

formation energy, as well as in locating specific formation energies. However, when predicting HOMO-LUMO energy levels, the greedy arrangement yields the same results as the evolutionary approach.

In the case of searching for defect configurations with minimal formation energy, it has been experimentally observed that the minimum formation energy is achieved when defects cluster in a single location. Results for structure with 8 defects (two of each kind) presented in Table 2 the example of few structures from optimised batch presented in Fig. 6. However, there are instances where the defects do not cluster together (e.g., the fourth structure in Fig. 6). This occurs because the algorithm gets trapped in a local energy minimum, whereas the optimal—minimal—energy is achieved when the defects are clustered into a single group. This behavior is driven by the interaction functions between the defects.

CONCLUSIONS AND OUTLOOK

The result of this work is the demonstration of an algorithm for constructing defects in MoS_2 crystals based on desired properties (formation energy and HOMO-LUMO), enabled by a system of discrete optimization and the results of symbolic regression. This method of rapid defect configuration generation has certain limitations. For instance, the accuracy of the method remains dependent on how the properties of the crystal with defects are evaluated. Currently, the method works for MoS_2, but it could potentially be generalized to other materials. However, if symbolic regression results are used, it would be necessary to study the formulas for other pairwise defect

interactions. Alternatively, a pretrained neural network could be used, though this would significantly slow down the generation process.

FUNDING

The article was prepared within the framework of the project "Mirror Laboratories" of the HSE University, Russia.

CONFLICT OF INTEREST

The authors of this work declare that they have no conflicts of interest.

REFERENCES

1. P. Ares and K. S. Novoselov, Nano Mater. Sci. **4**, 3 (2022). https://doi.org/10.1016/j.nanoms.2021.05.002
2. P. K. Chow, R. B. Jacobs-Gedrim, J. Gao, T.-M. Lu, B. Yu, H. Terrones, and N. Koratkar, ACS Nano **9**, 1520 (2015). https://doi.org/10.1021/nn5073495
3. F. Xia, H. Wang, D. Xiao, M. Dubey, and A. Ramasubramaniam, Nat. Photonics **8**, 899 (2014). https://doi.org/10.1038/nphoton.2014.271
4. F. Withers, O. Del Pozo-Zamudio, A. Mishchenko, A. P. Rooney, A. Gholinia, K. Watanabe, T. Taniguchi, S. J. Haigh, A. K. Geim, A. I. Tartakovskii, et al., Nat. Mater. **14**, 301 (2015). https://doi.org/10.1038/nmat4205
5. W. Du, C. Li, J. Sun, H. Xu, P. Yu, A. Ren, J. Wu, and Zh. Wang, Laser Photonics Rev. **14**, 2000271 (2020). https://doi.org/10.1002/lpor.202000271
6. S. Das, D. Pandey, J. Thomas, and T. Roy, Adv. Mater. **31**, 1802722 (2019). https://doi.org/10.1002/adma.201802722
7. C. J. Court, B. Yildirim, A. Jain, and J. M. Cole, J. Chem. Inf. Model. **60**, 4518 (2020). https://doi.org/10.1021/acs.jcim.0c00464
8. A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, Nature **624**, 80 (2023). https://doi.org/10.1038/s41586-023-06735-9
9. Q. Ma, G. Ren, K. Xu, and J. Zh. Ou, Adv. Opt. Mater. **9**, 2001313 (2021). https://doi.org/10.1002/adom.202001313
10. J. Jiang, T. Xu, J. Lu, L. Sun, and Zh. Ni, Research **2019**, 4641739 (2019). https://doi.org/10.34133/2019/4641739
11. I.-J. Chen, M. Aapro, A. Kipnis, A. Ilin, P. Liljeroth, and A. S. Foster, Nat. Commun. **13**, 7499 (2022). https://doi.org/10.1038/s41467-022-35149-w
12. K. Khan, A. Kh. Tareen, M. Aslam, R. Wang, Yu. Zhang, A. Mahmood, Zh. Ouyang, H. Zhang, and Zh. Guo, J. Mater. Chem. C **8**, 387 (2020). https://doi.org/10.1039/C9TC04187G
13. E. Maine and P. Seegopaul, Nat. Mater. **15**, 487 (2016). https://doi.org/10.1038/nmat4625
14. J. Hafner, J. Comput. Chem. **29**, 2044 (2008). https://doi.org/10.1002/jcc.21057
15. R. G. Parr, Sh. R. Gadre, and L. J. Bartolotti, Proc. Natl. Acad. Sci. U. S. A. **76**, 2522 (1979). https://doi.org/10.1073/pnas.76.6.2522
16. J. Hénin, T. Lelièvre, M. R. Shirts, O. Valsson, and L. Delemotte, Living Journal of Computational Molecular Science **4**, 1583 (2022). https://doi.org/10.33011/livecoms.4.1.1583
17. Sh. Li, Yu. Liu, D. Chen, Yi. Jiang, Zh. Nie, and F. Pan, WIREs Computational Molecular Science **12**, e1558 (2022). https://doi.org/10.1002/wcms.1558
18. T. Xie and J. C. Grossman, Phys. Rev. Lett. **120**, 145301 (2018). https://doi.org/10.1103/PhysRevLett.120.145301
19. S. Kim, J. Noh, G. H. Gu, A. Aspuru-Guzik, and Yo. Jung, ACS Cent. Science **6**, 1412 (2020). https://doi.org/10.1021/acscentsci.0c00426
20. T. Xie, X. Fu, O.-E. Ganea, R. Barzilay, and T. Jaakkola, arXiv Preprint (2021). https://doi.org/10.48550/arXiv.2110.06197
21. S. Popov, M. Lazarev, V. Belavin, D. Derkach, and A. Ustyuzhanin, PeerJ Comput. Sci. **9**, e1241 (2023). https://doi.org/10.7717/peerj-cs.1241
22. Ch. Chen, W. Ye, Yu. Zuo, Ch. Zheng, and Sh. P. Ong, Chem. Mater. **31**, 3564 (2019). https://doi.org/10.1021/acs.chemmater.9b01294
23. P. Huang, R. Lukin, M. Faleev, N. Kazeev, A. R. Al-Maeeni, D. V. Andreeva, A. Ustyuzhanin, A. Tormasov, A. H. Castro Neto, and K. S. Novoselov, npj 2D Mater. Appl. **7**, 6 (2023). https://doi.org/10.1038/s41699-023-00369-1
24. M. Lazarev, in *Proceedings of the Conference AI4AM, Barcelona, 2024* (2024).