Symbolic regression for defects interactions in MoS_2 and WSe_2

Mikhail Lazarev[1], Andrey Ustyuzhanin[2]
[1]HSE University, Myasnitskaya Ulitsa, 20, Moscow, Russia
[2]Constructor University, Campus Ring 1, Bremen, 28759, Germany

mvlazarev@hse.ru

The advent of neural network (NN) methodologies and expansive databases of materials has incited the deployment of deep learning techniques for atomistic predictions. Machine Learning (ML) algorithms, underpinned by datasets derived via Density Functional Theory (DFT), have found widespread application. These ML strategies expedite the design of novel materials by forecasting material properties with a fidelity approaching that of ab-initio calculations, albeit at significantly reduced computational costs. The past few years have witnessed the introduction of several rapid and precise deep learning architectures. Among these, graph NNs, such as MEGNet [1], CGCNN [2], SchNet [3], and GemNet [4], have proven to be particularly effective. Nonetheless, this approach is not without its limitations, including the requirements for bid data volumes, the high cost of model training, lack of result interpretability, and generalization challenges. Meanwhile, symbolic expressions elucidate a clear relationship between observations and the target variable.

Historically, genetic algorithms have predominantly served as the methods for symbolic regression [5]. Due to their inherent advantages, this paradigm has been employed across various domains within materials science [6][7][8]. However, recent years have seen the development of numerous symbolic regression techniques founded on NN paradigms [9][10][11][12]. In the current study, we employed SEGVAE [13], owing to its simplicity and efficiency in processing small datasets. The SEGVAE algorithm, based on a Variational Autoencoder (VAE), is adept at identifying a suite of formulas that describe the observed data. The fundamental operational schema of the algorithm is depicted in Fig 1. We advocate for the application of SEGVAE to delineate the interaction of defects in 2D materials and their physical attributes. The utilization of symbolic regression techniques would enable the discovery of novel functional representations for the dependency of properties on the defect structure within materials. In this investigation, we employed a dataset of two-dimensional materials featuring defects [14], wherein the properties of these materials were simulated using the VASP software. The simulation approach is anchored in a physical model that utilizes the Density Functional Theory (DFT) methodology.

Employing the SEGVAE method for symbolic regression, we successfully identified formulas characterizing the pairwise interactions of defects within the MoS_2 and WSe_2 crystals. An illustrative depiction of how the properties of a structure with two defects are influenced by the positioning of the defect pair in MoS_2 is shown in fig. 2 (a)(b). An example for a structure with three defects and all correspondent interactions is illustrated in fig. 2 (c)(d). For this, a small dataset of defect pairs of each type was selected to approximate interaction law by a formula derived from the SEGVAE algorithm. The resulting formula for the formation energy of a structure for a material with N defects can be approximated as: $E_{formation} = \sum_{i\in N} E_i + \frac{1}{2}\sum_{i,j\in N} V_{i,j}(r)$, where E_i is the formation energy of i-th defect, $V_{i,j}(r)$ is the value of interaction function between i-th and j-th defects, r distance between the defects and N is a number of defects.

The primary feature of this approach lies in its generalizability and interpretability. We used only a few hundred structural examples to learn all pairwise interactions to construct a formula for structures with an arbitrary number of defects. While NNs used almost 15000 structures to train. The functional form of the discovered dependency ensures the interpretability of the results. Remarkably, the found energy values significantly surpass many NN approaches, only fail compared to Sparse (MEGNet) as shown in fig. 3 whereas quality metric we used mean absolute error (MAE) of formation energy per site: $E_{per\_site} = \frac{E_{formation}}{N}$

We also tested the pairwise decomposition method to determine the band gap defined as difference between LUMO and HOMO. To fit band gap value for arbitrary number of defects in structure we used the following formula: $Band\_gap = \min_{i,j\in N}\left(Pair\_band\_gap_{i,j}(r)\right)$, where $Pair\_band\_gap_{i,j}(r)$ is the value of band gap function of the structure with i-th and j-th defects and r distance between the defects. As with the energy, SEGVAE was used to derive a function for the band gap value dependency on the distance between two defects. The formulas thus derived serve as approximations, given the lack of an analytical formula or law in nature that accurately describes such a dependency. Nonetheless, this methodology proficiently captures the target values, outperforming many other sophisticated algorithms or similar to MEGNet in terms of accuracy. The comparison results for band gap of this approach with other NN methods are presented in fig. 3. One more undauntable advantage of our method compared to NNs is the speed of obtaining predicted values.

References

[1] Chi Chen et al. “Graph networks as a universal machine learning framework for molecules and

crystals”. In: Chemistry of Materials 31.9 (2019), pp. 3564–3572.

[2] Park, Cheol Woo, and Chris Wolverton. "Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery." Physical Review Materials 4, no. 6 (2020): 063801.

[3] Kristof Schutt et al. “Schnet: A continuous-filter convolutional neural network for modeling quantum interactions”. In: Advances in neural information processing systems 30 (2017).

[4] Johannes Gasteiger, Florian Becker, and Stephan Gunnemann. “Gemnet: Universal directional graph neural networks for molecules”. In: Advances in Neural Information Processing Systems 34 (2021), pp. 6790–6802.

[5] Dominic P Searson, David E Leahy, and Mark J Willis. “GPTIPS: an open source genetic programming toolbox for multigene symbolic regression”. In: Proceedings of the International multiconference of engineers and computer scientists. Vol. 1. Citeseer. 2010, pp. 77–80.

[6] Yiqun Wang, Nicholas Wagner, and James M Rondinelli. “Symbolic regression in materials science”. In: MRS Communications 9.3 (2019), pp. 793–805.

[7] Eibar Flores et al. “Learning the laws of lithium-ion transport in electrolytes using symbolic regression”. In: Digital Discovery 1.4 (2022), pp. 440–447.

[8] Mu He and Lei Zhang. “Machine learning and symbolic regression investigation on stability of MXene materials”. In: Computational Materials Science 196 (2021), p. 110578.

[9] Matthias Werner et al. “Informed equation learning”. In: arXiv preprint arXiv:2105.06331(2021).

[10] Silviu-Marian Udrescu et al. “AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity”. In: Advances in Neural Information Processing Systems 33 (2020), pp. 4860–4871.

[11] Raban Iten et al. “Discovering physical concepts with neural networks”. In: Physical review letters 124.1 (2020), p. 010508.

[12] T Nathan Mundhenk et al. “Symbolic regression via neural-guided genetic programming population seeding”. In: arXiv preprint arXiv:2111.00053 (2021).

[13] Sergei Popov et al. “Symbolic expression generation via Variational Auto-Encoder”. In: arXiv preprint arXiv:2301.06064 (2023).

[14] Pengru Huang et al. “Unveiling the complex structure-property correlation of defects in 2D materials based on high throughput datasets”. In: npj 2D Materials and Applications 7.1 (2023), p. 6.

Figures

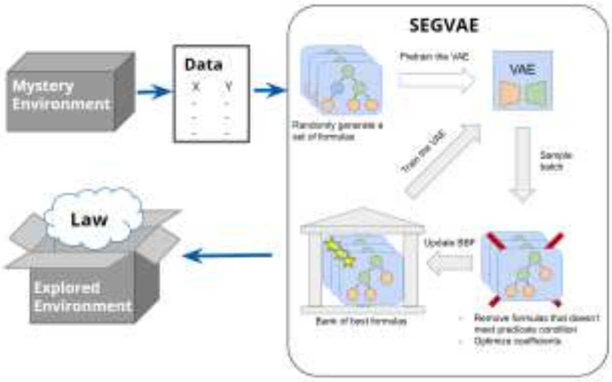


Figure 1. The SEGVAE application, architecture, and training scheme. Adapted from [13]

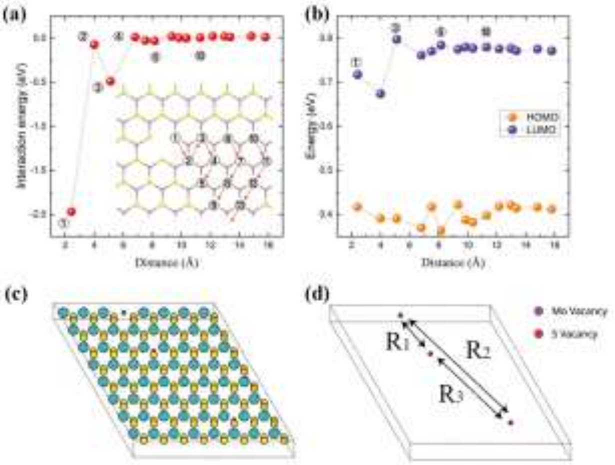
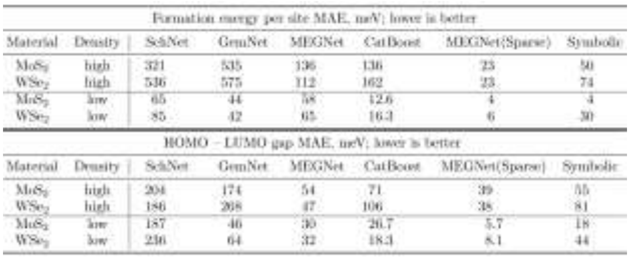


Figure 2. Defects in MoS_2 and it's properties. (a) interaction energy, (b) HOMO, and LUMO as a function of the distance between the Mo vacancy and the S vacancy. The inset labels the positions of the nearest S sites to the Mo vacancy. (c) MoS_2 structure with 3 defects, 1 Mo, 2 S vacancies. (d) defects separate from the structure. (a)(b) Taken from [14]



Formation energy per site MAE, meV; lower is better							
Material	Density	SchNet	GemNet	MEGNet	CatBoost	MEGNet(Sparse)	Symbolic
MoS_2	high	321	535	136	136	23	50
WSe_2	high	536	575	112	162	23	74
MoS_2	low	65	44	58	12.6	4	4
WSe_2	low	85	42	65	16.3	6	30

HOMO – LUMO gap MAE, meV; lower is better							
Material	Density	SchNet	GemNet	MEGNet	CatBoost	MEGNet(Sparse)	Symbolic
MoS_2	high	204	174	54	71	39	55
WSe_2	high	186	268	47	106	38	81
MoS_2	low	187	46	30	26.7	5.7	18
WSe_2	low	236	64	32	18.3	8.1	44

Figure 3. Performance of the different methods in terms of the mean absolute error (MAE in meV) on 2d materials dataset with defects [14]. Symbolic is our method.