Accurate and robust cluster expansion of multicomponent systems via comprehensive physics-guided featurization

Xufa Huang[a], **Leng Ze Tang**[a], **Chryston Boo**[b], **Yun Liu**[b],

Kedar Hippalgaonkar[a], **Zhidong Leong**[b], **Teck Leong Tan**[b]

[a] *School of Materials Science and Engineering, Nanyang Technological University, 50 Nanyang Ave, #01-30, Block N4.1, Singapore 639798, Singapore* hu0002fa@e.ntu.edu.sg, tang0482@e.ntu.edu.sg, kedar@ntu.edu.sg

[b] *Institute of High Performance Computing, Agency for Science, Technology and Research, 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Singapore* chryston_Boo@ihpc.a-star.edu.sg, liu_Yun@ihpc.a-star.edu.sg, leong_zhidong@ihpc.a-star.edu.sg, tantl@ihpc.a-star.edu.sg

1. Introduction

Cluster expansion (CE) is a popular surrogate model for capturing structure-property relationships and modeling alloy formation energies [1-7]. CE follows a generalized Ising model [8] shown in Eq. 1, where it expands the formation energy $E(\sigma)$ of an alloy configuration, σ, in terms of atomic clusters, c, such that the cluster correlation functions $\Phi_c(\sigma)$ serve as a basis set and the effective cluster interactions (ECIs), V_c, are the coefficients.

$$E(\sigma) = \sum_c V_c \Phi_c(\sigma). \qquad (1)$$

Figure 1 illustrates that the traditional CE approach works well for systems with low atomic size mismatch (ASM) such as Mo-Nb, where long range order contributions from higher-ordered clusters are negligible. In alloy systems like Mo-Zr, where significant ASM is present, traditional CE fails to capture the alloy's structural-property behavior due to high lattice distortion.

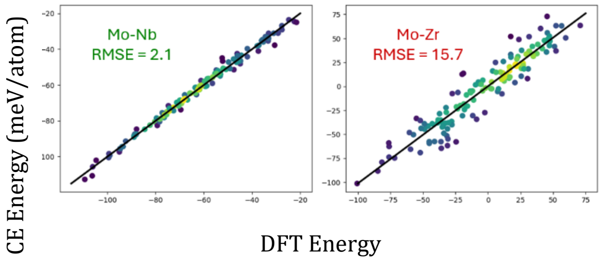

Fig. 1: Parity plots showing that CE is accurate for Mo-Nb (left), a representative low ASM binary system, but fails for Mo-Zr (right), a representative of high ASM binary system.

In this work, we adopt physics-guided featurization to construct accurate and robust CE for multicomponent systems including bulk binary alloys (binary constituents of Mo-V-Nb-Ti-Zr) and alloyed perovskite systems (Pb-based ABX_3).

2. Method

We propose the novel integration of a diverse set of descriptors from the Matminer library [9] with traditional CE in modeling formation energies. These descriptors (e.g., radial distribution function, sine coulomb matrix etc.) are curated based on the physics of diverse materials databases. While clusters in CE primarily capture short-range order interactions, Matminer descriptors extend this capability to better capture long-range order interactions. This synergy between Matminer and CE allows for extracting a more comprehensive set of structural, chemical and geometrical features, which serve as superior predictors of formation energies in high ASM systems. We further propose a high-throughput recursive feature elimination (HTRFE) machine learning pipeline for robust feature selection over traditional CE.

2.1 Related work

Recent works have sought to enhance the robustness of traditional CE by introducing grouped regularized regression [6, 10-13]. These methods improve feature selection by grouping important features together, handling complex correlated dependencies and ensuring the collective selection of interdependent features for configurational energy prediction. However, manually defining groups may inadvertently introduce biases or overlook important interactions among features. Beyond regression approaches, Bayesian optimization [14] and neural networks [15] have also been explored as enhancement over traditional CE. Meanwhile, Matminer descriptors have been extensively employed in surrogate models to predict properties such as bandgap [16, 17], elasticity [18], thermal conductivity [19] etc. for diverse classes of materials. This work is the first to integrate Matminer descriptors within the CE framework, expanding the scope of physics-guided featurization in CE.

3. Results

Fig. 2 illustrates the performance of different modeling approaches, with traditional CE combined with lasso (blue) serving as the baseline model. Upon addition of Matminer features without the HTRFE pipeline, traditional CE's performance declined due to overfitting (yellow). Applying the proposed pipeline allowed us to extract the important features, leading to significant improvement in prediction performance across all ten binary alloys (green). Further analysis of feature importance weights revealed that the prediction performance remained high using only the set of four most important feature sets (purple). These results highlight the necessity of both the HTRFE pipeline and Matminer descriptors for achieving accurate and robust predictions of formation energy. To assess the transferability of our proposed HTRFE pipeline with the integration of Matminer features and CE, we extended the approach to four Pb-based ABX_3 perovskite systems with varying degrees of ionic size mismatch, where A is Cs, B is Pb, which is substituted by the respective alkaline earth metals, and X is Br. Fig. 3 shows that our proposed method consistently outperforms traditional CE in the perovskite systems.

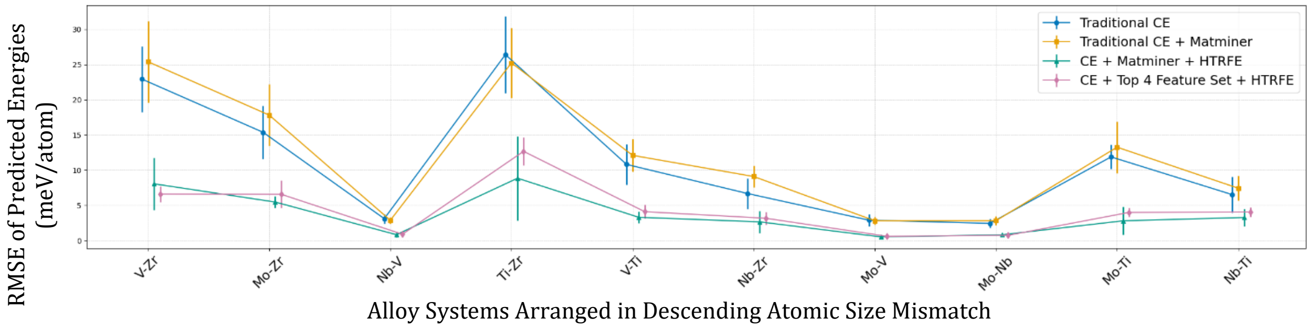

Fig. 2: Compared to traditional CE (blue), the addition of Matminer descriptors (yellow) worsens the overfitting. The proposed HTRFE pipeline trained on CE clusters and Matminer descriptors (green) results in major error reduction of 38% to 77%, averaging 56% across the ten binary alloys, with absolute improvement of more than 10 meV/atom for some high atomic size mismatch systems like V-Zr, Mo-Zr and Ti-Zr. The refined HTRFE with only top four feature classes shows comparable performance across all ten binary alloy systems (purple).

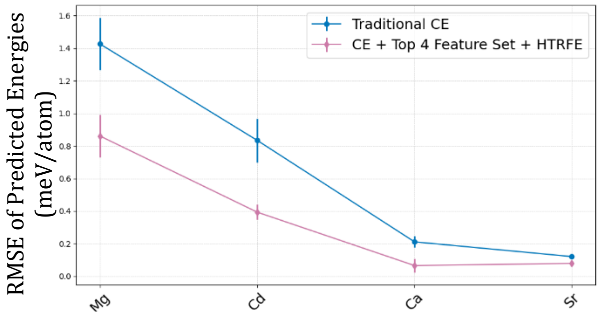

Fig. 3: Compared to traditional CE (blue), our proposed HTRFE pipeline with the integration of Matminer features and CE reduced the error by 34% to 68%, averaging 49% across the four perovskites, with significant absolute improvements in the high ASM Mg and Cd systems.

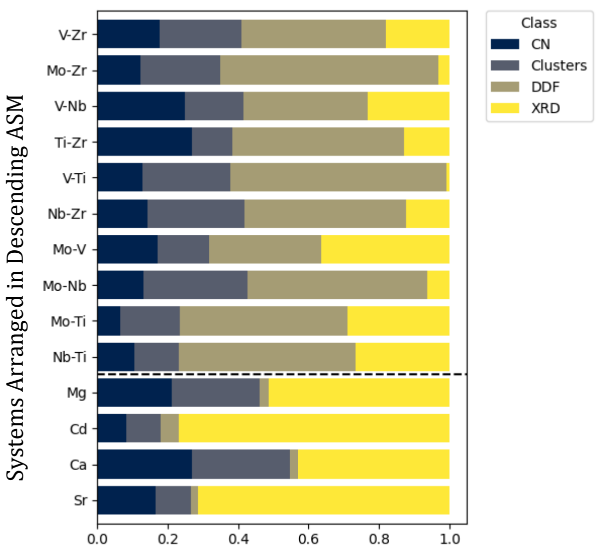

Fig. 4: Normalized feature importance for formation energy prediction in alloy and perovskite systems. Bars represent dashed line separates alloy systems (top) from perovskite systems (bottom), highlighting their differing dependencies on long-range (XRD, CN) versus short-range (Clusters, DDF) structural descriptors.

By analyzing the feature importance in each system shown in Fig. 4, we found that while clusters are unsurprisingly important, three additional classes of descriptors are consistently selected across all ten systems: —coordination number (CN), XRD, and dihedral-angle distribution function (DDF). These findings highlight the method's ability to capture both long-range and short-range order across diverse systems.

4. Discussion

Our results demonstrate distinct roles of various descriptors in predicting formation energy in perovskite and binary alloy systems. DDF and cluster descriptors dominate in alloys due to their heightened sensitivity to short-range interactions [20]. XRD descriptors (peak positions, intensities, broadening) effectively capture long-range lattice distortions in perovskites, where its stability relies predominantly on maintaining long-range structural order arising from deviations in tolerance factors [21] and octahedral tilting [22]. CN descriptors quantify chemical bonding environments, which are essential to the adherence to bond valence sum [23]. The demonstrated feature relevance could guide the selection of stability predictors for novel materials, potentially reducing reliance on computationally expensive DFT calculations.

5. Conclusion

Our study demonstrates that integrating Matminer descriptors with CE significantly enhances the prediction accuracy of formation energy models across both alloy and perovskite systems. By systematically selecting key features through the proposed HTRFE pipeline, we capture both short-range (Clusters, DDF) and long-range (XRD, CN) effects, enabling fast, robust and accurate modeling for a diverse set of structures. Our results show that perovskite stability is predominantly governed by long-range periodicity, while alloys rely on short-range interactions. The method effectively generalizes beyond binary alloys, reducing prediction errors in binary alloys by 56% and in perovskite systems by an average of 49%. Our findings highlight the importance of physics-guided feature selection in extending traditional CE-based models to complex, high ASM materials, offering a scalable and transferable framework for applying CE to increasingly complex systems.

Acknowledgments
The authors thank the Agency for Science, Technology and Research (A*STAR) for providing the A*STAR Graduate Scholarship (AGS) and financial support. The DFT computations in this article were performed on the resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

[1] Lerch, D., Wieckhorst, O., Hart, G. L. W., Forcade, R. W., & Müller, S. (2009). UNCLE: a code for constructing cluster expansions for arbitrary lattices with minimal user-input. *Modelling and Simulation in Materials Science and Engineering*, *17*(5), 055003. https://doi.org/10.1088/0965-0393/17/5/055003

[2] van de Walle, A., Asta, M., & Ceder, G. (2002). The Alloy Theoretic Automated Toolkit: A user guide. In *arXiv [cond-mat.stat-mech]* (Issue 4, pp. 539–553). https://ceder.berkeley.edu/publications/avdv_Calphad_26(4)_539_2002.pdf

[3] Ångqvist, M., Muñoz, W. A., Rahm, J. M., Fransson, E., Durniak, C., Rozyczko, P., Rod, T. H., & Erhart, P. (2019). icet - A Python library for constructing and sampling alloy cluster expansions. In *arXiv [cond-mat.mtrl-sci]*. http://arxiv.org/abs/1901.08790

[4] Chang, J. H., Kleiven, D., Melander, M., Akola, J., Garcia-Lastra, J. M., & Vegge, T. (2019). CLEASE: a versatile and user-friendly implementation of cluster expansion method. Journal of Physics. Condensed Matter: An Institute of Physics Journal, 31(32), 325901. https://doi.org/10.1088/1361-648X/ab1bb

[5] Dana, A., Mu, L., Gelin, S., Sinnott, S., & Dabo, I. (2023). Cluster expansion by transfer learning for phase stability predictions. Computational Materials Science. https://doi.org/10.1016/j.commatsci.2024.113073

[6] Leong, Z., & Tan, T. L. (2019). Robust cluster expansion of multicomponent systems using structured sparsity. In arXiv [cond-mat.mtrl-sci]. http://arxiv.org/abs/1910.08086

[7] Rigamonti, S., Troppenz, M., Kuban, M., Hübner, A., & Draxl, C. (2024). CELL: a Python package for cluster expansion with a focus on complex alloys. *Npj Computational Materials*, *10*(1). https://doi.org/10.1038/s41524-024-01363-x

[8] Huang, W., Kitchaev, D. A., Dacek, S., Rong, Z., Urban, A., Cao, S., Luo, C., & Ceder, G. (2016). Finding and proving the exact ground state of a generalized Ising model by convex optimization and MAX-SAT. In *arXiv [cond-mat.dis-nn]*. http://arxiv.org/abs/1604.06722

[9] Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N. E. R., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., Chard, K., Asta, M., Persson, K. A., Snyder, G. J., Foster, I., & Jain, A. (2018). Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, *152*, 60–69. https://doi.org/10.1016/j.commatsci.2018.05.018

[10] Barroso-Luque, L., Zhong, P., Yang, J. H., Xie, F., Chen, T., Ouyang, B., & Ceder, G. (2022). Cluster expansions of multicomponent ionic materials: Formalism and methodology. *Physical Review. B*, *106*(14). https://doi.org/10.1103/physrevb.106.144202

[11] Yang, J. H., Chen, T., Barroso-Luque, L., Jadidi, Z., & Ceder, G. (2022). Approaches for handling high-dimensional cluster expansions of ionic systems. *Npj Computational Materials*, *8*(1), 1–11. https://doi.org/10.1038/s41524-022-00818-3

[12] N. Simon and R. Tibshirani, Standardization and the group lasso penalty, Stat. Sinica 22, 983 (2012).

[13] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, A sparse-group lasso, J. Comput. Graphical Stat. 22, 231 (2013).

[14] Mueller, T., & Ceder, G. (2009). Bayesian approach to cluster expansions. Physical Review. B, Condensed Matter and Materials Physics, 80(2). https://doi.org/10.1103/physrevb.80.024103

[15] Natarajan, A. R., & Van der Ven, A. (2018). Machine-learning the configurational energy of multicomponent crystalline solids. Npj Computational Materials, 4(1). https://doi.org/10.1038/s41524-018-0110-y

[16] Borlido, P., Schmidt, J., Huran, A. W., Tran, F., Marques, M. A. L., & Botti, S. (2020). Exchange-correlation functionals for band gaps of solids: benchmark, reparametrization and machine learning. *Npj Computational Materials*, *6*(1), 1–17. https://doi.org/10.1038/s41524-020-00360-0

[17] Marchenko, E. I., Fateev, S. A., Petrov, A. A., Korolev, V. V., Mitrofanov, A., Petrov, A. V., Goodilin, E. A., & Tarasov, A. B. (2020). Database of two-dimensional hybrid perovskite materials: Open-access collection of crystal structures, band gaps, and atomic partial charges predicted by machine learning. *Chemistry of Materials: A Publication of the American Chemical Society*, *32*(17), 7383–7388. https://doi.org/10.1021/acs.chemmater.0c02290

[18] Revi, V., Kasodariya, S., Talapatra, A., Pilania, G., & Alankar, A. (2021). Machine learning elastic constants of multi-component alloys. *Computational Materials Science*, *198*(110671), 110671. https://doi.org/10.1016/j.commatsci.2021.110671

[19] Chen, L., Tran, H., Batra, R., Kim, C., & Ramprasad, R. (2019). Machine learning models for the lattice thermal conductivity prediction of inorganic materials. *Computational Materials Science*, *170*(109155), 109155. https://doi.org/10.1016/j.commatsci.2019.109155

[20] Naghdi, A., Domínguez-Gutiérrez, F. J., Huo, W. Y., Karimi, K., & Papanikolaou, S. (2024). Dynamic nanoindentation and short-range order in equiatomic NiCoCr medium-entropy alloy lead to novel density wave ordering. *Physical Review Letters*, *132*(11), 116101. https://doi.org/10.1103/PhysRevLett.132.116101

[21] Li, Z., Yang, M., Park, J.-S., Wei, S.-H., Berry, J. J., & Zhu, K. (2016). Stabilizing perovskite structures by tuning tolerance factor: Formation of formamidinium and cesium lead iodide solid-state alloys. *Chemistry of Materials: A Publication of the American Chemical Society*, *28*(1), 284–292. https://doi.org/10.1021/acs.chemmater.5b04107

[22] Shao, Y., Gao, W., Yan, H., Li, R., Abdelwahab, I., Chi, X., Rogée, L., Zhuang, L., Fu, W., Lau, S. P., Yu, S. F., Cai, Y., Loh, K. P., & Leng, K. (2022). Unlocking surface octahedral tilt in two-dimensional Ruddlesden-Popper perovskites. *Nature Communications*, *13*(1), 138. https://doi.org/10.1038/s41467-021-27747-x

[23] Brown, I. D. (2009). Recent developments in the methods and applications of the bond valence model. *Chemical Reviews*, *109*(12), 6858–6919. https://doi.org/10.1021/cr900053k