Polymorphism - a property of the single molecule or an emergent phenomenon: a machine learning study

Itamar Wallwater[a], **Yonatan Dubi**[a], **Ari Pakman**[b], **Anat Milo**[a]

[a] *Department of Chemistry, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel* itamarwa@post.bgu.ac.il, jdubi@bgu.ac.il, anatmilo@bgu.ac.il

[a] *Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel* pakman@bgu.ac.il

* Presenting author

1. Introduction

Polymorphism, the ability of a molecule to crystallize in multiple distinct structures, significantly influences material properties in fields like pharmaceuticals [1] and materials science [2]. Understanding the crystallization of molecules, and specifically the appearance of polymorphs, is a great challenge to modern chemistry, with both fundamental and practical aspects [3, 4, 5].

Here, motivated by the proven ability of Machine-Learning (ML) algorithms to perform classification tasks, we harness ML-based tools and existing chemical datasets to ask the following question: **can the existence of polymorphs of a molecular crystal be predicted based solely on properties of the single molecule?** To that end, we have trained and tested a variety of ML binary classification models relying on the vast crystallographic data found at the Cambridge Structural Database (CSD). Based on the available data, we find that our algorithm can predict the existence of polymorphism with an average accuracy of $\sim 65\%$ at best, indicating that fundamentally, crystallization is an emergent phenomenon, and namely, a characteristic that cannot be deduced by examining the elementary constituents of the system alone [6, 7].

However, the nature of the data (and its inherent biases) and the fact that the ML algorithms tend to overestimate the number of polymorphs raises the interesting possibility, that in fact many of the molecules which are thought to have only one crystal structure, actually have more than one, and that the algorithm is, in a way, more accurate than the data itself. We invite the community to test our predictions through experiments.

2. Substantial section

Using data extracted from the CSD, we curated a dataset of organic molecular crystals, filtering out metal-organic frameworks and co-crystals. A fundamental problem arose in the categorization of molecules to mono- and poly-morphs as for a molecule to be labeled as polymorphic, it has to have (at least) two distinct structures crystallized, characterized (via, e.g., x-ray crystallography), and recorded into the database. Often, there is no apparent interest in searching for poly-morphs once a molecule has been crystallized, resulting in a big bias in the data.

To ensure a balanced dataset despite the inherent underrepresentation of polymorphic molecules (2% of the dataset), we employed Random Over-Sampling (ROS) and a Cross-Validation (CV) framework for model training and evaluation. Molecular features were derived from computational chemistry methods such as Extended Tight Binding (XTB) and Density Functional Theory (DFT), as well as cheminformatics descriptors and molecular fingerprints. We applied five supervised learning models: Logistic Regression (LG), Multi-Layer Perceptron (MLP), k-Nearest Neighbors (kNN), Random Forest (RF), and Support Vector Machine (SVM). These models were assessed using Accuracy, ROC-AUC, Specificity, and Recall, with results compared across different feature sets. Despite rigorous optimization, the best-performing models achieved a maximum test accuracy of $\sim 65\%$, only slightly better than a random classifier.

Additionally, we explored Positive-Unlabeled (PU) Learning [8], which compensates for dataset bias by considering that some monomorphic classifications might be artifacts of incomplete searches for alternative crystal forms. PU models tended to classify a larger fraction of molecules as polymorphic, aligning with independent estimates that polymorphism is more widespread than traditionally recorded ($\sim 30\%$ instead of the $\sim 6\%$ seen in CSD) [9, 10, 11].

The implication of these results is either that (i) single particle data is not enough for the classification task and therefore polymorphism is emergent, or alternatively, (ii) single data is enough for this task, and we are able to predict poly-morphs not yet found in the lab.

When models independently suggest results that differ from historical data, this divergence can serve as a catalyst for re-evaluating entrenched assumptions, driving experimental innovation. We encourage the community to leverage the data we have gathered and the models we have developed. A collaboration can take several forms like sharing experimental data that could refine and validate these models (alternative methods such as electron diffraction, exemplified by techniques like 3D electron diffrac-

tion [12] or CryoED [13], offer promising avenues for overcoming the limitations imposed by XRD), or by raising questions about whether specific molecules are polymorphic under conditions not yet explored. If the reader is engaged in molecular crystallization, please - send us your data and the molecule you are crystallizing, and we will examine if it is a mono- or polymorph. Through collective efforts, we can enhance the predictive accuracy of these models, find new crystal structures, and expand the understanding of polymorphism. For inquiries or contributions, please contact us at itamarwa@post.bgu.ac.il

2.1 Related work

1. Prediction of polymorphism using deep learning - see Michael Shatruk team work towards identifying coordination sites [14] or works like Lauren Takahashi team that aim to predict crystal structures [15]

2. Prediction of polymorphism using chemical calculation methods - in this area most of the work is centered around ranking the energy landscape of different molecular arrangments like in the work done by Alexandre Tkatchenko lab [16]

3. Experimental discovery of poly-morphs - in this field, extensive work is done to enhance the current understanding of known polymorphs and why they appear like for Ritonavir [17] and ROY [18] [5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile]

2.2 Figures and tables

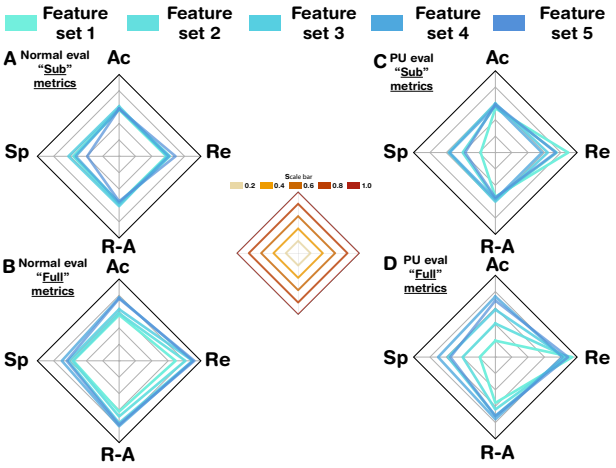

Fig. 1: Radar plots summarizing the evaluation results using MLP across the different feature sets. (A) Radar plot of "sub" metric using normal evaluation, (B) Radar plot of "Full" metric using normal evaluation, (C) Radar plot of "sub" metric using PU evaluation, (D) Radar plot of "Full" metric using PU evaluation. From these plots we can see that the choice of feature set did not change the accuracy or Roc-AUC, meaning that the overall performance was not affected. The difference between the features seems to be between specificity and recall, meaning that the models are more lenient towards classifying all molecules as non-polymorphic or polymorphic respectively. A key difference between plots A, C and B, D is the high recall seen in plots C,D, showing that using PU learning results in models that tends to classify high percentage of molecules as polymorphic

References

[1] Ashwini Nangia. Conformational polymorphism in organic crystals. *Accounts of Chemical Research*, 41:595–604, 5 2008.

[2] Matthew Boyes, Adriana Alieva, Jincheng Tong, Vaiva Nagyte, Manuel Melle-Franco, Thomas Vetter, and Cinzia Casiraghi. Exploiting the surface properties of graphene for polymorph selectivity. *ACS Nano*, 14:10394–10401, 8 2020.

[3] Gregory JO Beran. Modeling polymorphic molecular crystals with electronic structure theory. *Chemical reviews*, 116(9):5567–5613, 2016.

[4] Joel Bernstein. *Polymorphism in molecular crystals 2e*, volume 30. International Union of Crystal, 2020.

[5] Srinivasulu Aitipamula. Polymorphism in molecular crystals and cocrystals. *Advances in Organic Crystal Chemistry: Comprehensive Reviews 2015*, pages 265–298, 2015.

[6] Norman Sieroka, Tammo Lossau, and Tim Neudecker. Emergent properties in chemistry-relating molecular properties to bulk behavior. *Chemistry–A European Journal*, 30(25):e202303868, 2024.

[7] Vanessa A. Seifert. Open questions on emergence in chemistry. *COMMUNICATIONS CHEMISTRY*, 5(1), APR 7 2022.

[8] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

[9] Colin R. Groom and Frank H. Allen. The cambridge structural database in retrospect and prospect, 1 2014.

[10] Aurora J Cruz-Cabeza, Susan M Reutzel-Edens, and Joel Bernstein. Facts and fictions about polymorphism. *Chemical Society Reviews*, 44(23):8619–8635, 2015.

[11] Aurora J Cruz-Cabeza, Neil Feeder, and Roger J Davey. Open questions in organic crystal polymorphism. *Communications Chemistry*, 3(1):142, 2020.

[12] Mauro Gemmi and Enrico Mugnaioli. 3d electron diffraction: The nanocrystallography revolution. *ACS Central Science*, 5:1315–1329, 8 2019.

[13] Christopher G. Jones, Michael W. Martynowycz, Johan Hattne, Tyler J. Fulton, Brian M. Stoltz, Jose A. Rodriguez, Hosea M. Nelson, and Tamir Gonen. The cryoem method microed as a powerful tool for small molecule structure determination. *ACS Central Science*, 4:1587–1592, 11 2018.

[14] Kevin Ryan, Jeff Lengyel, and Michael Shatruk. Crystal structure prediction via deep learning. *Journal of the American Chemical Society*, 140:10158–10168, 8 2018.

[15] Keisuke Takahashi and Lauren Takahashi. Creating machine learning-driven material recipes based on crystal structure. *Journal of Physical Chemistry Letters*, 10:283–288, 1 2019.

[16] Johannes Hoja, Hsin-Yu Ko, Marcus A Neumann, Roberto Car, Robert A Distasio, and Alexandre Tkatchenko. Reliable and practical computational description of molecular crystal polymorphs, 2019.

[17] Stephan D. Parent, Pamela A. Smith, Dale K. Purcell, Daniel T. Smith, Susan J. Bogdanowich-Knipp, Ami S. Bhavsar, Larry R. Chan, Jordan M. Croom, Haley C. Bauser, Andrew McCalip, Stephen R. Byrn, and Adrian Radocea. Ritonavir form iii: A coincidental concurrent discovery. *Crystal Growth and Design*, 23:320–325, 1 2023.

[18] Manolis Vasileiadis, Andrei V. Kazantsev, Panagiotis G. Karamertzanis, Claire S. Adjiman, and Constantinos C. Pantelides. The polymorphs of roy: Application of a systematic crystal structure prediction technique. *Acta Crystallographica Section B: Structural Science*, 68:677–685, 12 2012.