# SWORD: A compositional disorder-aware crystal representation for cross-dataset curation and disordered novelty assessment

*Yuyao Huang*[1] *Wei Nong* [1] *Shuya Yamazaki* [1] *Martin Hoffmann Petersen* [2] *Kedar Hippalgaonkar* [1,2]

[1][School of Materials Science and Engineering, Nanyang Technological University, Singapore 639798]
[2][*Berkeley Education Alliance for Research in Singapore (BEARS) CREATE Tower, Singapore 138602*]
*Correspondence to: [Kedar Hippalgaonkar] kedar@ntu.edu.sg*

## 1. Introduction

In crystallography, disorder refers to deviations from an ideally ordered atomic arrangement within a periodic lattice, typically manifested as partial atomic site occupancies (e.g., substitutional mixing, vacancies). Modulating disorder can break the symmetry constraints of ordered lattices, enabling phenomena such as high-entropy stabilization, ultralow thermal conductivity, enhanced transport, and improved catalytic activity. In the Inorganic Crystal Structure Database (ICSD), nearly half of the experimentally reported crystalline materials show disorder. Yet most computational and AI-driven material pipelines remain largely disorder-agnostic, with ordered and disordered crystals represented and validated in fundamentally different ways. Recent commentaries on GNoME and MatterGen's generative model highlight that many "new" compounds may simply be symmetry-broken, artificially ordered structures that are experimentally disordered at finite temperature [1], [2]. This challenge could invalidate benchmarks of structural novelty. A key bottleneck is the lack of disorder-native structure representations for scalable indexing, comparison, and novelty assessment. Most widely used encodings are optimized for fully ordered crystals and do not treat partial occupancies as first-class information, which limits the ability to define similarity and novelty in disordered materials. This in turn hinders large-scale organization, deduplication, and annotation of disordered-structure databases and, consequently, downstream data-driven research.

To address this issue and facilitate data-driven research on disordered crystalline materials, we introduce Symmetry and Wyckoff-sequence of Ordered and Disordered crystals (SWORD), a string representation compatible with both ordered and disordered structures. Analogous to the AFLOW prototype, SWORD represents structures using Wyckoff letters, the space-group number, and composition (Fig. 1).

## 2. Related work

Despite extensive efforts to catalog and categorize unique crystal structures, challenges remain in realizing a universal classification scheme in a computationally rigorous and scalable way. Existing approaches include indexable string-based prototype or fingerprint representations, such as AFLOW prototypes [3], CLOUD [4], SLICES [5], and graph-based methods such as LeMat BAWL [6]. However, none of these representations natively encode partial occupancies, which limits their ability to describe disordered structures. A complementary line of work focuses on equivalence testing via pairwise structure matching, which determines whether two structures are the same through direct comparisons. However, the computational cost scales poorly when applied to large, database-wide comparisons. When disorder is included, tools such as MatterGen's disorder StructureMatcher [7] may require enumerating and comparing many candidates, causing the computational cost to further grow exponentially with database size and the complexity of the disordered–ordered configuration space, making such approaches impractical for large-scale cross-dataset applications.

## 3. Main results

Unlike the anonymous formula used in the AFLOW prototype, SWORD adopts a Wyckoff site-resolved, element-wise sequence representation in which partial occupancies are inherently encoded. Disordered structures that share the same disordered sites but differ in stoichiometry are further distinguished by the degree of mixing (DOM), which quantifies compositional mixing on partially occupied sites. Benchmark tests show strong robustness under atomic and lattice perturbations as well as invariance under site-translations, atom permutations, and symmetry-equivalent transformations.

Based on SWORD, database-scale deduplication can be carried out on datasets containing mixed ordered and disordered entries such

as ICSD, yielding a standardized disorder-aware cleaning workflow and a curated set of unique structures. More importantly, SWORD enables a more rigorous assessment of uniqueness and novelty by explicitly accounting for compositional disorder. Newly generated structures can be screened against both ordered and disordered crystallographic records using a mapping algorithm that links an ordered configuration to candidate disordered phases. As illustrated in Fig. 2 using the $TaCr_2O_6$ structure generated by MatterGen as an example, the candidate was initially labelled as novel but later found to belong to a known disordered phase. With SWORD, the likelihood of false novelty and uniqueness assessments is reduced without expensive pairwise matching. The deduplication of the existing databases including ICSD could reduce the data bias for training AI models, making SWORD an essential tool for data curation and evaluation.

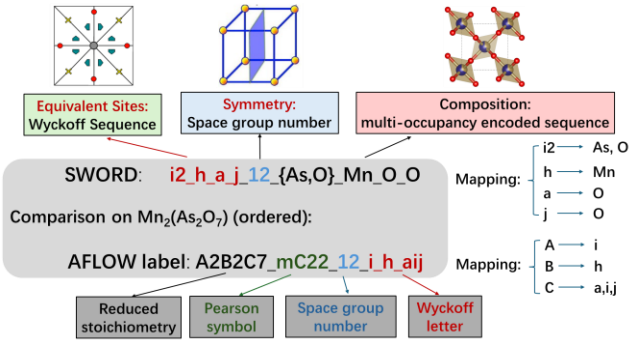


Fig. 1: Schematics of the SWORD representation compared to the AFLOW prototypes.

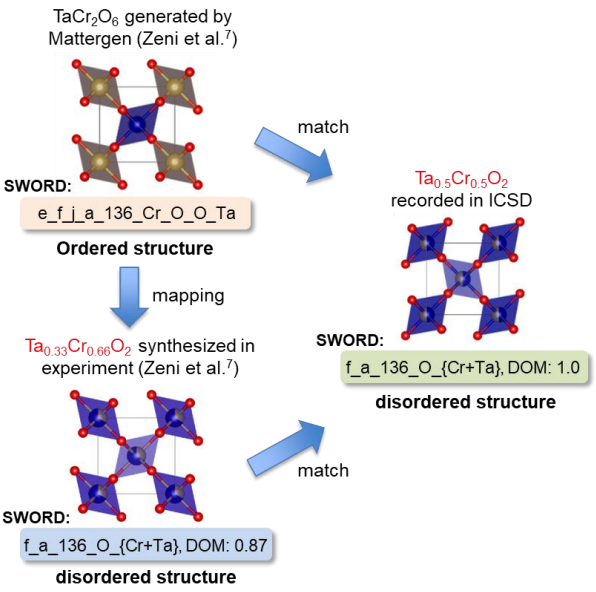


Fig. 2: Schematic of the ordered–disordered linkage enabled by SWORD

## References

[1] Anthony K. Cheetham; Ram Seshadri. *Artificial Intelligence Driving Materials Discovery? Perspective on the Article: Scaling Deep Learning for Materials Discovery*. Chemistry of Materials, 3490–3495, 2024.

[2] Mikkel Juelsholt. *Continued Challenges in High-Throughput Materials Predictions: MatterGen predicts compounds from the training dataset*. ChemRxiv, 2026.

[3] David Hicks; Cormac Toher; Denise C. Ford; Frisco Rose; Carlo De Santo; Ohad Levy; Michael J. Mehl; Stefano Curtarolo. *AFLOW-XtalFinder: a reliable choice to identify crystalline prototypes*. npj Computational Materials, 7, 30, 2021.

[4] Changwen Xu; Shang Zhu; Venkatasubramanian Viswanathan. *CLOUD: A Scalable and Physics-Informed Foundation Model for Crystal Representation Learning*. arXiv, 2025.

[5] Hang Xiao; Rong Li; Xiaoyang Shi; Yan Chen; Liangliang Zhu; Xi Chen; Lei Wang. *An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning*. Nature Communications, 14, 7027, 2023.

[6] Martin Siron; Inel Djafar; Ali Ramlaoui; Etienne du Fayette; Amandine Rossello; Edvin Fako; Matthew McDermott; Felix Therrien; Luis Barroso-Luque; Flaviu Cipcigan; Philippe Schwaller; Thomas Wolf; Alexandre Duval. *LeMat-Bulk: aggregating, and de-duplicating quantum chemistry materials databases*. arXiv, 2025.

[7] Claudio Zeni; Robert Pinsler; Daniel Zügner; Andrew Fowler; Matthew Horton; Xiang Fu; Zilong Wang; Aliaksandra Shysheya; Jonathan Crabbé; Shoko Ueda; Roberto Sordillo; Lixin Sun; Jake Smith; Bichlien Nguyen; Hannes Schulz; Sarah Lewis; Chin-Wei Huang; Ziheng Lu; Yichi Zhou; Han Yang; Hongxia Hao; Jielan Li; Chunlei Yang; Wenjie Li; Ryota Tomioka; Tian Xie. *A generative model for inorganic materials design*. Nature, 639, 624–632, 2025