Addressing data quality issues and redundancies across quantum chemistry databases for building better datasets for materials discovery: LeMat-Bulk

Martin Siron[1], Inel Djafar[1], Etienne D.L. Fayette[1], Amadine Rossello[1], Ali Ramlaoui[1], Alexandre Duval[1]
[1]Entalpic AI, Paris, France

{martin.siron, alexandre.duval}@entalpic.ai

Abstract

The rapid expansion of material science databases presents unprecedented opportunities to leverage vast volumes of quantum chemistry data. These large computational databases represent a great resource to train predictive machine learning models to make fast and accurate predictions of materials properties, as well as to train generative models to search in the combinatorial space of possible material candidates. Recent advancements in the ML community, enabled by the increased size of available datasets, have the potential to transform the discovery of novel materials with tailored properties. Initiatives like Materials Project [1, 2], OQMD [3], and Alexandria [4] have expanded the scope of computational materials science and fueled progress in the community. However, they also introduced issues of duplication, data integration, and interoperability, complicating efforts and limiting the efficiency of the machine learning community. More broadly, there are challenges related to limited high-quality data, inconsistent computational parameters, and lack of benchmarking for material novelty persist.

To address these challenges, we introduce **LeMat-Bulk**, a unified dataset combining Density Functional Theory (DFT) calculations from the Materials Project, OQMD, and Alexandria. This dataset encompasses over 5.3 million materials across three **DFT** functionals, including the largest repository of PBESol and SCAN functional calculations (~500k). Our methodology standardizes DFT calculations across databases with varying parameters, resolving inconsistencies and enhancing cross-compatibility.

Standardization involved reconciling pseudopotentials, Hubbard U corrections and spin-polarization settings, all of which are critical for consistency; to this end we excluded incompatible calculations from these datasets. We ensured uniform structural data by adopting the Optimade [5] specification and standardizing property names across databases. To augment missing charge information, we performed Bader charge calculations for over 53k materials within the Materials Project dataset. By addressing key barriers to compatibility and providing tools for data integration, LeMat-Bulk establishes a standardized foundation for leveraging large-scale materials datasets. Previous efforts observed chemical biases in the database, which were partly reduced. Rare-earth elements now tend to form more compounds with other rare earths rather than oxides, and similarly with transition metals. Such an increased compositional diversity is crucial for machine learning models, as it enhances their ability to generalize effectively by providing a more balanced and representative foundation for training.

Furthermore, one key issue in materials databases is that of data redundancy, due to duplicates which is in part due to the complexity in representing these structures. To prevent redundancy and streamline data integration, we propose a **hashing function** that generates identifiers for materials by capturing their structural and compositional properties. Material fingerprints are calculated based on the ECoN [6] bonding algorithm to construct a bonded graph structure of the most primitive unit cell and the Weisfeiler-Lehman (WL) hash. To further discriminate between different materials, we incorporate the space group number and the reduced composition in the fingerprint.

Our fingerprint approach proved effective by identifying over 340,000 duplicate structures. This was then validated by 81% of the matched structure groups with the same PBE functionals showing energy differences below 0.250 eV/atom. For structures with large energy discrepancy but matching fingerprint, DFT calculations revealed that many of these entries could be relaxed to the same structure. In comparison with existing deduplication methods, our approach demonstrates superior performance in handling symmetry and translation operations. It is also less sensitive to atomistic and lattice vector noise. It also delivers more than three orders of magnitude better computational efficiency compared to Pymatgen's StructureMatcher [7] and other structure matching and pairwise similarity-based methods. This methodology enables robust connections between databases with varying properties and calculation parameters. By matching Alexandria materials with the Materials Project database, for example, a user could connect the rich properties calculated by Materials Project to Alexandria entries. Our cross-functional analysis revealed important trends across energy, magnetization, and fermi energy.

This enables identification of unique materials within the dataset but is also valuable for generative modeling in materials science where the goal is to design novel materials with specific properties. The lack of well-benchmarked computational approaches to define novelty has posed a significant limitation to advancing generative models making challenging to quantify performance. Combined with the inability to experimentally validate many generated materials, researchers face both theoretical and practical barriers, hindering progress in material discovery. Comprehensive benchmarking under atomic noise,

lattice strain, and symmetry transformations underscores our fingerprint's efficacy in identifying duplicates and disordered materials. The benchmark highlights its robustness across a wide range of test cases compared to existing methods such as StructureMatcher, SLICES [8] and CLOUD [9].

By ensuring that each material is distinctly represented, our approach facilitates the discovery and verification of new materials, enhancing the reliability of generative workflows. This serves as a centralized resource for accessing vast and compatible data collections, facilitating advancements in generative materials and accelerating the transition to scalable AI-driven materials discovery. These advancements allow for more efficient discovery pipelines, where AI can rapidly generate, evaluate, and suggest materials with tailored properties, accelerating progress in fields like battery technology, semiconductor manufacturing, and the design of new catalysts for sustainable energy applications, such as hydrogen production or carbon dioxide capture [10].

References

[1] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, et al., Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL materials 1 (2013).

[2] A. Jain, J. Montoya, S. Dwaraknath, N. E. Zimmermann, J. Dagdelen, M. Horton, P. Huck, D. Winston, S. Cholia, S. P. Ong, et al., The materials project: Accelerating materials design through theory-driven data and tools, Handbook of Materials Modeling: Methods: Theory and Modeling , 1751 (2020).

[3] J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti, and M. A. L. Marques, Crystal graph attention networks for the prediction of stable materials, Science Advances 7 (2021).

[4] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd), Jom 65, 1501 (2013).

[5] C. W. Andersen, R. Armiento, E. Blokhin, G. J. Conduit, et al., Optimade, an api for exchanging materials data, Scientific data 8, 217 (2021).

[6] Hoppe, Effective coordination numbers (econ) and mean fictive ionic radii (mefir), Zeitschrift f¨ur Kristallographie-Crystalline Materials 150, 23 (1979).

[7] S. P. Ong, W. D. Richards, et al. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, Computational Materials Science 68, 314 (2013).

[8] C. Xu, S. Zhu, and V. Viswanathan, CLOUD: A scalable scientific foundation model for crystal representation learning, in Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges (2024).

[9] H. Xiao, R. Li, et al. An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning, Nature Communications 14, 7027 (2023).

[10] Zitnick, C. Lawrence, et al. "An introduction to electrocatalyst design using machine learning for renewable energy storage." *arXiv preprint arXiv:2010.09435* (2020).

Figures

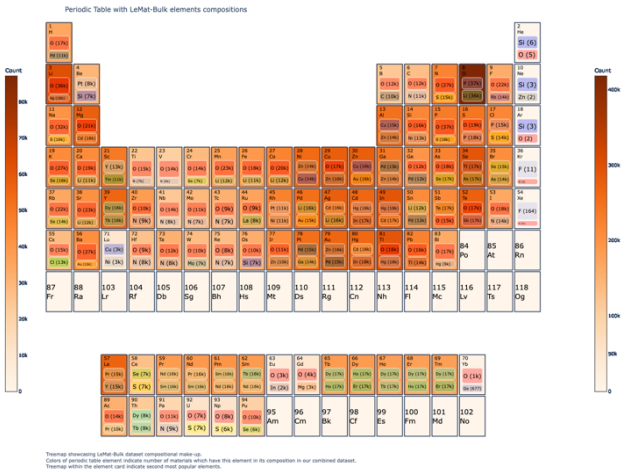

Figure 1. Periodic Table illustrating the elemental coverage of LeMat-Bulk. Highlighted elements represent those with substantial representation reflecting the diversity and depth of the dataset across chemical space.

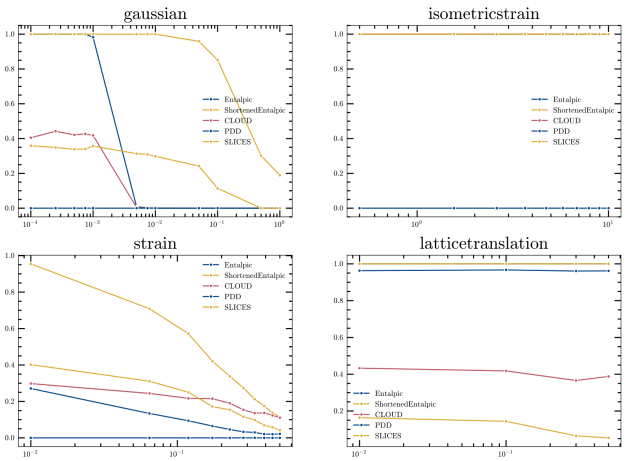

Figure 2. Average Success Rate of Hashing Method Under Random Transformations and Deformations (Including Noising, Strains, and Translations).