LEMAT-BULK: AGGREGATING, AND DE-DUPLICATING QUANTUM CHEMISTRY MATERIALS DATABASES

Martin Siron, Inel Djafar, Etienne du Fayet, Amandine Rossello,
Ali Ramlaoui, Alexandre Duval
Entalpic
Paris, France
{martin.siron,alexandre.duval}@entalpic.ai

ABSTRACT

The rapid expansion of material science databases enables the training of predictive machine learning models that deliver fast, accurate estimates of materials properties, as well as generative models that explore the vast combinatorial space of material candidates. Initiatives like the Materials Project (Jain et al., 2013; 2020), OQMD (Saal et al., 2013), and Alexandria (Schmidt et al., 2021; 2024) have greatly expanded the scope of computational materials science and fueled progress in the materials science community. However, they also introduced challenges related to duplication, data integration, and interoperability which complicates efforts to develop scalable machine learning models. 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 ($\sim$500k). Our methodology standardizes DFT calculations across databases with varying parameters, resolving inconsistencies and enhancing cross-compatibility. Besides, we propose and benchmark a hashing function (BAWL) built on Ongari et al. (2022) that generates identifiers for crystalline inorganic materials by capturing their structural and compositional properties[1].

1 INTRODUCTION

The discovery of new materials has the potential to drive major advances in battery technology, semiconductor manufacturing, and catalytic processes, to name a few (Zitnick et al., 2020). However, the chemical space has been theorized to span potentially 10^{60} (Lipinski et al., 1997) materials, thus exploring it remains a critical challenge, especially when relying on human intuition and manual lab experiments.

Developing methods capable of performing high-throughput screening of the material space is essential to speed up discovery. In this perspective, Machine Learning (ML) models have been increasingly used to approximate Density Functional Theory (DFT) computations (Kohn et al., 1996), while being orders of magnitude faster. Geometric Graph Neural Networks (GNNs) (Schütt et al., 2017; Liao et al., 2023; Duval et al., 2023), in particular, are very effective at predicting materials' properties because they can create atomic representations encoding both spatial configurations and atomic properties. However, despite being significantly cheaper than DFT, ML models are still not fast enough to exhaustively screen the vast space of candidate materials. A more effective approach may lie in generative models that can efficiently explore promising regions of this space, rather than attempt to screen it in its entirety. Progress in developing high-quality ML models—whether for property prediction or generative tasks—has been hindered by the fragmentation of available datasets.

[1]Code available at https://github.com/LeMaterial/lematerial-fetcher and https://github.com/LeMaterial/lematerial-hasher

While large-scale quantum materials databases already exist, and recent efforts have aimed to improve interoperability between them (Andersen et al., 2021; Vita et al., 2023), their practical integration remains challenging. Instead of consolidating and reusing existing results from sources, initiatives have repeated density functional theory (DFT) computations on similar or identical structures (Schmidt et al., 2021; Shoghi et al., 2023)—resulting in unnecessary duplication of computational effort. These databases are further limited by integration issues (e.g., inconsistent formats, mismatched field definitions, and incompatible calculation settings), compositional biases, and a constrained scope (Sommer et al., 2025). Compounding these problems is a lack of standardized identifiers linking equivalent or related materials across databases. This fragmented and inefficient landscape impedes researchers in AI4Science and materials informatics from fully capitalizing on the available data (Hegde et al., 2023).

To address some of these challenges, we introduce *LeMat-Bulk*, a unified materials dataset based on the Materials Project (Jain et al., 2013; 2020), OQMD (Saal et al., 2013) and Alexandria (Schmidt et al., 2021; 2024) databases while ensuring compatibility of DFT parameters (*e.g.*, pseudo-potentials, Hubbard U parameters, spin polarization, DFT functional) and removing inconsistent data points. LeMat-Bulk is a dataset built with consistent property names across materials ensuring compatibility between entries. It acts as a unified and well formatted way for researchers to train large foundation models and to explore the chemical space with higher resolution. We also identify and remove duplicates, a crucial step to mitigate database redundancy, which can introduce biases and inefficiencies. To do so, we propose a modification of a hashing method by Ongari et al. (2022), which we name *BAWL* (Bonding Algorithm Weisfeiler-Lehman) to generate a unique identifier for each material structure, identifying over 340k duplicate structures. We further validate 81% of these duplicates-computed with the same DFT functional-have an energy difference below 0.25 eV/atom, which is common to estimate metastability (Aykol et al., 2018). We benchmark our hashing approach against alternative de-duplication methods to demonstrate its effectiveness on random perturbations and disordered structure identification.

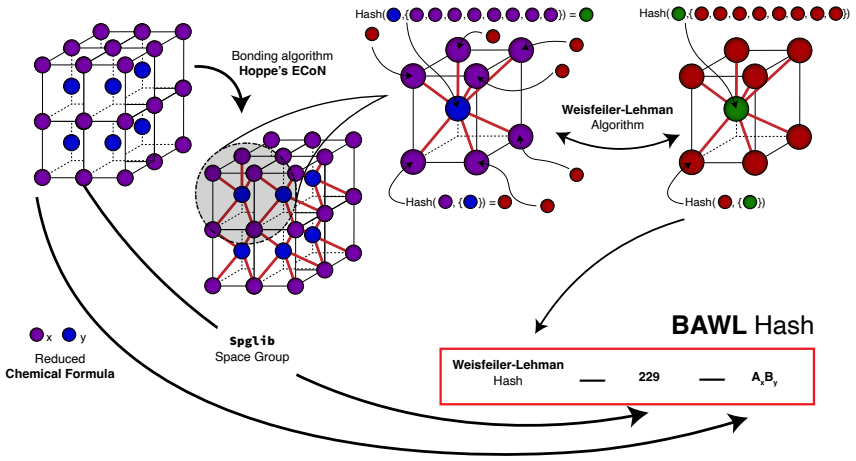


Figure 1: Illustration of our BAWL hashing method.

2 Related Work

The design of generative models for inorganic materials increasingly highlights the importance of well-defined measures for novelty and diversity of the candidates generated (Merchant et al., 2023). Currently, various methods exist for detecting duplicate materials. Pymatgen's StructureMatcher (Ong et al., 2013) compares structures by normalizing them, checking if the lattices can be transformed into each other, and comparing the sites by sorting them and permuting them to allow for optimal matching. Although it is widely used in the materials science community, it struggles with disordered structures and scales quadratically with the number of materials to compare (without any composition-based filtering). Other approaches for comparing crystals include, but are not limited to, SLICES (Xiao et al., 2023) which uses a graph-based approach to encode the geometric configuration, and CLOUD (Xu et al., 2024) which uses a clustering algorithm. Vectorizing structures

such as PDD (Widdowson & Kurlin, 2022) or using GNNs to compare structures (Yang et al., 2022) are also common approaches. However, benchmarks of how these methods perform against affine transformations which respect symmetry and noise lack, limiting their general applicability.

3 METHOD

We introduce LeMat-Bulk, a dataset which encompasses over 5.3 million materials across three DFT functionals, including the largest repository of PBESol and SCAN functional calculations ($\sim$500k).

Dataset. To ensure a consistent and high-quality dataset for machine learning applications, we standardize functionals, pseudopotentials, Hubbard U corrections, and spin-polarization settings across all entries. These parameters play a critical role in first-principles calculations and greatly affect thermodynamic and other material properties computed. Calculations that did not meet these unified criteria were excluded to remove incompatibility in these parameters. We then harmonize structural data using the OPTIMADE specification (Andersen et al., 2021) and standardize property names across the various databases. To ensure consistency and completeness across datasets, we also compute Bader charges for over 53,000 materials in the Materials Project, thereby adding charge information to a level comparable with that available in OQMD and Alexandria. Notably, combining all these datasets, reduces biases of any single database.

Hasher. To systematically identify and remove duplicate materials, we introduce a hashing procedure that generates unique identifiers (fingerprints) for each structure built on top of Ongari et al. (2022). First, the ECoN (Effective Coordination Number) bonding algorithm (Hoppe, 1979) constructs a bonded graph of the material's most primitive unit cell including the species encoded in the node. We then apply the Weisfeiler-Lehman (WL) graph isomorphism hash (Shervashidze et al., 2011) to capture each graph's structural features. To further distinguish among materials with similar topologies, we incorporate both the space group number-identified using `Spglib` (Togo et al., 2018)—and the reduced composition into the final fingerprint. By concatenating these elements with the WL hash, we obtain a comprehensive identifier that is used to flag duplicates. In Section 4, we compare the performance of this full fingerprint—and a shortened version (Short-BAWL) which omits the space group number—against other existing methods. We illustrate this process in Figure 1. The full BAWL hash is used to deduplicate the dataset.

4 EXPERIMENTS

In this section, we evaluate the effectiveness of our hashing function on the LeMat-Bulk dataset and compare its performance against several established methods, including Pymatgen's StructureMatcher (Ong et al., 2013), SLICES (Xiao et al., 2023), and CLOUD (Xu et al., 2024). In addition to these conventional approaches, we explore GNN-based similarity metrics, called EqV2-sim, for detecting and confirming duplicate structures by leveraging an EquiformerV2 model (Liao et al., 2023) trained on OMAT24 (Barroso-Luque et al., 2024) and computing cosine similarity metrics between embedding pairs. More details are provided in the Appendix A.2. Our analysis focuses on three main aspects: robustness to structural perturbations, robustness under symmetry and translation operations, and handling disordered structures. Computational efficiency is reported in Appendix A.5.

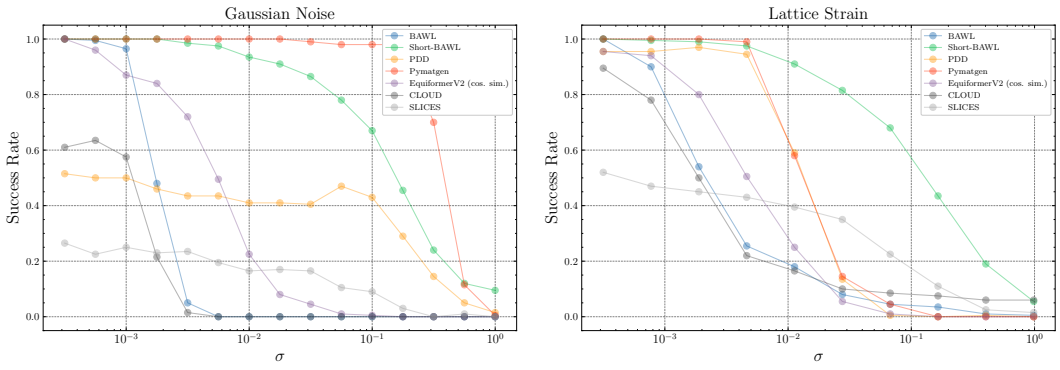
Robustness to Structural Perturbations. We first investigate how these algorithms react to small distortions in atomic positions and lattice parameters. 100 structures were randomly selected from the LeMat-Bulk dataset and transformed. The success rate of each method in matching the original structure with its perturbed counterpart is shown in Figure 2. Different methods exhibit varying sensitivity to Gaussian noise on fractional coordinates, with BAWL and Pymatgen maintaining high identification rates at larger perturbations with StructureMatcher being less sensitive, while others show increased sensitivity to lower levels of noise. Similarly, under lattice vector noise (Figure 2b), we find that Short-BAWL has the least sensitivity, even under high noise to lattice vectors. Pymatgen falls between both BAWL-based methods in terms of sensitivity. The discrepancies demonstrated between BAWL and Short-BAWL are due the symmetry identification in SPGLIB.

Robustness to Lattice Translation and Symmetry Operations A fingerprint method should be invariant under symmetry operations of the lattice and translations. In our comparison of fingerprint method we found that both BAWL and Short-BAWL respected all translation and symmetry operations. SLICES had a lower success rate for smaller noise. All of the hashing method different from BAWL (*i.e.* CLOUD, SLICES and PDD) are way less robust to site translations (Appendix A.3).

Disordered Structures. Disordered or partially occupied sites are often encountered in real-world materials. Our results (Table 1 detailed in Appendix A.4) highlight that while some methods, such as PDD, maintain moderate accuracy in recognizing similar disordered structures, other approaches like Pymatgen and SLICES fail to generalize to these cases. Short-BAWL shows improved performance over its full version, due to symmetry-detection issues from `Spglib`.

	BAWL	**Short-BAWL**	**EqV2-sim**	**PDD**	**Pymatgen**	**CLOUD**	**SLICES**
Disordered Structure	0.14 ± 0.18	0.30 ± 0.33	0.61 ± 0.38	0.56 ± 0.46	0.00 ± 0.01	0.46 ± 0.33	0.00 ± 0.00

Table 1: Comparison of different methods matching pairs of disordered structures. The success rate is reported on all the combinations of pairwise similarity for different chemical formulas. We then aggregate these results over all chemical formulas and report the mean success rate and standard deviation in the second row (*Disordered Structures*). We also tested the success rate for correctly disctriminating between random pairs of chemically different materials which all models succeeded in. More details in Table 3.



(a) Atomic Positions

(b) Lattice Vectors Strain

Figure 2: Success rate of structure identification methods under different different perturbations. On the left, we show the performance of each method under Gaussian noise added to the atomic positions. The noise is sampled *per atom* from a normal distribution with a mean of 0 and a standard deviation σ. On the right, we show the performance under lattice strain from a Gaussian distribution with a mean of 0 and a standard deviation σ applied independently to each lattice vector.

5 Conclusion

The unification of materials science datasets is crucial for advancing AI-driven materials discovery. LeMat-Bulk represents a step toward this goal by integrating large-scale DFT databases while addressing duplication and parameter inconsistencies issues. The applications for such a dataset include—but are not limited to—computing more accurate energies above the convex hull by generating more reliable phase diagrams, and providing a larger coherent source of data for training ML and generative models.

The proposed full BAWL hashing method offers an efficient solution for identifying duplicate structures, used to significantly improve data integrity across different sources in *LeMat-Bulk*. Using the full BAWL fingerprint, we created a separate *LeMat-Bulk Unique* with the lowest energy material among duplicates. For generative models, such a hashing function can be relevant to define improved novelty and diversity metrics.

Future work should explore extending this approach to additional datasets, by including other types of materials data such as trajectories or specialized datasets from catalysis for example. Refining hashing techniques for enhanced accuracy, and further benchmarking against alternative deduplication strategies to create a standardized benchmark are necessary steps for improving generative models and property prediction in materials science.

ACKNOWLEDGMENTS

We acknowledge HuggingFace and their team including Thomas Wolff for providing compute for calculating Bader charges and hosting of the database, and Leandro Von Werra for calculating Bader charges. We acknowledge fruitful discussions from Zack Ulissi, and Matt Horton.

REFERENCES

Casper W Andersen, Rickard Armiento, Evgeny Blokhin, Gareth J Conduit, Shyam Dwaraknath, Matthew L Evans, Ádám Fekete, Abhijith Gopakumar, Saulius Gražulis, Andrius Merkys, et al. Optimade, an api for exchanging materials data. *Scientific data*, 8(1):217, 2021.

Muratahan Aykol, Shyam S Dwaraknath, Wenhao Sun, and Kristin A Persson. Thermodynamic limit for synthesis of metastable inorganic materials. *Science advances*, 4(4):eaaq0148, 2018.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C. Lawrence Zitnick, and Zachary W. Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint arXiv: 2410.12771*, 2024.

Alexandre Duval, Simon V. Mathis, Chaitanya K. Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D. Malliaros, Taco Cohen, Pietro Liò, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems. *arXiv preprint arXiv: 2312.07511*, 2023.

Vinay I. Hegde, Christopher K. H. Borg, Zachary del Rosario, Yoolhee Kim, Maxwell Hutchinson, Erin Antono, Julia Ling, Paul Saxe, James E. Saal, and Bryce Meredig. Quantifying uncertainty in high-throughput density functional theory: A comparison of aflow, materials project, and oqmd. *Phys. Rev. Mater.*, 7:053805, May 2023. doi: 10.1103/PhysRevMaterials.7.053805. URL `https://link.aps.org/doi/10.1103/PhysRevMaterials.7.053805`.

Rudolf Hoppe. Effective coordination numbers (econ) and mean fictive ionic radii (mefir). *Zeitschrift für Kristallographie-Crystalline Materials*, 150(1-4):23–52, 1979.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Anubhav Jain, Joseph Montoya, Shyam Dwaraknath, Nils ER Zimmermann, John Dagdelen, Matthew Horton, Patrick Huck, Donny Winston, Shreyas Cholia, Shyue Ping Ong, et al. The materials project: Accelerating materials design through theory-driven data and tools. *Handbook of Materials Modeling: Methods: Theory and Modeling*, pp. 1751–1784, 2020.

Walter Kohn, Axel D Becke, and Robert G Parr. Density functional theory of electronic structure. *The Journal of Physical Chemistry*, 100(31):12974–12980, 1996.

Yidong Liao, Brandon Wood, Abhishek Das, and T. Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2306.12059.

Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1):3–25, 1997. ISSN 0169-409X. doi: https://doi.org/10.1016/S0169-409X(96)00423-1. In Vitro Models for Selection of Development Candidates.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Kirill Okhotnikov, Thibault Charpentier, and Sylvian Cadars. Supercell program: a combinatorial structure-generation approach for the local-level modeling of atomic substitutions and partial occupancies in crystals. *Journal of cheminformatics*, 8:1–15, 2016.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, February 2013. ISSN 09270256. doi: 10.1016/j.commatsci.2012.10.028. URL http://linkinghub.elsevier.com/retrieve/pii/S0927025612006295.

Daniele Ongari, Leopold Talirz, Kevin Maik Jablonka, Daniel W Siderius, and Berend Smit. Data-driven matching of experimental crystal structures and gas adsorption isotherms of metal–organic frameworks. *Journal of Chemical & Engineering Data*, 67(7):1743–1756, 2022.

James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65:1501–1509, 2013.

Jonathan Schmidt, Love Pettersson, Claudio Verdozzi, Silvana Botti, and Miguel A. L. Marques. Crystal graph attention networks for the prediction of stable materials. *Science Advances*, 7(49): eabi7948, 2021. doi: 10.1126/sciadv.abi7948. URL https://www.science.org/doi/abs/10.1126/sciadv.abi7948.

Jonathan Schmidt, Tiago FT Cerqueira, Aldo H Romero, Antoine Loew, Fabian Jäger, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, 48:101560, 2024.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, A. Tkatchenko, and K. Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Neural Information Processing Systems*, 2017.

Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

Nima Shoghi, Adeesh Kolluru, John R. Kitchin, Zachary W. Ulissi, C. L. Zitnick, and Brandon M. Wood. From molecules to materials: Pre-training large generalizable models for atomic property prediction. *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2310.16802.

Timo Sommer, Cian Clarke, and Max García-Melchor. Beyond chemical structures: lessons and guiding principles for the next generation of molecular databases. *Chemical Science*, 16(3):1002–1016, 2025.

Atsushi Togo, Kohei Shinohara, and Isao Tanaka. Spglib: a software library for crystal symmetry search. *arXiv preprint arXiv: 1808.01590*, 2018.

Joshua A Vita, Eric G Fuemmeler, Amit Gupta, Gregory P Wolfe, Alexander Quanming Tao, Ryan S Elliott, Stefano Martiniani, and Ellad B Tadmor. Colabfit exchange: Open-access datasets for data-driven interatomic potentials. *The Journal of Chemical Physics*, 159(15), 2023.

Daniel Widdowson and Vitaliy Kurlin. Resolving the data ambiguity for periodic crystals. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=4wrB7Mo9_OQ.

Hang Xiao, Rong Li, Xiaoyang Shi, Yan Chen, Liangliang Zhu, Xi Chen, and Lei Wang. An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning. *Nature Communications*, 14(1):7027, 2023.

Changwen Xu, Shang Zhu, and Venkatasubramanian Viswanathan. CLOUD: A scalable scientific foundation model for crystal representation learning. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024. URL https://openreview.net/forum?id=geZ5LQOCSj.

Yilin Yang, Mingjie Liu, and John R. Kitchin. Neural network embeddings based similarity search method for atomistic systems. *Digital Discovery*, 1:636–644, 2022. doi: 10.1039/D2DD00055E. URL http://dx.doi.org/10.1039/D2DD00055E.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Ryota Tomioka, and Tian Xie. Mattergen: a generative model for inorganic materials design, 2024. URL https://arxiv.org/abs/2312.03687.

C. Lawrence Zitnick, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Thibaut Lavril, Aini Palizhati, Morgane Riviere, Muhammed Shuaibi, Anuroop Sriram, Kevin Tran, Brandon Wood, Junwoong Yoon, Devi Parikh, and Zachary Ulissi. An introduction to electrocatalyst design using machine learning for renewable energy storage. *arXiv preprint arXiv: 2010.09435*, 2020.