A APPENDIX

A.1 DATASET COVERAGE

Functional	Rows	Unique BAWL rows	Alexandria	Materials Project	OQMD
PBE	5,335,298	5,005,017	4,628,422	138,931	567,945
PBESol	447,824	432,633	415,419	32,405	0
SCAN	422,840	417,666	415,419	7,421	0
Non Compatible	519,627	-	0	10,646	508,981

Table 2: Summary of LeMat-Bulk's functional distribution. *Unique BAWL rows* reports the number of unique structures found using the BAWL hash. The number of structures from the source database for every functional is also reported. *Non Compatible* rows use a non compatible pseudopotential or Hubbard U but which can still be leveraged in specific Machine Learning workflows that don't consider interoperatbility.

We provide a detailed breakdown of the materials included in LeMat-Bulk, highlighting the coverage and the number of materials. The number of rows is reported in Table 2, along with the unique rows as identified by the BAWL fingerprint. Figure 3c shows the distribution of materials across the Materials Project database. This periodic table show the two elements that every other element is more likely to form a compound with. Figure 3d shows the distribution of materials across the LeMat-Bulk database. In the Materials Project, we observe that oxygen is the most common element in the database, forming compounds with a wide range of elements. In contrast, the LeMat-Bulk database shows a more balanced distribution of materials across the periodic table, with a higher proportion for almost all elements.

A.2 HASHING FUNCTIONS PARAMETERS

We detail the different configurations used to assert the equivalence of two structures using the different algorithms presented in section 4.

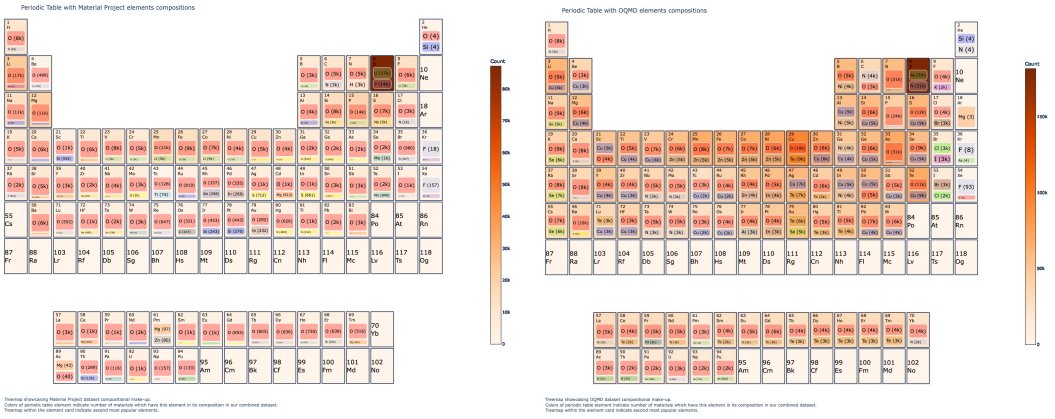
Hashers. For BAWL, CLOUD and SLICES, which return a string (hash) for a given input structure, we consider two structures to be equivalent if their hashes are identical.

PDD. For PDD, we consider two structures to be equivalent if the euclidean distance between their embeddings is below a certain threshold.

EqV2-sim. For EquiformerV2, we consider two structures to be equivalent if the cosine similarity between their embeddings is above a certain threshold. This threshold was hand-picked to maximize the number of structures matched by both BAWL and EquiformerV2 while minimizing the number of false positives. Experiments were conducted using a Euclidean distance instead of cosine similarity, but the embeddings length were highly correlated with the number of atoms in the structure, which made the distance metric less reliable. The embeddings were extracted at the entry of the energy prediction head of the model trained on OMAT24 ($l = 0$ of the spherical harmonics) and summed over all atoms in the structure. Experiments using the mean of the embeddings instead of the sum or using an untrained architecture of model were also conducted, showing poor performance in this perturbation benchmark.

A.3 STRUCTURAL PERTURBATIONS

In addition to Gaussian perturbations of the atomic positions and the lattice, we also investigate the performance of the different methods under isometric strain, translation and space group symmetry operations. Isometric strain consists of applying a uniform strain to the lattice vectors of the structure, while translation consists of moving the structure along a random direction. We show the results of these experiments in Figure 4. Unsurprisingly, a lot of methods succeed perfectly in identifying the same structure under translations because they are invariant to this operation. However, some methods like CLOUD, SLICES and PDD struggle with translations. For isometric strain,

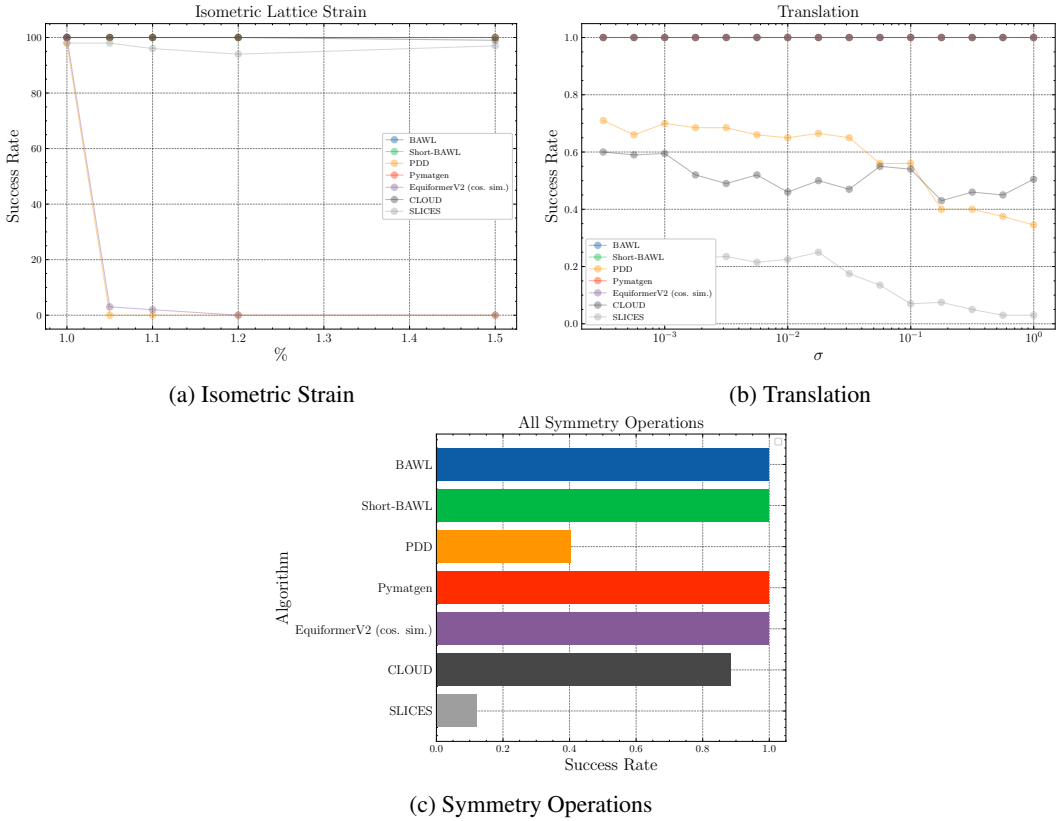


(a) Materials Project (b) OQMD

(c) Alexandria (d) LeMat-Bulk

Figure 3: Distribution of materials across the periodic table in the Materials Project, OQMD and Alexandria and *LeMat-Bulk* databases.

BAWL, Pymatgen and CLOUD perform perfectly well too. However, it is worth noting that vector based methods like EquiformerV2 and PDD fail to generalize to this perturbation. While better threshold tuning could potentially improve their performance, it is clear that they are not robust to this type of perturbation.



(a) Isometric Strain

(b) Translation

(c) Symmetry Operations

Figure 4: Success rate of structure identification methods under more perturbations than Figure 2.

A.4 Disordered Structures

The benchmark for disordered structures is computed by generating a dataset from a number of materials with the Supercell software package (Okhotnikov et al., 2016). For every material, we generate a large number of disordered structures by randomly substituting partial occupancies to atomic positions in the bulk structure based on symmetry constraints. The materials are curated from the Supercell study itself with various disorder types from materials ($Sn_{0.5}Pb_{0.5}Te$, $MnCaCO_3$, $Ti_{1-x}Mg_xN$), proton disorder in Ice Ih, and cationic site disorder ($Ca_2Al_2SiO_7$, CZTSe, $FeSbO_4$, $MgAlFeO_4$, PZT, Rb-PST-1, $SrSiAlO_x$). We then compute, for every chosen material, the success rate for identifying that two random disordered structures of the same material are indeed identified to be the same by the structure checker. To ensure that the test doesn't just set all observed pairs to be equivalent, we also compute the success rate of correctly flagging two different molecules as being different. The full results are reported in Table 3.

A.5 Computational Efficiency

Figure 5 shows the computational efficiency of each method as the number of structures increases. We also extrapolate the compute time to the amount of time it would take to compare the entire *LeMat-Bulk* against itself, based on the observed run times of our experiments. While exact graph-based methods (e.g., StructureMatcher) are initially fast, they scale poorly with the number of structure (at least quadratically). The runtime for BAWL has a bigger slope than other algorithms such

Composition	BAWL	Short-BAWL	PDD	Pymatgen	EqV2-sim	CLOUD	SLICES
$Ca_8Al_8Si_4O_{28}$	0.04 ± 0.00	0.18 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.00 ± 0.00	0.87 ± 0.00	0.04 ± 0.00	0.00 ± 0.00
$Fe_2Cu_4Sn_2Se_5S_3$	0.10 ± 0.00	0.30 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.00 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.40 ± 0.00	0.00 ± 0.00
$Fe_4Sb_4O_{16}$	0.05 ± 0.00	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	0.00 ± 0.00	0.48 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
H_8O_4	0.42 ± 0.15	0.70 ± 0.17	0.00 ± 0.00	0.00 ± 0.00	0.04 ± 0.01	$\mathbf{0.54 \pm 0.15}$	0.00 ± 0.00
$Mg_8Al_8Fe_8O_{32}$	0.00 ± 0.00	0.00 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.00 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.30 ± 0.06	0.00 ± 0.00
$Sn_4Te_8Pb_4$	0.00 ± 0.00	0.18 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.04 ± 0.00	0.57 ± 0.00	0.04 ± 0.00	0.00 ± 0.00
$Sr_5Al_{10}Si_6O_{32}$	0.09 ± 0.01	0.09 ± 0.01	0.25 ± 0.02	0.00 ± 0.00	0.63 ± 0.07	$\mathbf{0.90 \pm 0.09}$	0.00 ± 0.00
$Zr_4Ti_4Pb_8O_{24}$	0.09 ± 0.00	0.27 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.02 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.02 ± 0.00	0.00 ± 0.00

Table 3: Detailed comparison of different methods for matching pairs of disordered structures per chemical formula. The success rate is reported on all the combinations of pairwise similarity for every chemical formula. We report the average success rate and standard deviation over 5 runs (of 40 structures) when a given composition has more than 40 structures.

as CLOUD or SLICES but is still below EqV2-sim on the hardware used to run these experiments. Hashing-based functions scale linearly for pairwise comparisons in the number of structures, making them all viable for large-scale materials databases. This balance between efficiency and accuracy enables rapid screening in real-world materials discovery applications.

B DFT PARAMETERS

We detail in Table 4 and Table 5 the DFT parameters used in *LeMat-Bulk* along with the original dataset amongst Alexandria, Materials Project and OQMD where they are extracted from.

Table 4: Pseudopotentials adopted in LeMat-Bulk

Element	LeMat-Bulk	Alexandria	Materials Project	OQMD
Ac	PAW_PBE Ac 06Sep2000	✓	✓	✓
Ag	PAW_PBE Ag 06Sep2000	✓	✓	✓
Al	PAW_PBE Al 04Jan2001	✓	✓	✓
Ar	PAW_PBE Ar 07Sep2000	✓	✓	✓
As	PAW_PBE As 06Sep2000	✓	✓	✓
At	PAW_PBE At_d	× (not in DB)	× (not in DB)	✓
Au	PAW_PBE Au 06Sep2000	✓	✓	✓
Ba	PAW_PBE Ba_sv 06Sep2000	✓	✓	✓
Be	PAW_PBE Be_sv 06Sep2000	✓	✓	✓
Bi	PAW_PBE Bi 08Apr2002	✓	✓	✓
B	PAW_PBE B 06Sep2000	✓	✓	✓
Br	PAW_PBE Br 06Sep2000	✓	✓	✓
Ca	PAW_PBE Ca_sv 06Sep2000	✓	✓	× (Ca_pv)
Cd	PAW_PBE Cd 06Sep2000	✓	✓	✓
Ce	PAW_PBE Ce 28Sep2000	✓	✓	× (Ce_3)
Cl	PAW_PBE Cl 17Jan2003	✓	✓	✓
Co	PAW_PBE Co 06Sep2000	✓	✓	✓
C	PAW_PBE C 08Apr2002	✓	✓	✓
Cr	PAW_PBE Cr_pv 07Sep2000	✓	✓	× (Ce_3)
Cs	PAW_PBE Cs_sv 08Apr2002	✓	✓	✓
Cu	PAW_PBE Cu_pv 06Sep2000	✓	✓	✓
Dy	PAW_PBE Dy_3 06Sep2000	✓	✓	✓
Er	PAW_PBE Er_3 06Sep2000	✓	✓	✓
Eu	PAW_PBE Eu 08Apr2002	✓	✓	× (Eu_2)
Fe	PAW_PBE Fe_pv 06Sep2000	✓	✓	✓
F	PAW_PBE F 08Apr2002	✓	✓	✓
Ga	PAW_PBE Ga_d 06Sep2000	✓	✓	✓
Gd	PAW_PBE Gd 08Apr2002	✓	✓	✓
Ge	PAW_PBE Ge_d 06Sep2000	✓	✓	✓

Continued on next page

Table 4 – *Continued from previous page*

Elements	LeMat-Bulk	Alexandria	Materials Project	OQMD
He	PAW_PBE He 05Jan2001	✓	✓	✓
Hf	PAW_PBE Hf_pv 06Sep2000	✓	✓	✓
Hg	PAW_PBE Hg 06Sep2000	✓	✓	✓
Ho	PAW_PBE Ho_3 06Sep2000	✓	✓	✓
H	PAW_PBE H 15Jun2001	✓	✓	✓
In	PAW_PBE In_d 06Sep2000	✓	✓	✓
I	PAW_PBE I 08Apr2002	✓	✓	✓
Ir	PAW_PBE Ir 06Sep2000	✓	✓	✓
K	PAW_PBE K_sv 06Sep2000	✓	✓	✓
Kr	PAW_PBE Kr 07Sep2000	✓	✓	✓
La	PAW_PBE La 06Sep2000	✓	✓	✓
Li	PAW_PBE Li_sv 23Jan2001	✓	✓	✓
Lu	PAW_PBE Lu_3 06Sep2000	✓	✓	✓
Mg	PAW_PBE Mg_pv 06Sep2000	✓	✓	✓
Mn	PAW_PBE Mn_pv 07Sep2000	✓	✓	× (Mn)
Mo	PAW_PBE Mo_pv 08Apr2002	✓	✓	✓
Na	PAW_PBE Na_pv 05Jan2001	✓	✓	✓
Nb	PAW_PBE Nb_pv 08Apr2002	✓	✓	✓
Nd	PAW_PBE Nd_3 06Sep2000	✓	✓	✓
Ne	PAW_PBE Ne 05Jan2001	✓	✓	✓
Ni	PAW_PBE Ni_pv 06Sep2000	✓	✓	✓
N	PAW_PBE N 08Apr2002	✓	✓	✓
Np	PAW_PBE Np 06Sep2000	✓	✓	✓
O	PAW_PBE O 08Apr2002	✓	✓	✓
Os	PAW_PBE Os_pv 20Jan2003	✓	✓	✓
Pa	PAW_PBE Pa 07Sep2000	✓	✓	✓
Pb	PAW_PBE Pb_d 06Sep2000	✓	✓	✓
Pd	PAW_PBE Pd 05Jan2001	✓	✓	✓
Pm	PAW_PBE Pm_3 07Sep2000	✓	✓	✓
P	PAW_PBE P 17Jan2003	✓	✓	✓
Po	PAW_PBE Po	× (not in DB)	× (not in DB)	✓
Pr	PAW_PBE Pr_3 07Sep2000	✓	✓	✓
Pt	PAW_PBE Pt 05Jan2001	✓	✓	✓
Pu	PAW_PBE Pu 06Sep2000	✓	✓	✓
Rb	PAW_PBE Rb_sv 06Sep2000	✓	✓	✓
Re	PAW_PBE Re_pv 06Sep2000	✓	✓	✓
Rh	PAW_PBE Rh_pv 06Sep2000	✓	✓	× (Rh)
Ru	PAW_PBE Ru_pv 06Sep2000	✓	✓	× (Ru)
Sb	PAW_PBE Sb 06Sep2000	✓	✓	✓
Sc	PAW_PBE Sc_sv 07Sep2000	✓	✓	✓
Se	PAW_PBE Se 06Sep2000	✓	✓	✓
Si	PAW_PBE Si 05Jan2001	✓	✓	✓
Sm	PAW_PBE Sm_3 07Sep2000	✓	✓	✓
Sn	PAW_PBE Sn_d 06Sep2000	✓	✓	✓
S	PAW_PBE S 17Jan2003	✓	✓	✓
Sr	PAW_PBE Sr_sv 07Sep2000	✓	✓	✓
Ta	PAW_PBE Ta_pv 07Sep2000	✓	✓	✓
Tb	PAW_PBE Tb_3 06Sep2000	✓	✓	✓
Tc	PAW_PBE Tc_pv 06Sep2000	✓	✓	✓
Te	PAW_PBE Te 08Apr2002	✓	✓	✓
Th	PAW_PBE Th 07Sep2000	✓	✓	✓
Ti	PAW_PBE Ti_pv 07Sep2000	✓	✓	× (Ti)
Tl	PAW_PBE Tl_d 06Sep2000	✓	✓	✓
Tm	PAW_PBE Tm_3 20Jan2003	✓	✓	✓
U	PAW_PBE U 06Sep2000	✓	✓	✓

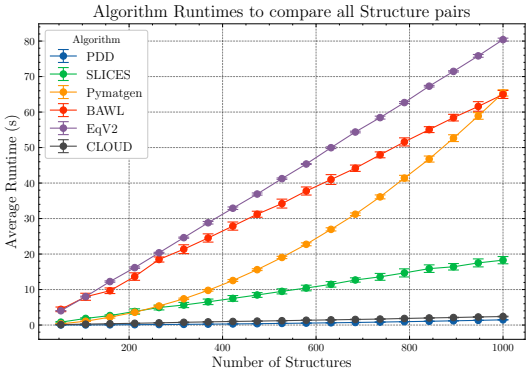
Continued on next page

Table 4 – *Continued from previous page*

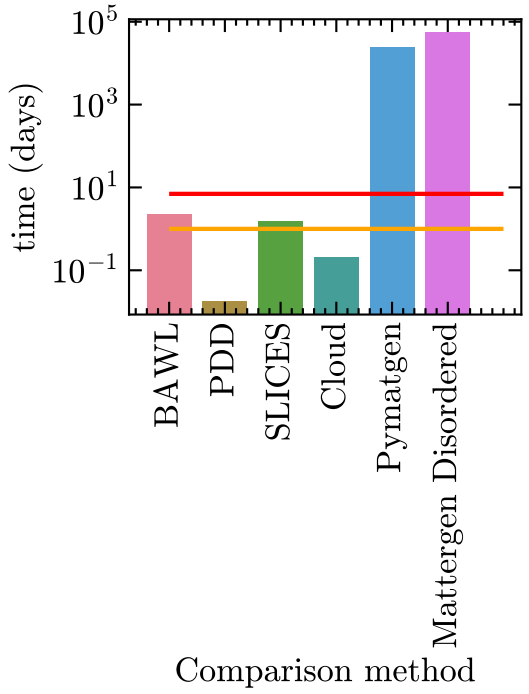
Elements	LeMat-Bulk	Alexandria	Materials Project	OQMD
V	PAW_PBE V_sv 07Sep2000	✓	× (V_pv)	× (V)
W	PAW_PBE W_pv 06Sep2000	✓	✓	✓
Xe	PAW_PBE Xe 07Sep2000	✓	✓	✓
Yb	PAW_PBE Yb 24Feb2003	✓	× (Yb_3)	× (Yb_2)
Y	PAW_PBE Y_sv 06Sep2000	✓	✓	✓
Zn	PAW_PBE Zn 06Sep2000	✓	✓	✓
Zr	PAW_PBE Zr_sv 07Sep2000	✓	✓	✓

Element	**System**	**Hubbard U (eV)**
Co	Oxides, Fluorides	3.32
Cr	Oxides, Fluorides	3.7
Fe	Oxides, Fluorides	5.3
Mn	Oxides, Fluorides	3.9
Mo	Oxides, Fluorides	4.38
Ni	Oxides, Fluorides	6.2
V	Oxides, Fluorides	3.25
W	Oxides, Fluorides	6.2

Table 5: Hubbard-U values adopted in LeMat-Bulk



(a) Runtimes across 5 seeds



(b) Extrapolated

Figure 5: Time taken to compare all combinations from a list of n structures picked randomly from *LeMat-Bulk*. Error bars show the time complexity measured over 5 different seeds when extracting the structures from the dataset. Extrapolated time in 5b is obtained by fitting a polynomial regression based on 5a and estimating the time it would take to compare all the pairwise materials in the database. No algorithm was explicitly parallelized (outside of existing implementations). An AMD Ryzen 5600G CPU was used for algorithms running on the CPU and an RTX 3070 for EqV2-sim. Mattergen Disordered is the Structure matcher implemented in (Zeni et al., 2024).