Supplementary Information

Space-Group-Guided Template Retrieval for Composition-to-Structure Prediction

Yen-Ju Wu, Yibin Xu

[1] Data-driven Inorganic Materials Group, Center for Basic Research on Materials, National Institute for Materials Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

*Corresponding author: wu.yenju@nims.go.jp

This Supplementary Information reports the AtomWork-Adv. (AWA) evaluation of the main-text Materials Project (MP) benchmark, the complete retrieval comparison using the top 10, 3, and 1 predicted space groups, the dataset composition distributions, and the chemical-role assignments used by the substitution engine. Abbreviations follow the main text: SG, space group; LOEO, leave-one-entry-out; SOAP, smooth overlap of atomic positions; RMSD, root-mean-square deviation.

S1. Cross-database evaluation on AtomWork-Adv. (AWA)

The leave-one-entry-out (LOEO) protocol described in the main text was repeated on AWA using the same downstream settings: 500 target entries sampled with random_state = 42, up to 50 ranked structure templates considered per target, substitution attempted in rank order until the first feasible mapping, and MatterSim v1.0.0 relaxation. The AWA SG classifier used the same XGBoost hyperparameters and five-fold stratified cross-validation protocol as the MP classifier.

Table S1. AWA leave-one-entry-out benchmark corresponding to main-text Table 2. The valid subset contains structures with converged relaxation and $|\Delta V/V| < 15\%$. K denotes the number of top predicted SGs used to guide template retrieval. RMSD is reported for the evaluable ordered subset.

Metric	**Unconstrained**	**SG-guided K=10**	**SG-guided K=1**
Substitution success	376/500 (75.2%)	380/500 (76.0%)	350/500 (70.0%)
Relaxation valid	280/376 (74.5%)	282/380 (74.2%)	270/350 (77.1%)
SG match on valid subset	122/280 (43.6%)	164/282 (58.2%)	196/270 (72.6%)
SOAP cosine mean on valid subset	0.907	0.901	0.952
RMSD median on ordered subset (Å)	0.076	0.071	0.086

On AWA, SG-guided retrieval using the top predicted SG gives an SG match rate of 72.6% on the valid subset, compared with 43.6% for unconstrained retrieval, an increase of 29.0 percentage points. The corresponding MP increase is 22.9 percentage points. The same positive direction is therefore observed in both databases, despite differences in data source, composition distribution, and prevalence of partially occupied structures.

S2. Space-group match across compositional complexity on AWA

Table S2 stratifies the AWA LOEO results by the number of constituent elements, corresponding to the MP analysis in main-text Table 3.

Table S2. AWA SG match by number of constituent elements, corresponding to main-text Table 3. N valid is reported for unconstrained (unc), K = 10, and K = 1 retrieval. The $n \geq 5$ estimates provide trend-level support because the valid subsets remain limited, especially for $n = 7$. Three additional AWA targets at $n \geq 8$ have converged relaxations and are counted in the Table S1 totals; they are omitted from this table because each bin would contain a single case per strategy.

n_elements	**N valid (unc / K=10 / K=1)**	**Unc SG %**	**K=10 SG %**	**K=1 SG %**	**Δ (K=1 − unc)**
2	35 / 29 / 33	31.4	48.3	63.6	+32.2
3	121 / 121 / 125	47.1	59.5	76.0	+28.9
4	66 / 72 / 65	42.4	61.1	75.4	+33.0
5	38 / 38 / 32	47.4	63.2	68.8	+21.4
6	15 / 16 / 9	40.0	43.8	66.7	+26.7
7	2 / 3 / 3	50.0	66.7	66.7	+16.7

The SG-guided advantage is positive in every represented bin, ranging from +16.7 to +33.0 percentage points. The positive direction also remains for n ≥ 5. Compared with the MP bins reported in the main text, AWA provides more valid cases for the five- and six-element subsets, although the seven-element subset remains too small for a strong bin-wise conclusion.

S3. Retrieval comparison using the top 10, 3, and 1 predicted space groups

Table S3 reports the complete retrieval comparison on MP and AWA. The table compares unconstrained retrieval with SG-guided retrieval using the top 10, 3, and 1 predicted SGs.

Table S3. Retrieval comparison for unconstrained retrieval and SG-guided retrieval using the top 10, 3, and 1 predicted SGs on MP and AWA. SG match is evaluated on each strategy's valid subset. Retrieval guided by the top predicted SG gives the highest SG match in both databases. On AWA, retrieval guided by the top three predicted SGs delivers the highest substitution success; on MP its substitution coverage is essentially tied with retrieval guided by the top 10 predicted SGs.

Pool	Strategy	Substitution success	SG match (valid)
MP	Unconstrained	85.2%	40.1%
MP	K=10	86.6%	43.1%
MP	K=3	86.4%	49.6%
MP	K=1	81.4%	63.0%
AWA	Unconstrained	75.2%	43.6%
AWA	K=10	76.0%	58.2%
AWA	K=3	79.4%	68.2%
AWA	K=1	70.0%	72.6%

Retrieval guided by the top predicted SG gives the highest SG match in both databases. Retrieval guided by the top three predicted SGs gives the highest substitution success on AWA; on MP its substitution coverage is essentially tied with the top 10 predicted SGs while still delivering higher SG match, making it a practical choice when candidate yield is important. In contrast, using the top 10 predicted SGs is less effective for MP because descriptor-similarity ranking can still select templates from an incorrect SG within the larger candidate set.

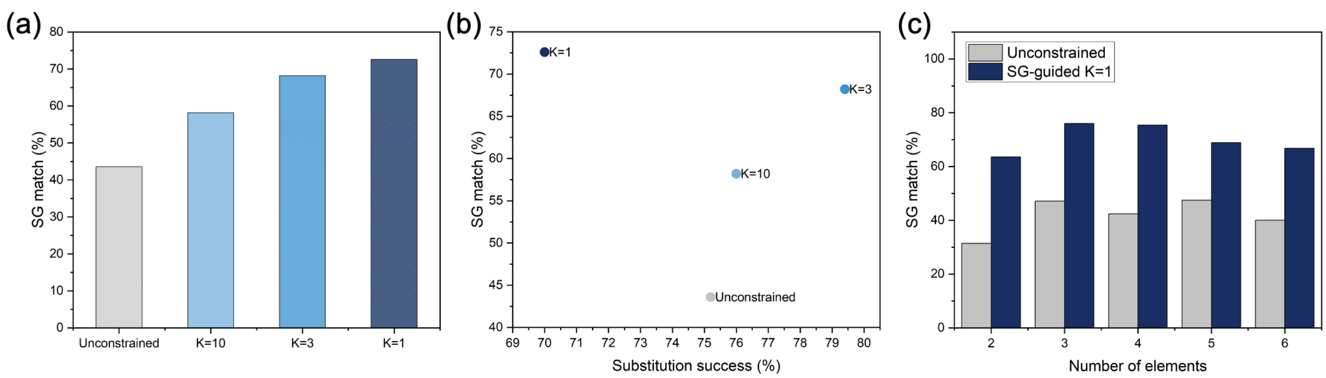

Figure S1. AWA evaluation of SG-guided retrieval. (a) SG match on the valid subset for unconstrained retrieval and SG-guided retrieval using the top 10, 3, and 1 predicted SGs. (b) Relationship between substitution success and SG match. Retrieval guided by the top three predicted SGs gives the highest substitution success, whereas retrieval guided by the top predicted SG gives the highest SG match. (c) SG match versus the number of constituent elements for unconstrained and top-1 SG-guided retrieval. The n = 7 bin is omitted because only three valid top-1 SG-guided cases were available; full values are reported in Table S2.

S4. Dataset composition distributions

Table S4 lists the composition distribution of the AWA and MP pools by number of constituent elements.

Table S4. Composition-complexity distribution of the AWA and MP pools by number of constituent elements. Based on the displayed bins, AWA contains a larger fraction of entries with at least five elements than MP (24.8% versus 13.1%), whereas MP is dominated by ternary entries after the stability filter.

n_elements	AWA %	MP %
1	0.3	0
2	8.8	12.3
3	35.9	45.7
4	30.2	28.9
5	16.3	9.8
6	5.6	3.2
7	1.6	0.1
8	0.6	0
9	0.3	0
10	0.4	0

AWA contains more multinary entries than MP, especially for compositions with five or more constituent elements, whereas MP after the stability filter is dominated by ternary entries. This difference supports the use of AWA as a further experimental cross-database evaluation of the MP-centered benchmark.

S5. Chemical-role assignments

Table S5 lists the nine chemical roles used by the substitution engine. Each element is assigned to one role according to its empirical tendency to occupy similar symmetry-equivalent sites. These assignments define the role capacities and scale-factor checks described in Methods 4.4.

Table S5. Chemical-role assignments used by the substitution engine. Tier 2 permits selected role mergers: alkali with alkaline-earth, alkaline-earth with lanthanide, and transition metal with post-transition metal.

Role	Elements
Alkali metals	Li, Na, K, Rb, Cs, Fr
Alkaline-earth metals	Be, Mg, Ca, Sr, Ba, Ra
Transition metals	Sc-Zn, Y-Cd, Hf-Hg
Lanthanides	La-Lu
Actinides	Ac-Lr
Metalloids	B, Si, Ge, As, Sb, Te
Post-transition metals	Al, Ga, In, Sn, Tl, Pb, Bi
Reactive nonmetals	H, C, N, O, F, P, S, Cl, Se, Br, I
Noble gases	He, Ne, Ar, Kr, Xe, Rn