# Supplementary Information

## 1 Supplementary Information

Every procedure is in the Methods; the S-sections below hold the tables and robustness analyses that the Methods subsections and the main-text callouts defer to. S0 is the index.

### 1.1 S0. Reader's map—which section answers which callout

**Table S1**

| Looking for | Where |
|---|---|
| model provenance: checkpoints, loaders, envs | S1 |
| test-set strata, counts, wave manifest | S2 |
| relaxation protocol, formation energies, references | Methods B |
| single-point control, eqV2 | Methods C |
| extended bootstrap intervals (phase 2) | S3 |
| database DFT settings, +U exposure | S4 |
| floor detail, MP metallic cut, Hegde | S5 |
| trim-cutoff sweep | S6 |
| Rung-0 constants, held-out, CHGNet decomposition | S7 |
| the full leaderboard (Fig. 12) | S8 |
| hull threshold sweep, like-for-like ceiling | S9 |
| symprec sweep | S10 |
| polymorph detail, Alexandria ceiling | S11 |
| elastic no-+U recut, log-space, metals-only | S12 |
| flag anatomy, panel robustness | S13 |
| "do the flags distort the headline?" | S14 |
| E_f binning, matched-E_hull, near-hull, OQMD cross-DB | S15 |
| magnetism stratum | S16 |
| cost / Pareto (Fig. 11) | S17 |
| data and code availability | S18 |
| AI assistance statement | S19 |

### 1.2 S1. Model provenance (`model_provenance.md`)

Model selection and the run protocol are Methods A. The table gives each potential's exact loader, checkpoint, and isolating environment (`envs/<name>`), rebuilt from the registry `src/umlip_bench/potentials.py`; package version pins and checkpoint hashes ship with the released dataset.

**Table S2**

| Model | Package | Loader / checkpoint | Env (`envs/`) |
|---|---|---|---|
| SevenNet-MF | `sevenn` | `7net-mf-ompa`, modal `omat24` | `sevenn` |
| GRACE-2L | `tensorpotential` (TF) | `grace_fm("GRACE-2L-OMAT", pad_atoms_number=40)` | `grace` |
| MACE-OMAT | `mace-torch` | `mace_mp(model="medium-omat-0", default_dtype="float64")` | `core` |
| DPA-3 | `deepmd` | `DP(model="dpa3.pth")` | `deepmd` |
| UMA | `fairchem` (v2) | `uma-s-1p1`, task `omat` | `fairchem` |
| Orb-v3 | `orb-models` | `orb_v3_conservative_inf_omat` (float32-high) | `orb` |
| eSEN | `fairchem-core` 1.10.0 | `facebook/OMAT24 : esen_30m_oam.pt` (OCPCalculator) | `fairchem-v1` |
| eqV2 | `fairchem-core` 1.10.0 | `facebook/OMAT24 : eqV2_86M_omat_mp_salex.pt` | `fairchem-v1` |
| MACE-MPA | `mace-torch` | `mace_mp(model="medium-mpa-0", default_dtype="float64")` | `core` |
| MatterSim-5M | `mattersim` | `MatterSim-v1.0.0-5M.pth` | `mattersim` |
| M3GNet | `matgl` | `M3GNet-PES-MatPES-PBE-2025.2` (stress_unit eV/Å$^3$) | `matgl` |
| CHGNet | `chgnet` | `CHGNetCalculator` (0.3.0) | `core` |
| SevenNet-0 † | `sevenn` | `7net-0` | `sevenn` |
| MatterSim-1M † | `mattersim` | default `MatterSimCalculator` | `mattersim` |

† superseded within-family checkpoint; shown only in the closing leaderboard table (Fig. 12) and the eight-model elastic comparison. Three reproducibility caveats: the eSEN [1] and eqV2 [2] checkpoints come from the **gated** `facebook/`

OMAT24 Hugging Face repo [3] and load only through the legacy OCPCalculator in fairchem-core 1.10.0 (a separate environment from UMA's fairchem v2); MACE runs in float64 because float32 stress noise stalls FIRE near minima; and M3GNet's stress_unit must be eV/Å$^3$ or ASE cell relaxation breaks silently.

## 1.3 S2. Test-set strata and counts

Test-set construction—scope, filters, strata, and sampling—is Methods A. **Strata and counts** (final manifest, results/analysis/per_stratum.csv):

**Table S3**

| Stratum | What | Count |
|---|---|---|
| LIB2 | binary intermetallic prototype decorations | 202,724 |
| LIB3 | ternary intermetallic prototype decorations | 1,178,736 |
| unary | elemental reference states | 30,808 |
| E3 (lanthanide) | 4f-alloy "edge of validity" probe | 9,153 |
| E1 / E1-ael | standard-DFT (no-U) nonmetals | 5,818 / 470 |
| E2 / E2-ael | +U entries (geometry only) | 19,882 / 2,749 |
| icsd_ael | AEL elastic subset | 325 core † |
| overlap | ICSD alloys matched to a Materials Project entry | 1,191 |
| ln_unary / ln_alloy | trivalent-lanthanide references / alloys | 956 / 3,189 |
| quaternary | metal-only 4-component alloys | 32 |
| **Total** | | **1,456,033 (body rounds to 1.46M)** |

† The AEL-backed elastic set is 325 in the wave-1 core and grows to 3,544 once the E1/E2 AEL sub-strata are added; the elastic comparison of the main text's §5 uses ∼3,500 AEL structures across all admitted chemistries.

**Construction in five waves.** The manifest accumulated in five monotone waves (Methods A), entries selected by ascending md5(auid) so each wave is a strict superset of the last:

**Table S4**

| Wave | Strata introduced | Entries |
|---|---|---|
| 1 | unary, lib2, lib3, icsd_ael, overlap | 88,430 |
| 2 | e1, e2, e3 | 16,956 |
| 3 | exhaustive expansion of the prototype libraries | 1,343,099 |
| 4 | quaternary | 32 |
| 5 | ln_unary, ln_alloy | 7,516 |
| **Total** | | **1,456,033** |

From wave 3 on, geometries are taken inline from AFLUX [4] (positions_fractional plus cell parameters); the earlier waves use fetched POSCARs.

Two data-availability facts shaped the design. AFLOW's own ICSD oxides are 97% already in MP, so an AFLOW oxide benchmark would add a second DFT answer but no new chemistry—hence the alloy-first scope. And AFLOW reports enthalpy_formation_atom for none of its +U intermetallics (0 of 4,579 LIB-matching +U entries), so the +U story is a geometry-and-methodology story, never a +U formation-energy comparison. Production ran the 1,456,033-entry manifest (∼1.46M structures) against up to fourteen models at a ∼0.03% hard-failure rate.

## 1.4 S3. Extended bootstrap intervals

The system-resampled bootstrap behind every "±" in the paper is Methods D. **Phase-2 intervals** (scripts/bootstrap_ci_phase2.py; same conventions—system-resampled, B = 2000, percentiles 2.5/97.5) extend it to the reference-free and elastic metrics:

**Table S5**

| Metric | Model | Estimate | 95% CI |
|---|---|---|---|
| Residual force, median max (eV/Å) | eqV2 | 0.00038 | [0.00037–0.00039] |
| Residual force, median max (eV/Å) | eSEN | 0.0217 | [0.0214–0.0221] |
| Residual force, median max (eV/Å) | UMA | 0.0401 | [0.0394–0.0409] |
| Elastic G $r$, all-cut linear | MACE-MPA | 0.160 | [0.06–0.89] |
| Elastic G $r$, no-+U | MACE-MPA | 0.965 | [0.939–0.979] |
| Symprec retention gap UMA−GRACE, 0.1 Å | — | −0.017 | [−0.043, +0.007] |
| Symprec retention gap UMA−GRACE, 0.01 Å | — | −0.160 | [−0.207, −0.117] |
| Familiar−novel gap (meV/atom) | eqV2 | +11.7 | [10.6–12.9] |
| Familiar−novel gap (meV/atom) | CHGNet | +47.6 | [44.8–50.4] |

The residual-force top-three ordering is non-overlapping; MACE-MPA's all-cut linear G $r$ is outlier-dominated (the quantified log-space caveat of S12) while its no-+U recovery is robust; the 0.1 Å symprec gap crosses zero while the 0.01 Å gap is real (symprec uses n = 300, resampled over structures); and the familiar-vs-novel gaps all exclude zero.

## 1.5 S4. DFT settings of the four databases

The floor comparison rests on the four databases sharing PBE but differing in protocol. Settings compiled from the primary sources (`docs/database_protocol_notes.md`):

**Table S6**

| Setting | AFLOW (LIB2/3) | Alexandria (PBE 3D) | OQMD | Materials Project |
|---|---|---|---|---|
| XC | PBE | PBE | PBE | PBE(+U) |
| ENCUT | AFLOW HT-standard | 520 eV | max ENMAX of POTCARs | 520 eV |
| k-points | AFLOW HT-standard | KPPRA ≈ 1000 | KPPRA 4000→8000 | KPPRA ≈ 1000 |
| +U | none (alloys) | oxides/fluorides (MP-style U) | correlated-TM oxides | oxides/fluorides (MP2020) |
| Elemental refs | raw DFT bulk elements | MP-compatible | FERE-style fitted | fitted / corrected |
| Post-hoc corrections | none | MP-compatible | FERE | MP2020 anion/+U mixing |
| VASP | HT-standard | 5.2 | 5.3.2 | — |

AFLOW's exact plane-wave and k-point settings are its high-throughput standard [5, 6] and we do not re-tabulate them. The operative rows are the last two: they are why Materials Project [7] is the outlier of Table 2 (Methods E), and why Alexandria—MP-pseudopotential-compatible yet raw-PBE on metals—lands with AFLOW, not with MP.

**+U exposure by scored population** (`scripts/ldau_groupby.py`). Every population the paper scores, grouped by AFLOW's `ldau_type` flag; the operative fact is that the 1.38M core alloy libraries carry none:

**Table S7**

| Population | +U fraction |
|---|---|
| Full benchmark (1.46M) | 1.90% |
| e2 stratum | 100% |
| e3 stratum | 53% |
| ln_unary | 11% |
| lib2/lib3 alloy libraries, unary, AFLOW∩MP overlap | 0% |
| Alexandria-matched floor set | 1.13% |
| OQMD-matched floor set | 11.2% |
| AEL elastic set | 78% (2,749 of 3,544) |
| Bad-reference flags (14,832 raw; 1,671 aligned) | 0% |

The no-+U cut of the elastic section (S12) is the remaining 795 structures.

## 1.6 S5. DFT-vs-DFT noise floor

The floor's construction—fingerprint matching, estimator choice, and the two reality checks—is Methods E. **Pairwise disagreement** (Table 2 with the columns the body omits):

**Table S8**

| Pair | median $\|\Delta\|$ (meV/atom) | trimmed MAE (meV/atom) | removable per-element offset | shared metallic structures |
|---|---|---|---|---|
| AFLOW–Alexandria | 6.0 | 23.9 | none ($R^2 \leq 0.10$) | ∼343,000 |
| AFLOW–OQMD | 8.1 | 26.3 | none ($R^2 \leq 0.10$) | ∼12,900 |
| OQMD–Alexandria | 7.7 | 27.8 | none ($R^2 \leq 0.10$) | ∼34,000 |
| AFLOW–Materials Project | 24.7 | not reported † | ∼25 meV/atom residual, not removable ($R^2$ ≈ 0.10) | ∼4,800 |

$R^2 \leq 0.10$ in every row means composition explains almost none of the structure-to-structure disagreement; for MP a ∼25 meV/atom median offset survives the fit.

† The AFLOW–MP trimmed-MAE floor is not reported: MP is the convention outlier, and its heavy tail is dominated by removable convention and bad-reference artifacts that a trimmed mean would misread as a raw-PBE floor. Quoting one would require first passing MP's outlier tail through the bad-reference adjudication of Methods I; the tail-robust median (24.7) and the non-removable ∼25 meV/atom residual are reported instead.

**Per-element offset test (Methods E)—element sets.** The metallic cut uses the nonmetal-exclusion set H, He, B, C, N, O, F, Ne, Si, P, S, Cl, Ar, Ge, As, Se, Br, Kr, Te, I, Xe, Po, At, Rn; the Fig. 1 floor-versus-E_hull binning uses a 22-element variant that counts Ge as a metal (both appear in the pipeline). The AFLOW∩MP metallic per-element $R^2 = 0.097$ was recomputed 2026-07-07 in the adjudication-robustness audit (`scripts/adjudication_robustness.py`).

**Floor versus distance from the hull (Fig. 1).** Binned against formation energy and against E_hull, the raw-PBE pairs hold a 5–11 meV/atom band out to E_f ≈ +750 meV/atom and E_hull ≈ 800 meV/atom (AFLOW–Alexandria, the largest sample, at 5–9 across its entire range); only the sparsest high-E_f AFLOW–OQMD bins ($n \leq 552$) rise to 12–15 meV/atom. The reference does not degrade appreciably far from equilibrium, and no single database is the noisy one.

**The MP metallic-cut correction (Methods E), in numbers.** The "removable convention" the first, contaminated decomposition found was dominated by O, S, and N offsets specifically; Hegde et al. [8] report median formation-energy disagreements of 105 meV/atom for AFLOW–MP and 87 for MP–OQMD against only 19 for AFLOW–OQMD.

## 1.7 S6. Trim-cutoff sensitivity

The 1 eV/atom trim (Methods D) removes only exceptional events, but its placement matters for the continuum-tailed mid-board: between cutoffs 0.25 and 1.0 eV/atom MACE-MPA's trimmed MAE grows 62% and Orb-v3's 56%, so the mid-board order is cutoff-sensitive while the leader and the tier structure are not. Trimmed MAE (relaxed protocol, meV/atom) at representative cutoffs (columns in eV/atom; `scripts/trim_sensitivity.py` → `trim_sensitivity.csv`):

**Table S9**

| Model / DFT pair | 0.25 | 0.5 | 1.0 | 2.0 | 5.0 |
|---|---|---|---|---|---|
| eqV2 | 13.9 | 17.5 | 18.5 | 18.9 | 19.6 |
| SevenNet-MF | 21.5 | 26.4 | 28.2 | 28.9 | 29.6 |
| GRACE-2L | 23.2 | 28.1 | 29.8 | 30.4 | 31.3 |
| MACE-OMAT | 26.3 | 31.0 | 32.6 | 33.2 | 33.9 |
| DPA-3 | 26.6 | 31.3 | 33.0 | 33.6 | 34.3 |
| UMA | 28.0 | 32.9 | 35.0 | 35.8 | 36.6 |
| Orb-v3 | 22.9 | 30.3 | 35.8 | 37.6 | 38.6 |
| eSEN | 31.0 | 36.2 | 38.6 | 39.4 | 40.2 |
| MACE-MPA | 29.7 | 43.5 | 48.2 | 48.9 | 49.5 |
| MatterSim-5M | 43.8 | 49.5 | 51.6 | 52.3 | 53.1 |
| M3GNet | 51.7 | 59.1 | 63.2 | 64.6 | 65.4 |
| MatterSim-1M † | 52.6 | 59.6 | 61.8 | 62.6 | 63.3 |
| SevenNet-0 † | 70.5 | 83.4 | 87.0 | 88.1 | 88.8 |
| CHGNet | 98.8 | 121.1 | 126.5 | 127.6 | 128.4 |
| *AFLOW–Alexandria floor* | 14.3 | 20.3 | 23.9 | 25.2 | 25.5 |
| *AFLOW–OQMD floor* | 20.0 | 24.6 | 26.3 | 26.7 | 27.3 |
| *OQMD–Alexandria floor* | 18.6 | 23.3 | 27.8 | 29.3 | 29.9 |

† superseded checkpoints. The model ordering is invariant for every cutoff ≥ 1.0 eV/atom—below that, mid-board near-ties reorder (Orb-v3 moves three places between the 0.5 and 1.0 cutoffs; M3GNet and MatterSim-1M swap)—and the leader stays below the AFLOW–Alexandria floor throughout. At the untrimmed limit two checkpoints' single corrupted rows dominate (GRACE $6.6\times10^3$, SevenNet-0 $4.0\times10^7$ meV/atom), which is why raw means are never reported; the full sweep including the untrimmed limit is plotted in `docs/figs/fig_trim_sensitivity.png`.

## 1.8 S7. Rung-0

The fit, its capacity bound, and the system-disjoint held-out split are Methods F.

**Fitted per-element constants.** The offsets are largest for the magnetic 3d elements, whose elemental reference differs most between conventions—Mn −118, Cr −102, Fe −77 meV/atom—with the non-magnetic metals near zero. These are the same constants that collapse the "bad reference" count in S13.

**Held-out behavior (Methods F).** Rung-0 leaves the OMat models slightly worse—the held-out penalty runs +0.9 to +2.7 meV/atom (eqV2 18.5 → 20.9, SevenNet-MF 28.2 → 30.7, Orb-v3 35.8 → 38.5; S8)—while the MP-trained models shed a large convention on top of a large genuine error (CHGNet 126→73, M3GNet 63→57, MatterSim 52→43). The full before/after table is Table 3; the held-out uncorrected MAE is within 0.6 meV/atom of the full-set value, so the before→after figures in the text use the full-set raw MAE.

**CHGNet, decomposed.** CHGNet's large Rung-0 gain separates cleanly: a single global constant (−115.7 meV/atom, overbinding against its own unaries) accounts for most of it (126.1 → 82.3 meV/atom), and element-specific offsets carry the rest, largest on the refractory metals.

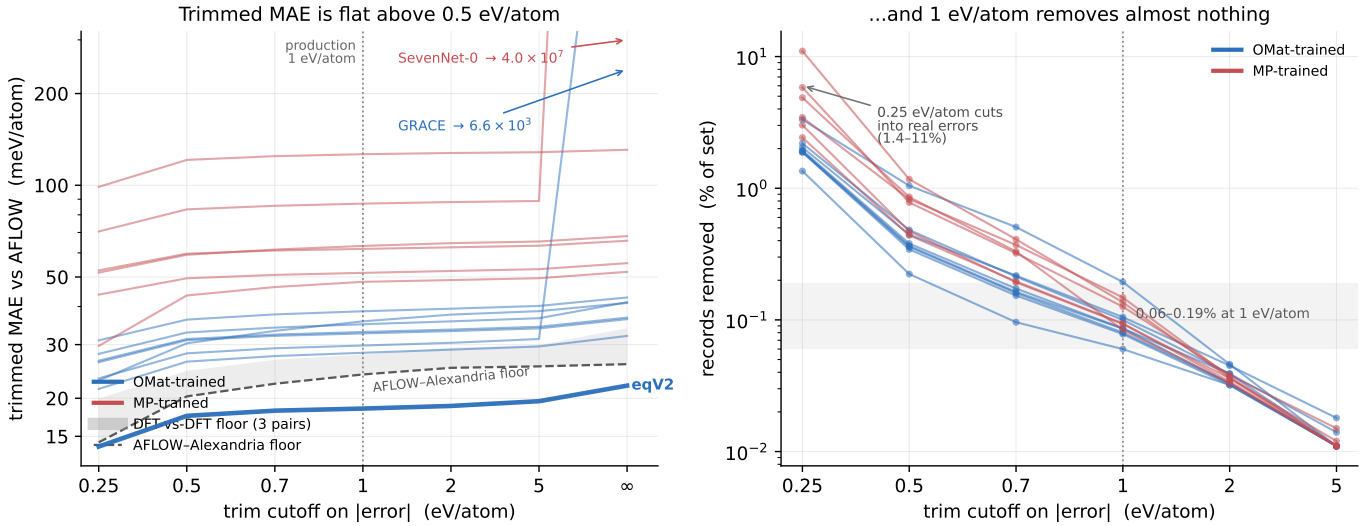


**Figure S1:** Trim-cutoff sensitivity: each model's trimmed MAE against the |error| cutoff, including the untrimmed limit where single corrupted rows dominate. The ranking is invariant for cutoffs at or above 1 eV/atom (Methods D).

**Table S10**

| Element | Centered offset (meV/atom) |
|---|---|
| Ru | −168 |
| W | −142 |
| Os | −128 |
| Mo | −99 |
| Cr | −89 |

These offsets correlate with neither the MP−AFLOW nor the Alexandria−AFLOW per-element reference differences (r = 0.11 / 0.18), so they are CHGNet-specific systematic error, not an inherited database convention. This is the paradox Rung-0 is built to resolve: model errors can be strongly composition-linear (held-out $R^2$ up to 0.62 for the convention carriers) while metallic DFT-pair disagreement is not ($R^2 \leq 0.06$), so the same one-constant-per-element fit strips much from the former and nothing from the latter (`rung0_chgnet_diagnosis.csv`).

**Robustness.** A 5-fold system-disjoint cross-validation reproduces the 70/30 corrected MAEs within 0.2 meV/atom (fold sd 0.13–0.52), so the split is not load-bearing; a Huber refit ($\delta = 1.345{\cdot}\sigma\_\mathrm{MAD}$) shifts absolute values down 0.4–5.6 meV/atom with rankings unchanged, and the OLS figures are quoted as the conservative choice (`scripts/rung0_robustness.py`).

## 1.9 S8. The full leaderboard (all metrics, fourteen models)

Figure 12 as a table: every model on every axis, with 95% system-resampled bootstrap half-intervals. Columns in paper order—raw MAE, single-point, after Rung-0, SP + Rung-0, hull-F1, E_hull MAE, residual force, polymorph $\rho$, no-+U shear. Dashes (—) mark metrics not run for a model—the shear column covers the eight elastic-task models on the no-+U cut, a panel that used the superseded MatterSim-1M checkpoint, so MatterSim-5M carries no shear value. The novel-vs-familiar gap is deliberately omitted as hull-distance-confounded (main-text §7).

**Table S11**

| Model | Raw MAE (meV/atom) | Single-point (meV/atom) | After Rung-0 (meV/atom) | SP + Rung-0 (meV/atom) | Hull-F1 | E_hull MAE (meV/atom) | Residual force (eV/Å) | Polymorph $\rho$ | No-+U shear r |
|---|---|---|---|---|---|---|---|---|---|
| eqV2 | 18.5 ± 0.4 | 14.7 ± 0.3 | 20.9 | 15.6 ± 0.7 | 0.930 ± 0.005 | 17.3 ± 0.8 | 0.0004 ± 0.0000 | 0.983 ± 0.000 | — |
| SevenNet-MF | 28.2 ± 0.5 | 19.4 ± 0.3 | 30.7 | 19.6 ± 0.7 | 0.887 ± 0.006 | 24.4 ± 0.8 | 0.0435 ± 0.0010 | 0.960 ± 0.001 | — |
| GRACE-2L | 29.8 ± 0.5 | 21.0 ± 0.3 | 31.1 | 21.4 ± 0.7 | 0.870 ± 0.007 | 26.8 ± 0.8 | 0.0420 ± 0.0009 | 0.955 ± 0.001 | 0.984 ± 0.005 |
| MACE-OMAT | 32.6 ± 0.5 | 24.2 ± 0.3 | 33.5 | 23.5 ± 0.7 | 0.850 ± 0.008 | 29.3 ± 0.8 | 0.0447 ± 0.0010 | 0.946 ± 0.001 | — |
| DPA-3 | 33.0 ± 0.4 | 23.9 ± 0.3 | 33.2 | 23.5 ± 0.6 | 0.842 ± 0.008 | 29.5 ± 0.8 | 0.0510 ± 0.0011 | 0.946 ± 0.001 | — |
| UMA | 35.0 ± 0.7 | 25.1 ± 0.7 | 29.2 | 14.2 ± 0.4 | 0.885 ± 0.010 | 23.7 ± 0.8 | 0.0401 ± 0.0007 | 0.966 ± 0.001 | 0.983 ± 0.008 |
| Orb-v3 | 35.8 ± 0.6 | 20.2 ± 0.3 | 38.5 | 19.6 ± 0.7 | 0.878 ± 0.007 | 22.3 ± 0.8 | 0.0449 ± 0.0009 | 0.964 ± 0.001 | 0.983 ± 0.006 |
| eSEN | 38.6 ± 0.8 | 30.0 ± 0.8 | 29.5 | 15.9 ± 0.7 | 0.891 ± 0.009 | 25.6 ± 0.8 | 0.0217 ± 0.0004 | 0.971 ± 0.001 | — |
| MACE-MPA | 48.2 ± 1.7 | 39.1 ± 1.7 | 35.3 | 23.1 ± 0.7 | 0.851 ± 0.008 | 42.4 ± 1.6 | 0.0466 ± 0.0011 | 0.944 ± 0.001 | 0.977 ± 0.007 |
| MatterSim-5M | 51.6 ± 0.5 | 39.1 ± 0.4 | 43.2 | 32.8 ± 0.7 | 0.773 ± 0.011 | 46.1 ± 0.8 | 0.0520 ± 0.0012 | 0.905 ± 0.001 | — |
| M3GNet | 63.2 ± 0.6 | 50.1 ± 0.5 | 56.9 | 46.4 ± 0.8 | 0.670 ± 0.012 | 54.0 ± 0.8 | 0.0958 ± 0.0024 | 0.868 ± 0.002 | 0.848 ± 0.034 |
| MatterSim-1M † | 61.8 ± 0.6 | 48.4 ± 0.5 | 55.6 | 45.3 ± 0.7 | 0.671 ± 0.012 | 54.6 ± 0.9 | 0.0893 ± 0.0023 | 0.868 ± 0.002 | 0.938 ± 0.030 |
| SevenNet-0 † | 87.0 ± 1.0 | 66.1 ± 0.8 | 64.4 | 55.1 ± 1.1 | 0.627 ± 0.012 | 69.8 ± 1.0 | 0.1067 ± 0.0029 | 0.833 ± 0.002 | 0.923 ± 0.017 |
| CHGNet | 126.5 ± 1.1 | 101.6 ± 0.9 | 72.9 | 61.3 ± 1.1 | 0.541 ± 0.013 | 92.5 ± 1.1 | 0.0980 ± 0.0024 | 0.825 ± 0.002 | 0.878 ± 0.030 |

± are 95% system-resampled bootstrap half-intervals (B = 2000; `bootstrap_ci_all.csv`, `bootstrap_ci_phase2.csv`, `bootstrap_ci_sprung0.csv`). † superseded checkpoints. The After-Rung-0 column carries no interval—no committed bootstrap covers the relaxed+Rung-0 combination—and the two Rung-0 columns are scored on the 30% held-out systems (per S7 the held-out uncorrected MAE is within 0.6 meV/atom of the full-set value, so the columns are comparable but not computed on identical sets). The shear column is the log-space r of S12 (`elastic_logspace.csv`). eqV2's

polymorph and residual-force intervals render as ±0.000 because the bootstrap collapses at a plateau (residual-force CI [0.00037, 0.00039]).

**The fully-controlled energy order, resolved.** Bootstrap CIs on the SP + Rung-0 column (Table 3), paired over the shared held-out systems (the correct test, since the models' errors on a common structure set are highly correlated; `scripts/bootstrap_ci_sprung0.py`), have marginal half-widths of ±0.4–0.7 meV/atom and paired differences that both exclude zero:

**Table S12**

| Paired difference (meV/atom) | Estimate | 95% CI |
|---|---|---|
| UMA − eqV2 | −1.38 | [−1.76, −0.99] |
| eqV2 − eSEN | −0.33 | [−0.39, −0.27] |

The fully controlled order is therefore a strict UMA < eqV2 < eSEN, with UMA formally lowest—but these intervals do not propagate the Rung-0 fit's own per-model noise (the +0.9 to +2.7 meV/atom held-out penalties of S7, the same few-meV scale as the differences), and the entire spread is under 2 meV/atom, far below the 24–28 meV/atom trimmed-MAE floor: the ordering is at best statistically resolvable and physically meaningless—the paper's own thesis, applied to its energy board.

## 1.10 S9. Hull-F1 robustness

Hull construction, the stability window, and the validation against pymatgen are Methods G.

**Threshold sweep** (`scripts/hull_f1_threshold_sweep.py`). eqV2 leads at every tolerance, and the chase pack (F1 0.84–0.89 at the 25 meV/atom operating point) swaps order across thresholds without resolving—eSEN leads it at 5 meV/atom, SevenNet-MF at 50. Hull-F1 at four stability windows:

**Table S13**

| Model | 5 meV/atom | 10 | 25 | 50 |
|---|---|---|---|---|
| eqV2 | 0.885 | 0.914 | 0.930 | 0.938 |
| eSEN | 0.859 | 0.874 | 0.891 | 0.898 |
| SevenNet-MF | 0.840 | 0.858 | 0.887 | 0.905 |
| UMA | 0.842 | 0.860 | 0.885 | 0.901 |
| Orb-v3 | 0.830 | 0.853 | 0.878 | 0.902 |
| GRACE-2L | 0.814 | 0.840 | 0.870 | 0.889 |
| MACE-MPA | 0.798 | 0.820 | 0.851 | 0.871 |
| MACE-OMAT | 0.786 | 0.811 | 0.850 | 0.873 |
| DPA-3 | 0.778 | 0.807 | 0.842 | 0.869 |
| MatterSim-5M | 0.724 | 0.748 | 0.773 | 0.804 |
| MatterSim-1M † | 0.599 | 0.624 | 0.671 | 0.730 |
| M3GNet | 0.581 | 0.618 | 0.670 | 0.732 |
| SevenNet-0 † | 0.545 | 0.573 | 0.627 | 0.688 |
| CHGNet | 0.477 | 0.506 | 0.541 | 0.607 |

**Like-for-like ceiling.** Table 4's Alexandria row (hull-F1 0.906) is computed on a different structure set than the model rows. Recomputed on the shared AFLOW∩Alexandria overlap (`scripts/alex_ceiling_shared.py`) the ceiling is 0.895 and the four leading models clear it: eqV2 0.943, eSEN 0.914, SevenNet-MF 0.914, UMA 0.910. eqV2 above the ceiling is protocol proximity, not super-DFT physics—fine-tuned on subsampled Alexandria, it matches AFLOW's raw-PBE convention more closely than a second independently run DFT database does.

## 1.11 S10. Space-group retention versus symprec

The tolerance problem and the four-model 300-structure sample are Methods G. Retention at each tolerance (`scripts/aggregate_symprec.py`):

**Table S14**

| Model | symprec 0.01 Å | 0.05 Å | 0.10 Å |
|---|---|---|---|
| UMA | 0.78 | 0.89 | 0.90 |
| Orb-v3 | 0.84 | 0.89 | 0.89 |
| GRACE-2L | 0.94 | 0.92 | 0.92 |
| CHGNet | 0.94 | 0.94 | 0.92 |

UMA's and Orb-v3's retention *rises* as the tolerance loosens while GRACE's and CHGNet's stays flat—the mechanism behind the gap collapse Methods G reports: the dramatic version of the gap is a symprec artifact, and only a small genuine fidelity difference remains.

## 1.12 S11. Polymorph ordering in detail

The metric—Spearman $\rho$ and ground-state hit rate per qualifying composition—is Methods G. Per-model medians at the three AFLOW-spread thresholds (`scripts/analyze_polymorph_ranking.py`):

**Table S15**

| Model | median $\rho$ (spread $\geq$ 25 / 50 / 100 meV/atom) | ground-state hit rate (25) |
|---|---|---|
| eqV2 | 0.98 / 0.98 / 0.98 | 0.57 |
| eSEN | 0.97 / 0.97 / 0.97 | 0.54 |
| UMA | 0.97 / 0.97 / 0.97 | 0.53 |
| Orb-v3 | 0.96 / 0.96 / 0.96 | 0.51 |
| SevenNet-MF | 0.96 / 0.96 / 0.96 | 0.51 |
| GRACE-2L | 0.95 / 0.95 / 0.96 | 0.51 |
| MACE-OMAT | 0.95 / 0.95 / 0.95 | 0.48 |
| DPA-3 | 0.95 / 0.95 / 0.95 | 0.48 |
| MACE-MPA | 0.94 / 0.94 / 0.94 | 0.48 |
| MatterSim-5M | 0.90 / 0.90 / 0.91 | 0.42 |
| M3GNet | 0.87 / 0.87 / 0.87 | 0.36 |
| MatterSim-1M † | 0.87 / 0.87 / 0.87 | 0.36 |
| SevenNet-0 † | 0.83 / 0.83 / 0.83 | 0.33 |
| CHGNet | 0.83 / 0.83 / 0.83 | 0.31 |

The two-tier divide is robust to the threshold: OMat-trained models at median $\rho$ 0.95–0.98, MACE-MPA at 0.94, the MP-trained generation at 0.83–0.91. The ground-state hit rate is humbling for all: asked to name the single lowest-energy polymorph, the leader is right 0.57 of the time and CHGNet 0.31, against a random-guess baseline of 0.073 (median 25 polymorphs per composition). The models capture the *shape* of the landscape while missing its exact minimum, because the lowest-lying polymorphs sit within the DFT-vs-DFT floor of one another: the Alexandria ceiling is top-1 0.880 at median $\rho$ 0.994, and eqV2 on the identical members reaches top-1 0.92, $\rho$ 1.00.

## 1.13 S12. Elastic response

The finite-strain protocol and the U-status stratification are Methods H. **Shear correlation before and after removing +U structures** (Fig. 6; `scripts/elastic_logspace.py`, `elastic_noU_wrangle.py`; the linear no-+U value for SevenNet-0, a superseded checkpoint, was dropped from the recut):

**Table S16**

| Model | Pearson r, all AEL | r, no-+U (∼795) | log–log slope (no-+U) | mult. bias (no-+U) |
|---|---|---|---|---|
| MACE-MPA | 0.16 | 0.97 | ∼1.0 | 0.97 |
| CHGNet | 0.51 | 0.83 | ∼1.0 | 0.48 |
| M3GNet | 0.46 | 0.72 | ∼1.0 | 0.71 |
| SevenNet-0 | ∼0.53 | — | ∼1.0 | 0.73 |
| UMA | 0.83 | 0.99 | 0.91–1.02 | 0.97–1.03 |
| Orb-v3 | 0.86 | 0.98 | 0.91–1.02 | 0.97–1.03 |
| GRACE-2L | 0.80 | 0.85 | 0.91–1.02 | 0.97–1.03 |

Bulk modulus tracks AFLOW for every model (r 0.78–0.98); the shear story is the +U story. On the no-+U subset every model's log–log slope lies between 0.91 and 1.02, so the residual "softening" of the older models is a uniform multiplicative bias (CHGNet 0.48×, M3GNet 0.71×, SevenNet-0 0.73×), not a slope deficit. MACE-MPA's headline r = 0.16 was itself partly a statistic's artifact: its full-set median multiplicative bias is 0.97 and its log-space correlation 0.92, so the collapse is what a small +U-concentrated population of catastrophic misses does to a linear correlation.

**Metals-only recut.** The AEL ground truth lives mostly on ICSD nonmetal compounds; only 1,259 of the 3,544 AEL structures (and 126 of the 795 no-+U ones) are purely metallic (`scripts/elastic_metals_recut.py`). Restricting to metals leaves the conclusion intact—+U-stratified shear correlations run 0.85–0.99 and the MP generation's +U sensitivity if anything sharpens—but at far lower statistical power, so the main text reports the full AEL set and this recut is the honest metals-only check.

## 1.14 S13. Bad references—anatomy of the flags

The consensus detector, its thresholds, the per-element re-referencing, the adjudication protocol, and the adjudication verdicts are all Methods I; this section holds what the flags *are*, plus one adjudication-provenance caveat.

**Flag geometry.** The failure is one-sided and structural—geometric, not energetic (`scripts/badref_characterize.py`, `badref_alex_geom_scan.py`). Of the 14,832 raw flags, 1,671 survive the per-element re-alignment (Methods I); AFLOW sits high in 99.7% of raw flags (median AFLOW − consensus +403 meV/atom raw, +1,085 aligned). The flagged structures depart sharply from the benchmark background in geometry, not just energy—volume per atom differs from the Alexandria partner by 22.3% against a 0.32% background, and model relaxations move the cell by 23–29% against 0.4–1.4%. The destination join is the near-proof of a wrong-branch reference: where Alexandria's volume differs from AFLOW's by more than 10%, 84% of model relaxations land within 10% of Alexandria's volume, and the flags' single-point errors (243–448 meV/atom) are roughly 3–5× smaller than their relaxed ones (2.7–4.9× per model)—AFLOW's energy is roughly right for a wrong structure.

Two systematic families dominate the aligned set. Three $AB_2$ *Fmmm* decorations (space group 69, Wyckoff patterns 1b,2i, 1a,2i, 1a,2f,2i) hold 460 of the 1,671 aligned flags—a 30.1% flag rate on a 1,526-member family, ∼260× the base rate, while the other 1,197 SG-69 structures carry only 7—a chemistry-dependent relaxation collapse leaving over-compressed published cells the models re-expand by a median +73%, enriched on Sn (7×), Be (4.5×), and Ba (4×). Duplicated broken unaries make up another 14% of aligned flags (identical stuck copies, E_f +450 to +1,204 meV/atom, volumes 11–21% above the ground state). The reverse population—where a third database sides with AFLOW and the models are wrong together—is 536 of 4,847 adjudicable raw flags (11%), concentrated in the 4d/5d refractories (Ir, Os, Tc, Re) with unremarkable geometry (median signed ΔV +0.6%): an energy-only collective miss consistent with shared training-reference conventions, on the same refractory elements that dominate CHGNet's fitted offsets (S7).

**Panel-expansion robustness.** Rebuilding the consensus detector on all fourteen models (`scripts/consensus_panel14.py`) gives 15,190 flags at 90.3% Alexandria pro-model against the production 8-model era's 14,832 at 88.9%, retaining 91.9% of the production flags (Jaccard 0.83); the paper's quoted counts stay with the production detector. The original adjudication matched 1,402 structures at 95%; its intermediate join tables were lost to scratch-space purging, and the committed strict-prototype reconstruction recovers 404 of them at a consistent 91%—both numbers are reported.

## 1.15 S14. Do the flags distort the headline numbers?

Barely, and symmetrically. The 1 eV/atom trim already removes about half of the raw flags; excluding the surviving 1,671 outright moves every model's trimmed MAE by less than 0.4 meV/atom. Excluding all 14,832 raw flags moves the models by 3.2–3.9 meV/atom and the like-for-like floor by 2.3—both sides shift together and no ranking changes (`scripts/build_badref_exclusions.py`). Because the flags are reference-convention artifacts present on both sides of the model-versus-floor comparison, removing them can neither manufacture nor erase a model's standing. Hull-F1 is immune: only 2–3 of the 24 hull-poisoning too-low flags survive any leader's kept set, and the recomputed F1 is unchanged to four decimals (`badref_hull_sensitivity.csv`). The two surgical exclusion rules are materialized in `badref_exclusions.csv`—the three-pattern SG-69 family (1,612 structures across all strata, a flag-then-inspect superset of the 1,526 counted above) and a keep-one-representative dedup of broken unary copies (11,477 entries)—but headline metrics in this paper are computed on the full test set, and the list ships with the reproduction repository as a robustness artifact whose adoption as a default filter is left to future revisions.

## 1.16 S15. Familiarity, novelty, and generalization

The familiarity fingerprint, the matched-E_hull control, the near-hull comparison, and the OQMD campaign design are Methods J.

**Formation-energy binning** (the SI form of Figs. 8–9; `scripts/analyze_err_vs_ef.py` → `results/analysis/error_vs_ef.csv`; figure `docs/figs/fig_err_vs_ef.png`). Binned in 50-meV slices of AFLOW E_f (73 bins per model with ≥100 structures each, −1.1 to +2.7 eV/atom), every model's median |error| climbs with formation energy: the Spearman correlation between per-bin median error and bin center is:

**Table S17**

| Model | $\rho$ | Model | $\rho$ |
|---|---|---|---|
| SevenNet-0 † | 0.99 | eSEN | 0.92 |
| M3GNet | 0.97 | GRACE-2L | 0.91 |
| MACE-MPA | 0.97 | eqV2 | 0.87 |
| CHGNet | 0.96 | MatterSim-5M | 0.84 |
| MACE-OMAT | 0.93 | Orb-v3 | 0.80 |
| DPA-3 | 0.93 | SevenNet-MF | 0.78 |
| MatterSim-1M † | 0.92 | UMA | 0.69 |

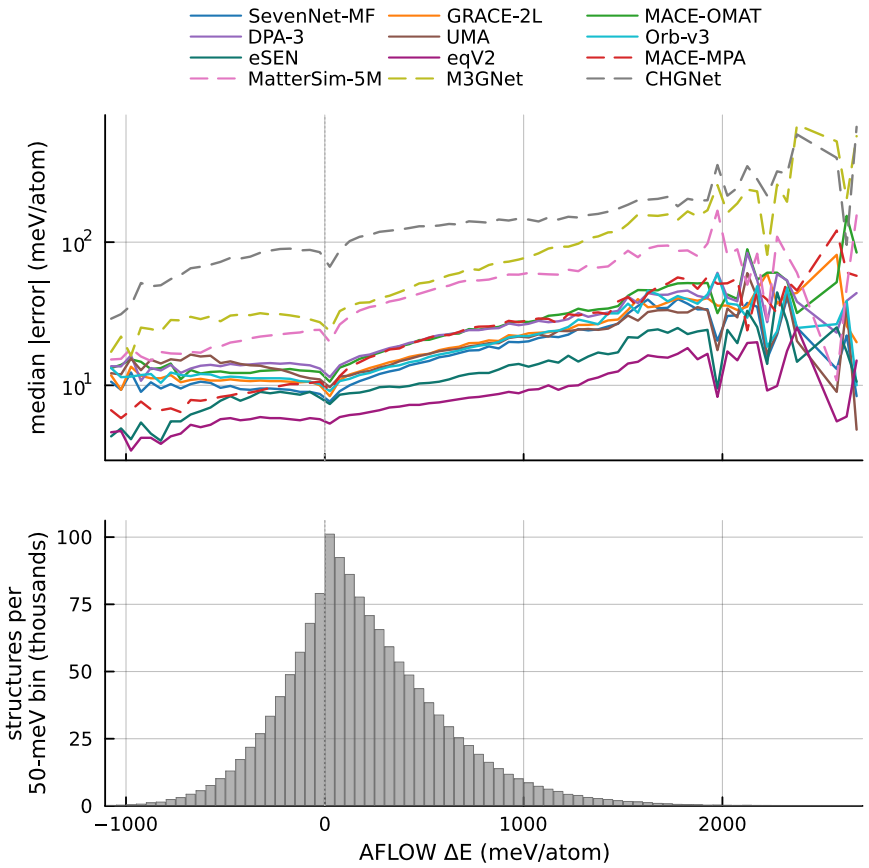


**Figure S2:** Median formation-energy error against AFLOW formation energy in 50-meV/atom bins (the SI form of main-text Figs. 8–9): every model's error climbs with formation energy (Methods J).

Ten of the fourteen sit at 0.84–0.97; the three low outliers (Orb-v3, SevenNet-MF, UMA) go noisy only in the sparse bins above +2 eV/atom—restricted to E_f $\leq$ 2 eV/atom, all fourteen sit at 0.81–0.99, the numbers the main text quotes.

The rise tracks the sign of the energy, not its magnitude: at matched |E_f| $\approx$ 1 eV/atom (bins centered −975 vs +975 meV/atom), median error is 1.4–4.7× larger on the unfavorable side (panel median 2.9×; MACE-MPA 6.7 → 27.6 meV/atom, CHGNet 36.1 → 146.0, eqV2 3.5 → 8.8; smallest UMA 15.4 → 21.9). And E_hull is the sharper axis (`scripts/analyze_err_vs_hull.py` → `err_vs_hull.csv`, `quartiles_ef.csv`, `quartiles_ehull.csv`): across the ten E_hull bins (edges 0–6,400 meV/atom) all fourteen models are monotonic to Spearman 0.99–1.00, and the lowest-to-highest E_hull quartile gives the main text's Fig.-9 numbers exactly—SevenNet-MF 7.3 → 16.4 meV/atom, MACE-MPA 7.5 → 22.6, CHGNet 51.2 → 133.6—while the same models binned on E_f quartiles start higher and rise less cleanly (9.1 → 16.3, 9.3 → 22.3, 81.3 → 127.4, the first two E_f quartiles non-monotonic for SevenNet-MF and CHGNet).

**Formation-energy distribution** (Fig. 8; `ef_dist_compare.csv`, `ef_overlap_compare.csv`, `oqmd_metals_ef_dist.csv`; full-OQMD pipeline in `scripts/oqmd_dump/`). On the committed 50-meV histograms: AFLOW's deduplicated metals (1,042,241 entries) put 71.2% of structures above zero with the median in the +150–200 meV/atom bin (the caption's unbinned +172); Alexandria's AFLOW-overlapping metals track that shape at 70.5% above zero; and OQMD's complete metallic set (111,901 entries from the qmpy dump) sits at 66.6% above zero with its median in the same bin (the caption's +159)—three independently constructed databases agree that the combinatorial libraries skew unfavorable. (Provenance: the exact medians +172/+159 and the 343k Alexandria-overlap and 111,901 OQMD counts come from the cluster-side unbinned dumps documented in `scripts/oqmd_dump/README.md`; the committed 50-meV histograms reproduce the binned distribution.) The scored benchmark itself (`ef_dist.csv`, 1,420,601 structures) is marginally less skewed: 68.8% above zero, median bin +100–150. The AFLOW–Alexandria floor stays flat along the same axis (per-structure median |ΔE_f| 5.1–7.5 meV/atom in every 250-meV E_f bin from −1.0 to +1.5 eV/atom, 8.4 only in the sparsest bin at n = 173; `floor_vs_ef.csv`, the E_f-axis companion of S5's Fig. 1), so the error that rises with formation energy belongs to the models, not to the benchmark.

**Matched-E_hull re-binning** (`scripts/matched_ehull_generalization.py`; the control of Methods J), per model: eqV2 +11 → +1.2, SevenNet-MF +15 → +1.0, GRACE +16 → +1.5; the convention carriers eSEN and UMA start at +20 and +10 and collapse only once Rung-0 is applied (eSEN +20.2 → −0.8, UMA +10.0 → +0.1). Neither confound alone accounts for the gap; both together erase it.

**Near-hull disagreement** (Fig. 10; `scripts/near_hull_dft_floor.py`; design in Methods J). The window's DFT

floor is median 5, mean 10 meV/atom. Beyond eqV2 riding that floor: MACE-MPA and UMA carry heavier tails of 0.1–1 eV/atom misses while SevenNet-MF, Orb-v3, and MACE-OMAT decay cleanly, and the DFT pair's own tail is flat past 100 meV/atom—even two careful databases carry a small catastrophic-disagreement population.

**OQMD cross-database campaign** (`scripts/analyze_oqmd_aligned.py`, `analyze_oqmd_metals.py`; design in Methods J), per model on the 12,409-compound covered set: UMA 21, SevenNet-MF 23, GRACE 25 meV/atom—the same leading group as the re-referenced AFLOW board; M3GNet 63 and CHGNet 64 trail—the leaderboard is a property of the models, not of AFLOW.

### 1.17 S16. Magnetism

Magnetic systems are a labeled stratum, never pooled into the nonmagnetic headline. MACE-MPA is the instructive case: it classifies stability as well as the OMat models yet carries its largest errors on magnetic intermetallics (magnetic vs non-magnetic trimmed MAE 90.4 vs 37.4 meV/atom (`results/analysis/magnetic_slice.csv`)), the same magnetic-3d reference convention that S7 and S13 remove. Reported separately so it never blurs the metallic headline.

### 1.18 S17. Compute cost and the Pareto frontier

Cost measurement—core-seconds, the matched CPU re-measure for the GPU-run models, and eqV2's exclusion—is Methods K. The efficient set depends on the metric (Fig. 11): Orb-v3 and UMA hold the mid-cost frontier for hull-F1, polymorph ordering, and forces; eSEN ($\sim$73 core-seconds, the most expensive) buys the high-accuracy tip of all three; DPA-3, MACE-OMAT, and SevenNet-MF carry the raw-energy frontier; the cheap M3GNet and MatterSim anchor every frontier, and CHGNet is dominated throughout.

### 1.19 S18. Data and code availability

The FAIR dataset and analysis-code release are described in Methods K. Each figure's generating script is named in its caption or the relevant SI section; model checkpoints and environment pins are in S1.

### 1.20 S19. Use of AI assistance

The use-of-AI disclosure is given in Methods §L.

## References


[1] X. Fu, B. M. Wood, L. Barroso-Luque, D. S. Levine, M. Gao, M. Dzamba, and C. L. Zitnick. Learning Smooth and Expressive Interatomic Potentials for Physical Property Prediction. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 17875–17893, 2025. doi: 10.48550/arXiv.2502.12147.

[2] Y.-L. Liao, B. Wood, A. Das, and T. Smidt. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2306.12059.

[3] L. Barroso-Luque, M. Shuaibi, X. Fu, B. M. Wood, M. Dzamba, M. Gao, A. Rizvi, C. L. Zitnick, and Z. W. Ulissi. Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models. 2024. doi: 10.48550/arXiv.2410.12771. Preprint.

[4] F. Rose, C. Toher, E. Gossett, C. Oses, M. Buongiorno Nardelli, M. Fornari, and S. Curtarolo. AFLUX: The LUX materials search API for the AFLOW data repositories. *Computational Materials Science*, 137:362–370, 2017. doi: 10.1016/j.commatsci.2017.04.036.

[5] S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, and D. Morgan. AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012. doi: 10.1016/j.commatsci.2012.02.005.

[6] C. E. Calderon, J. J. Plata, C. Toher, C. Oses, O. Levy, M. Fornari, A. Natan, M. J. Mehl, G. Hart, M. Buongiorno Nardelli, and S. Curtarolo. The AFLOW standard for high-throughput materials science calculations. *Computational Materials Science*, 108:233–238, 2015. doi: 10.1016/j.commatsci.2015.07.019.

[7] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. doi: 10.1063/1.4812323.

[8] V. I. Hegde, C. K. H. Borg, Z. del Rosario, Y. Kim, M. Hutchinson, E. Antono, J. Ling, P. Saxe, J. E. Saal, and B. Meredig. Quantifying uncertainty in high-throughput density functional theory: A comparison of AFLOW, Materials Project, and OQMD. *Physical Review Materials*, 7(5):053805, 2023. doi: 10.1103/PhysRevMaterials.7.053805.