Chemical filters for ultra-high-throughput materials screening and generation

Kinga O. Mastej, Panyalak Detrattanawichai, Hyunsoo Park, Anthony Onwuli, Masahiro Negishi, Aron Walsh

Department of Materials, Imperial College London | **k.mastej24@imperial.ac.uk**

NEURAL INFORMATION PROCESSING SYSTEMS

AI4Mat Workshop

NeurIPS 2026, Sydney

IMPERIAL

Motivation

Generative models — large candidate sets at low cost

Composition screening — SMACT: 12–26 ms per composition

Simulation — MLIP and DFT; high computational cost

Synthesis — experimental validation; highest cost

candidate numbers decrease and cost per candidate increases along the pipeline

1 SMACT: a chemical validity operator with tunable strictness

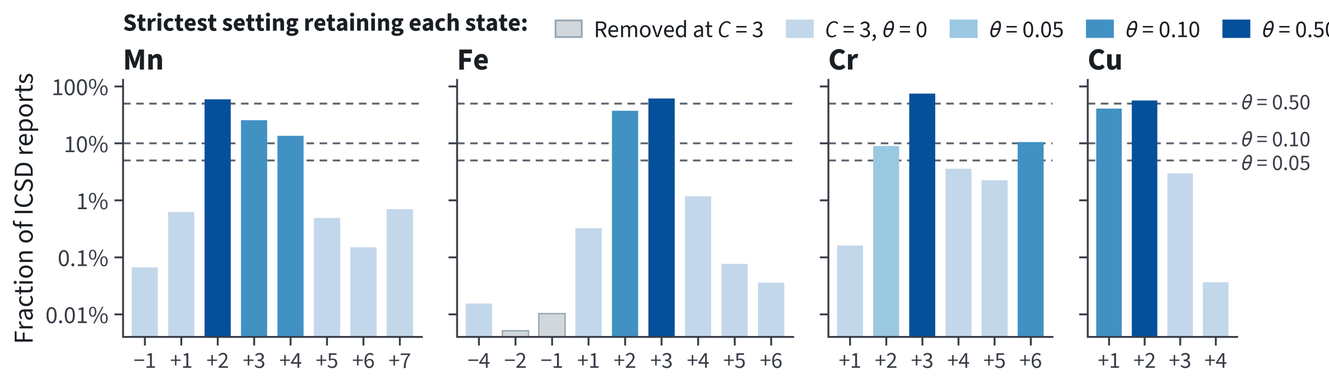
Existing binary check: a single valid assignment suffices

	1 · ICSD states	2 · Charge neutrality	3 · Electronegativity	Result
MgO	Mg^{+2} O^{-2}	✓ (+2) + (−2) = 0	✓ χ(O) 3.44 > χ(Mg) 1.31	valid
Mg_2O	Mg^{+2} Mg^{+2} O^{-2}	✗ 2(+2) + (−2) = +2	not evaluated	invalid

This work: thresholds on ICSD oxidation-state statistics

A state is retained if it is sufficiently documented (count $f \geq C$, consensus) and sufficiently common for its element (fraction $p \geq \theta$, commonality):

$$\Omega_{\text{filtered}}(C,\theta) = \left\{ E^q \,:\, f(E^q) \geq C \,\wedge\, p(E^q) \geq \theta \right\}, \qquad p(E^q) = \frac{f(E^q)}{\sum_{q'} f(E^{q'})}$$

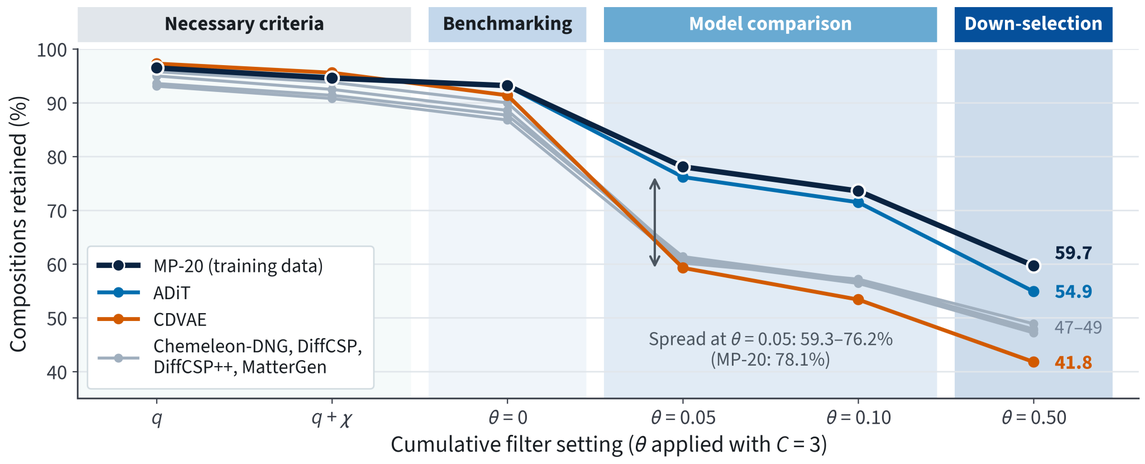


Fraction of ICSD reports for each oxidation state (log scale), coloured by the strictest setting under which it is retained. Rare oxidation states are not excluded *a priori*: Mn^{7+} (as in Mn_2O_7) accounts for 0.7% of Mn reports and is removed only from $\theta = 0.05$. At $\theta = 0.50$ only Mn^{2+}, Fe^{3+}, Cr^{3+} and Cu^{2+} remain.

```
SMACT 4.0  smact_validity("MnO2", icsd_filter=ICSD24FilterConfig(consensus=3, commonality=10)) # θ = 0.10, in %
```

2 Generators over-sample rare oxidation states

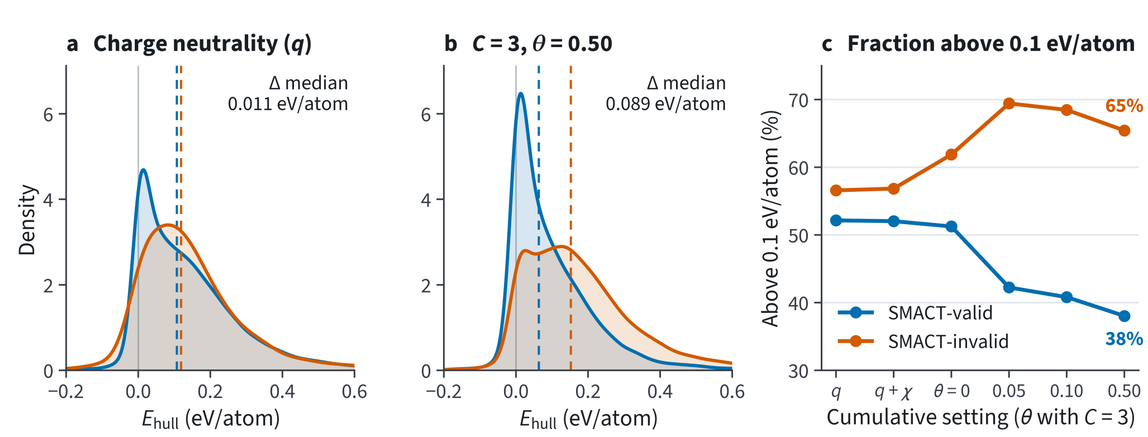
SMACT validity is a standard plausibility metric in generative-model benchmarks. Each model was sampled 10,000 times and screened under cumulative settings.



Here q denotes charge neutrality and χ the electronegativity rule. All six models reproduce both but fall below MP-20 under the commonality prior. Recommended: $C = 3$, $\theta = 0$ for benchmarking (93.2% of MP-20 retained); $\theta = 0.05$–0.10 for model comparison; $\theta = 0.50$ for synthesis-oriented down-selection. Retention should be reported relative to a reference set.

3 Rejected compositions are less stable

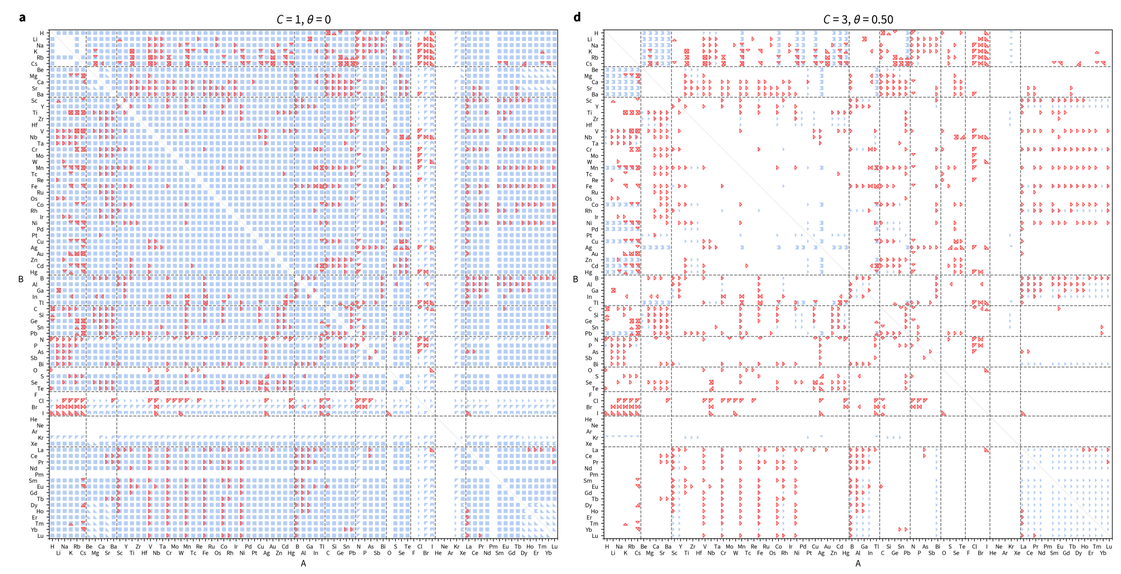
Energy above the convex hull for 58,469 structures from all six models, relaxed with MACE-MPA-0 and split *post hoc* into SMACT-valid and SMACT-invalid compositions.



At $\theta = 0.50$, 65% of invalid and 38% of valid compositions lie above 0.1 eV/atom; the median E_{hull} difference grows from 0.011 to 0.089 eV/atom with strictness (Mann–Whitney $p \ll 10^{-10}$). This is a distributional correlation, not a per-compound stability prediction.

4 Application: the ABX_3 perovskite design space

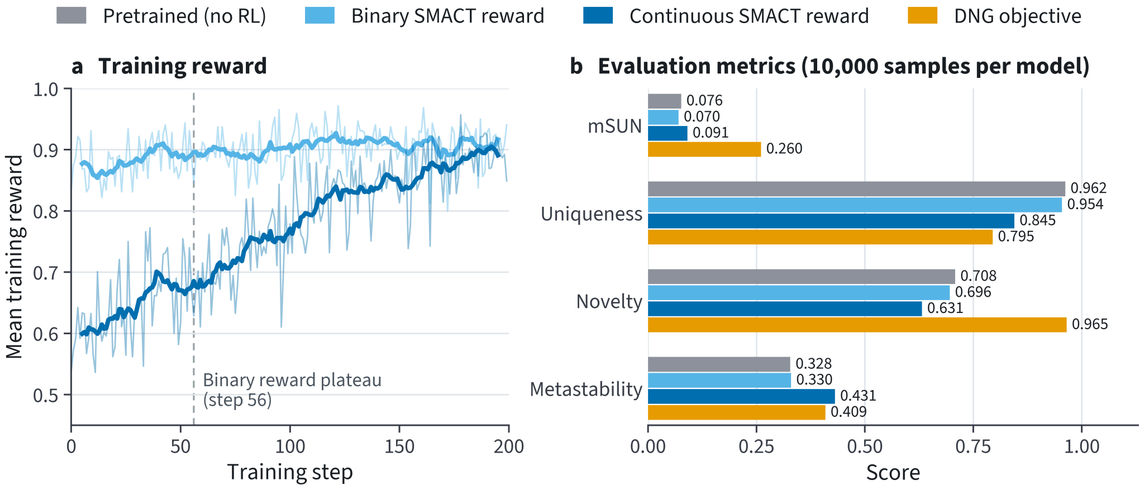
Admissible A–B combinations at the most permissive and strictest settings.



Blue: admissible; red: known Materials Project perovskites (X = O, Cl, Br, I). Admissible compositions fall from 11,091 to 781. Known perovskites are also removed (retention 90.7% → 34.9%), but the hypothetical space contracts more strongly, so their share rises from 5.4% to 29.4%.

5 Oxidation-state rewards steer generation

Chemeleon2, a latent diffusion model, was fine-tuned with GRPO for 400 steps using binary or continuous SMACT rewards, or the default DNG objective.



The binary reward plateaus by step 56; the continuous reward raises the metastable fraction from 0.33 to 0.43 at a cost in uniqueness and novelty. The DNG objective, which targets these metrics directly, attains the highest mSUN; SMACT rewards complement it as a chemically grounded auxiliary objective.

Conclusions

1. **Chemical plausibility is graded.** Consensus and commonality thresholds enable SMACT to interpolate between permissive and conservative screening.
2. **Generative models reproduce charge neutrality but over-sample rare oxidation states** relative to MP-20, a deficiency not captured by uniqueness, novelty or stability metrics; the compositions concerned are less stable.
3. **A continuous oxidation-state reward steers generation** towards metastable compositions at the expense of novelty; weighting against novelty objectives may tune this trade-off.

Google Scholar K. O. Mastej

Paper arXiv 2607.17910

Data and figures GitHub repository

SMACT code pip install smact

Open to research internships: generative models, MLIPs

References: SMACT, Davies *et al.*, JOSS 4, 1361 (2019); Chemeleon2, Park & Walsh, arXiv:2511.07158 (2025); MatterGen, Zeni *et al.*, Nature 639, 624 (2025); CDVAE, Xie *et al.*, ICLR (2022); ADiT, Joshi *et al.*, ICML (2025); DiffCSP, Jiao *et al.*, NeurIPS (2023); DiffCSP++, Jiao *et al.*, ICLR (2024); Chemeleon, Park *et al.*, Nat. Commun. 16, 4379 (2025); MACE, Batatia *et al.*, J. Chem. Phys. 163, 184110 (2025). Supported by EPSRC (EP/X037754/1) and the AI for Chemistry: AIchemy hub (EP/Y028775/1, EP/Y028759/1).