GEMBENCH: SEPARATING DISTRIBUTION MATCHING FROM DISCOVERY IN CRYSTAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models for crystals can serve two distinct aims: reproducing the distribution embodied in the training dataset, or discovering new stable materials beyond it. Classical generative metrics (Fréchet distances, kernel MMD, and related feature-space distances) evaluate only the first aim, while widely used S.U.N.-style metrics (fraction of stable, unique, and novel structures) are often reported as discovery progress without declaring the discovery aim. We make this split explicit as the *generative discovery dilemma*: the two aims share the same notion of stability, yet prescribe different optimal output distributions, so evaluation must declare the aim. We then introduce GEMBench (Generative Evaluation for Materials), which scores the two aims separately as GEM-match and GEM-disc under explicit choices of stability probe and crystal-similarity measure. A shared multi-probe stability panel (energy above hull, phonon, and finite-temperature checks) replaces the single energy-above-hull bit used by S.U.N.-style metrics. Both scores exclude near-duplicates of training-set crystals to avoid rewarding memorization. Beyond that, GEM-match rewards dataset-like distributional shapes, while GEM-disc rewards a large and diverse set of stable structures. On public generators trained on MP-20, GEM-match, GEM-disc, and CHGNet S.U.N. rank the same models differently. Which generator leads depends on the declared aim and on how crystal similarity is defined. In particular, the CHGNet S.U.N. leader ranks last on GEM-disc, because it concentrates on a few modes of the crystal space rather than discovering many distinct crystals.

1 INTRODUCTION

Classical generative metrics score agreement with a target distribution: Fréchet Inception Distance (FID) (Heusel et al., 2017), Inception Score (Salimans et al., 2016), kernel maximum mean discrepancy (MMD) (Gretton et al., 2012), and precision–recall in feature space (Sajjadi et al., 2018; Kynkäänniemi et al., 2019). Under that view, a successful generator reproduces the law embodied in the training dataset. Discovery adds a competing goal: exploration beyond the training dataset, with samples required to satisfy domain filters (Park et al., 2024; De Breuck et al., 2025). In crystal generation the stable, unique, and novel (S.U.N.) rate (Zeni et al., 2025) is the usual step toward that goal, and many models are tuned to raise it (Park & Walsh, 2026; Liu et al., 2026).

Stability dominates S.U.N. and acts as a task-agnostic gate before any distributional comparison, yet common yields reduce it to one hull bit (Zeni et al., 2025; Betala et al., 2025) while thermodynamic, phonon, and finite-temperature screens are not interchangeable (De Breuck et al., 2025; Han et al., 2025). Novelty is equally essential: crystal generators frequently reproduce structures from the training dataset rather than sample new ones. Together, stability and novelty have elevated S.U.N.-style yields to the dominant headline metric while weakening the connection to classical training-distribution metrics. We make this split explicit by formulating the generative aims underlying these two evaluation cultures. In crystal generation, an aim is fully specified only once the stability checks that define a stable crystal and a crystal-similarity measure for set-level geometry have been fixed. Both aims are restricted to novel structures, as outputs that reproduce the training dataset constitute memorization rather than generation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Matching is an aim of reproducing the law of the training dataset. For crystal generation it means agreement with the training dataset in both the fraction of generated crystals that pass the stability checks and the shape of the distribution of the stable part relative to the dataset under the chosen crystal-similarity measure.

Discovery is an aim of finding distinct stable objects. For crystal generation it means, under a fixed evaluation budget, attaining as high a fraction of generated crystals that pass the stability checks as possible while the stable part remains mutually dissimilar under the chosen crystal-similarity measure.

Generative discovery dilemma is the incompatibility of these optimality criteria. Under fixed stability checks and a crystal-similarity measure, matching rewards agreement with the dataset while discovery rewards breadth beyond it.

We consider unconditional generation only, and task-conditioned generation is out of scope. Section 2 situates prior metrics against the matching and discovery aims above. Section 3 defines GEMBench as a protocol that scores those aims under explicit stability checks and crystal similarity. Section 4 shows that the two aims and CHGNet S.U.N. reorder the same checkpoints differently, including a S.U.N. leader that ranks last on GEM-disc.

Contributions:

- A formulation of the *generative discovery dilemma*: matching and discovery prescribe different optima even when stability checks and the similarity measure are held fixed, so they must not be collapsed into one scalar.
- GEMBench, Generative Evaluation for Materials, which scores the two aims separately under explicit stability checks and set-level crystal similarity, with explicit combination rules GEM-match and GEM-disc.
- An empirical study of public checkpoints showing that matching, discovery, and CHGNet S.U.N. rank the same generators differently, depending on the declared aim and the similarity measure.

2 Related work

MatterGen made the stable, unique, and novel (S.U.N.) rate the default discovery headline (Zeni et al., 2025), and later models report the same product or close variants (Kazeev et al., 2025; Betala et al., 2025). S.U.N. is a single batch yield: it does not compare the stable part to the training-dataset law (*matching*), and for *discovery* it multiplies stability with discrete novelty and uniqueness flags, so a high product can still reflect near-duplicate motifs rather than the mutually dissimilar stable set in Section 1. Continuous variants fold the three legs into one tunable scalar (Negishi et al., 2026), masking collapse within the passing set. Uniqueness via `StructureMatcher` identity (Ong et al., 2013) can count two crystals as unique while they remain close in crystal space (De Breuck et al., 2025).

On the distributional side, prior crystal metrics rarely separate which crystals pass evaluation filters from how passing crystals are distributed in crystal space. The CDVAE template (Xie et al., 2021) reports validity and test-set coverage on the full sample and Wasserstein distances on density, energy, and element count on a *validity*-filtered subsample (minimum distances and charge neutrality). It does not score a stable subset, report training-set novelty, or evaluate matching on stable crystals or discovery breadth at fixed budget. LeMat-GenBench (Betala et al., 2025) reports the same S.U.N.-style yields and adds Jensen–Shannon, MMD, and Fréchet distances on compositions, densities, and MLIP embeddings, but computes those distances on validity-filtered structures rather than on the stability-passing set, pooling unstable with stable crystals and summarizing coarse marginals that need not reflect structure-level agreement with the training dataset. Classical image metrics (FID, kernel MMD, precision–recall) (Heusel et al., 2017; Salimans et al., 2016; Gretton et al., 2012; Sajjadi et al., 2018; Kynkäänniemi et al., 2019) share this pattern on de novo generation. Composition-conditioned CSP reconstruction (METRe, cRMSE) (Martirossyan et al., 2025) is out of scope (Section 1).

Generator leaderboards usually define stability as energy above hull after MLIP or DFT relaxation (Zeni et al., 2025; Betala et al., 2025), yet a hull-passing crystal can fail phonon or finite-

temperature checks (De Breuck et al., 2025; Han et al., 2025). PhononBench reports much lower dynamical-stability rates than hull-based S.U.N. on the same outputs (Han et al., 2025). Universal MLIPs make richer screens tractable (Deng et al., 2023; Batatia et al., 2023; Loew et al., 2025), but progress is still tracked along separate axes (reference marginals (Xie et al., 2021; Betala et al., 2025), hull yields (Zeni et al., 2025), phonon rates (Han et al., 2025; Loew et al., 2025)) without the joint readouts needed for the separable *matching* and *discovery* aims in Section 1. Formalizing those aims clarifies what existing leaderboards measure and what remains unreported when a model claims improvement on one headline number. GEMBench (Section 3) supplies the missing structure: one shared stability panel, calibrated set-level similarity kernels, explicit novelty filtering, and two scores (GEM-match, GEM-disc) that implement the matching and discovery optimality criteria without collapsing them into one scalar.

3 GEMBENCH

3.1 NOTATION

GEMBench (Generative Evaluation for Materials) assigns each generator two separate scores, GEM-match and GEM-disc, never averaged into one leaderboard. This subsection fixes shared notation in pipeline order. Section 3.2 and Section 3.3 give probe pass criteria and the similarity measure, and Sections 3.4 and 3.5 give the formulae. Figure 1 summarizes the full pipeline after the score definitions below.

Generation pool. A generation batch is the N crystals from one checkpoint evaluation (N=10,000 in the paper rankings), and both scores share it. D is the training dataset for novelty and matching, and G is the generator under evaluation, with $\Pr_G$ and $\Pr_D$ denoting empirical frequencies on the generation batch and on D (conditional pass rates in Section 3.4).

Eligibility. $V = 1$ if a crystal structure is parsable and multi-element, a minimal gate distinct from stability (Section 3.2) and from CDVAE/LeMat validity (Section 2). $N_D = 1$ if it is not matched to any structure in D by pymatgen's `StructureMatcher` on emitted geometry (Appendix C). ρ_{pre} is the fraction of the generation batch with $V = N_D = 1$; after shared prerelaxation of up to n such crystals (n=1000 in the rankings, Section 4.1), ρ_{post} is the fraction of that subset still novel versus D. The factor $\rho = \rho_{\text{pre}} \cdot \rho_{\text{post}}$ penalizes invalid outputs and memorized copies in both scores. Crystals failing the post-relaxation novelty filter are dropped; the remainder is the evaluated subset (Appendix C), and all stability and kernel readouts use it only.

Stability panel. A probe is a binary stability check on one crystal structure x. Write $S_T = 1$ if x passes probe T and $S_T = 0$ otherwise. The stability panel is a finite set $\mathcal{T}$ of probes. GEMBench's default is

$$\mathcal{T} = \{\text{Phonon},\ \text{Stable}_{50},\ \text{Stable}_{300},\ \text{Stable}_{600},\ \text{EaH}_0,\ \text{EaH}_{0.1}\}. \tag{1}$$

Pass criteria for each $T \in \mathcal{T}$ are in Section 3.2. At probe T, the accepted set is $\{x : S_T(x) = 1\}$, with cardinality m_T.

Training-data blocks. A training-data block is a cohort sharing novelty reference D and kernel calibration (Section 3.3), with rankings reported only within the block (Section 4.3). Three blocks are listed in Section 4.2.

3.2 PROBE PASS CRITERIA

Like widely used S.U.N.-style yields, many pipelines reduce crystal stability to a single energy-above-hull bit (Zeni et al., 2025; Betala et al., 2025). GEMBench instead evaluates each crystal under the six probes in (1): phonon stability, survival under short NPT molecular dynamics at three temperatures, and two energy-above-hull cuts. GEMBench freezes two MACE checkpoints (Batatia et al., 2023): MACE-small for every stability probe (shared prerelaxation, phonon forces, NPT MD, and hull energies against a frozen DFT Materials Project reference), and MACE-MP for node features in the embedding kernel only (Section 3.3, Appendix B.4). The two checkpoints are not interchangeable. All probes run after a shared cell relaxation with that MACE-small checkpoint, so generator and

dataset crystals are compared at a local minimum of the same potential. A static clash gate on the raw and relaxed geometry gates the panel: a failed crystal skips the remaining probes, so overlapping atoms cannot be relaxed into a passing structure during dynamics.

Phonon. The Phonon probe passes if no Brillouin-zone mesh mode of the relaxed crystal lies below -10^{-3} THz.

MD at 50, 300, 600 K (Stable$_{50}$, Stable$_{300}$, Stable$_{600}$). For each $T \in \{50, 300, 600\}$ K, probe Stable$_T$ passes if an NPT trajectory at T maintains hard integrity throughout (finite state, no internuclear clash below 0.50 Å, no volume or energy runaway) and, at 300 K and 600 K only, does not meet soft-melt criteria (liquid-like $g(r)$, diffusivity, or volume drift) after equilibration. Failure at a lower temperature stops the ladder.

Energy above hull (EaH$_0$, EaH$_{0.1}$). Probe EaH$_0$ passes if MACE-small relaxed energy per atom lies on or below a frozen DFT Materials Project hull (Jain et al., 2013) (< 0 eV/atom), and EaH$_{0.1}$ passes if it lies within 0.1 eV/atom of the hull.

Each probe is a distinct quality class (dynamical stability, survival at a given temperature, on-hull or metastable membership). The benchmark does not prioritize one class over another, so both scores use an unweighted mean over $\mathcal{T}$: a model that is strong on hull alone cannot dominate the mean. Full pass/fail thresholds, nesting of the MD and hull cuts, the dataset-rate floor that admits a probe into $\mathcal{T}$, the MD speedup retune, and catalog validation are given in Appendix A.

3.3 Similarity measure

Both aims in Section 1 require comparing *sets* of crystals under the stability panel. A generative aim is fully specified only once a crystal-similarity measure is fixed: it must capture the geometry of crystal sets. One could define a crystal space and a metric on it. We instead specify a positive-definite *crystal-similarity kernel* $\kappa(x, x')$ directly, which leaves more flexibility in metric design while supporting set-level statistics such as MMD and Vendi scores. Stability probes score one crystal at a time. They cannot tell whether two accepted sets follow the same distribution, or whether a set has collapsed onto a few motifs. For those questions we use κ.

Before trusting any kernel as a score term, we calibrate it against a frozen suite of controlled mode collapses along nine named margins (atom count, element count, crystal system, density, cell anisotropy, minimum internuclear distance, packing fraction, cell metric, and composition statistics), plus null-collapse tests that must read as similar. Suite construction, estimators, and fitted parameters are in Appendix B. The fit pool is the training dataset of each training-data block.

Embedding kernel. The embedding kernel uses a separate frozen MACE-MP network (Batatia et al., 2023) to map each crystal to the mean of its final-layer node features (Section 3.2, Appendix B.4). After pool standardization,

$$\kappa_{\mathrm{embed}}(x, x') = w \, \exp\!\big(-\|z(x) - z(x')\|^2/(2\ell^2)\big), \tag{2}$$

with lengthscale ℓ fixed at the pool median heuristic and scale w fit by bagged π^2 least squares (Appendix B). We refer to κ_{embed} as the *embedding kernel* throughout. Ranking tables label this column *Embedding*. Even after scaling w, this kernel tracks continuous geometry only weakly and misses several discrete and composition collapses, so we accompany it with a hand-crafted physical kernel.

Physical kernel. Each crystal is encoded by size, symmetry, packing, cell-geometry, and composition channels. Each channel is a Gaussian on a pool-standardized encoding, and

$$\kappa_{\mathrm{phys}}(x, x') = \sum_j w_j \kappa_j(x, x'), \qquad w_j \geq 0. \tag{3}$$

Form agreement in matching and mode counting in discovery both use the same κ, so the kernel is a shared similarity prior for both scores. The two kernels are reported as separate score columns. Ranking tables also give a Rank column that is the mean of the two per-kernel competition ranks (Section 4.3).

3.4 THE MATCHING SCORE (GEM-match)

Under the matching aim (Section 1), GEM-match operationalizes agreement with dataset D in pass fractions and in the shape of each accepted set under κ, with a penalty for copying D. For generator G and dataset D, write

$$\rho = \rho_{\text{pre}} \cdot \rho_{\text{post}}, \qquad \gamma_T = \Pr_G(S_T = 1 \mid V = N_D = 1), \qquad \beta_T = \Pr_D(S_T = 1), \tag{4}$$

where ρ is the novelty rate (invalid outputs and copies of D count against it, including matches that appear only after MACE-small prerelaxation, Appendix C, Table 9), γ_T is the generator pass rate on the evaluated novel subset, and β_T is the dataset pass rate. Rate agreement is

$$s_T = \text{clip}\left(1 - \frac{|\beta_T - \gamma_T|}{\beta_T}, 0, 1\right), \tag{5}$$

so $s_T = 1$ exactly when $\gamma_T = \beta_T$. Both under-stability and over-stability are punished.

Form agreement uses the kernel maximum mean discrepancy (MMD) of Gretton et al. (Gretton et al., 2012): a two-sample distance that is zero when the two laws agree under κ and grows as they separate. For laws P and Q on crystals,

$$\text{MMD}^2_\kappa(P, Q) = \mathbb{E}_{x,x' \sim P}[\kappa(x, x')] + \mathbb{E}_{y,y' \sim Q}[\kappa(y, y')] - 2\,\mathbb{E}_{x \sim P,\, y \sim Q}[\kappa(x, y)]. \tag{6}$$

For each probe T, under whichever of κ_{embed} or κ_{phys} is being scored, the bounded form term is

$$f_T = 1 - \text{clip}\big(\text{MMD}^2_\kappa(P_T, Q_T), 0, 1\big), \tag{7}$$

where P_T is the empirical law of generated crystals with $V = N_D = S_T = 1$, Q_T is the empirical law of dataset crystals with $S_T = 1$, and $f_T = 0$ for an empty accepted set. Because each kernel was calibrated to catch the mode collapses in Section 3.3, a collapse or shift of the accepted set is punished at a severity fixed before any generator was scored.

$$\boxed{\text{GEM-match} = \rho \cdot \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} s_T\, f_T.} \tag{8}$$

The product is the combination rule: rate and form must both match. A failure of either zeros the per-probe term. As a concrete contrast, a generation that matches the dataset pass rate β_T but collapses onto one motif gets $s_T \approx 1$ and $f_T \approx 0$, so the term is zero. A generation whose accepted set agrees in shape with the dataset but with $\gamma_T \ll \beta_T$ is limited by s_T. Keeping rate and form separate makes that distinction explicit (a single MMD on unequal set sizes would embed an implicit weighting between yield and shape). Crystal generators can copy structures from D, and the `StructureMatcher` novelty filter detects those copies, so GEMBench multiplies matching by ρ as an anti-overfitting factor: $\rho = 0$ scores 0.

3.5 THE DISCOVERY SCORE (GEM-disc)

Under the discovery aim (Section 1), GEM-disc operationalizes how many distinct novel stable modes a generator captures per evaluated sample, with copies of D excluded throughout. Distinctness is measured by the Vendi score (Friedman & Dieng, 2023): if λ are the eigenvalues of the kernel Gram K/m on an accepted sample of size m, the effective species count is

$$\text{VS} = \exp\Big(-\sum_i \lambda_i \log \lambda_i\Big). \tag{9}$$

The Gram uses a frozen transformation of the matching kernel (κ_{embed} or κ_{phys}). It is rescaled to similarity range $[0, 1]$, and for κ_{phys} its overall contrast is raised so that a typical crystal separates from most of a reference pool (frozen before any generator is scored, Appendix B.6). A Dirac mass gives $\text{VS} = 1$. Fully dissimilar crystals give $\text{VS} = m$. Families of near-duplicates lie in between. For each probe T, write VS_T for that count on the accepted set at T (novel versus D and $S_T = 1$), or 0 if the set is empty, and let n be the evaluated-subset size (Section 4.1). Then

$$\boxed{\text{GEM-disc} = \rho \cdot \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \frac{\text{VS}_T}{n},} \tag{10}$$

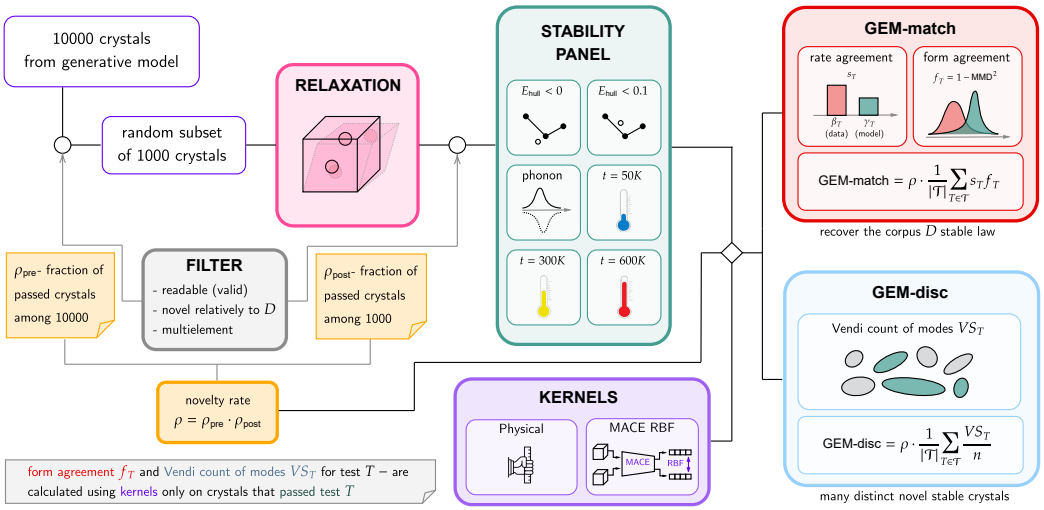

Figure 1: GEMBench pipeline for one generator G: novelty filtering and evaluated subset (N=10,000, n=1000), shared stability panel, and separate GEM-match / GEM-disc readouts under physical and embedding kernels.

with the same ρ as in GEM-match. Under the discovery aim (Section 1), success means filling the evaluation budget with novel crystals that pass the stability checks while keeping the passing set spread out under κ, not concentrated on a few modes. Equation (10) encodes that logic probe by probe: passing T is the gate ($S_T = 1$), mutual dissimilarity among accepted crystals is VS_T, and division by n fixes the budget. The factors ρ and the mean over $\mathcal{T}$ add eligibility and novelty, and panel-wide breadth. Unlike S.U.N.-style yields, GEM-disc counts distinct modes under κ, not unique structures under `StructureMatcher`. GEM-disc is a fixed-budget statistic: compare generators only at the same N, n, kernel calibration, and stability protocol. An empty accepted set scores 0, and a fully collapsed one scores ρ/n.

4 Experiments

Models are scored as published (none were tuned on GEMBench), not as upper bounds under aim-aware selection.

4.1 Evaluation protocol

Every model uses the same panel, kernels, and combination rules (Section 3). Following Section 3.1, each checkpoint is scored on one generation batch (N=10,000) and an evaluated subset of up to n=1000 novel-eligible crystals after shared prerelaxation and the post-relaxation novelty filter (Appendix C). Only the combination rule (GEM-match versus GEM-disc) changes between the two scores. Both scores are reported under the physical kernel and the embedding kernel as separate columns (Section 3.3).

4.2 Training-data blocks

Section 3.1 defines a training-data block and the evaluated-subset construction. We use three such blocks below. Blocks are grouped by training support so that novelty and kernel fit match what each model was trained on. A checkpoint must be scored against a reference that covers its training data: a model trained on a superset of MP-20 can appear novel with respect to MP-20 while copying its own training set. Within Blocks 1 and 2 the kernels are fitted on that block's training dataset (Appendix B). Block 3 reuses the Alex-MP-20 packs. All samples come from public repositories at pinned checkpoints.

Block 1: MP-20. Eligible checkpoints require no training data beyond MP-20 (Xie et al., 2021): Crystalite (Hadži Veljković et al., 2026), DiffCSP++ (Jiao et al., 2024), MiAD (Okhotin et al., 2025), MatterGen (Zeni et al., 2025), DiffCSP (Jiao et al., 2023), FlowMM (Miller et al., 2024), CrystalDiT (Yi et al., 2026), Wyformer-DiffCSP++ (Kazeev et al., 2025; Jiao et al., 2024), SymmCD (Levy et al., 2025), and OMatG (Höllmer et al., 2025). Novelty, matching dataset D, and kernel-fit pool are MP-20 train.

Block 2: Alex-MP-20. MatterGen and MiAD trained on Alex-MP-20 (Zeni et al., 2025): novelty, D, and kernel fit use Alex-MP-20 rather than Block 1 packs.

Block 3: CustomDataset (demonstration). This block is a demonstration cohort for checkpoints whose pipelines use training data beyond Alex-MP-20. Generative pretraining may stay on Alex-MP-20 while auxiliary objectives (reward models, trajectory data, or other supervision) draw from a larger superset (e.g. MP-traj). We do not have a unified reference corpus that covers all such supervision, so this block scores novelty only against Alex-MP-20, the largest dataset shared across the compared checkpoints, and reuses Block 2's kernel packs. Memorization outside D is not penalized. The public examples are Chemeleon2 (Park & Walsh, 2026) and CrysJEPA (Liu et al., 2026). Physical $\mathrm{GEM\text{-}match}$ is uninformative here (Table 1: both entries 0.000): the form term clips to zero on every probe under κ_{phys}. Ranks restart in this block and are not compared with Blocks 1 or 2. Only within-block embedding readouts are informative.

4.3 RANKINGS

Tables 1 and 2 report the two rankings, split by block. Within each block we sort models on each kernel column (higher score is better), assign competition ranks with tie bands of width $t = 2\hat{\sigma}$, and report Rank as the mean of the physical and embedding ranks (not aim fusion). Here $\hat{\sigma}$ is the subset-resampling standard deviation of scores on MatterGen (Appendix C). The value t is that transferred resolution band, not a per-model standard error (Section 4.5). Rows are ordered by mean Rank, with the physical score as tie-break. High $\mathrm{GEM\text{-}match}$ means the generation batch recovers the dataset's stable rates and shapes. High $\mathrm{GEM\text{-}disc}$ means many distinct novel stables per attempt. Do not average the two scores. Under the MP-20 and Alex-MP-20 headers of Tables 1 and 2, the D row scores dataset D as a generator with every crystal treated as novel ($\rho = 1$). It is not ranked.

Table 1: Matching ranking ($\mathrm{GEM\text{-}match}$) by training-data block under the physical and embedding kernels. The D row scores dataset D as a generator with $\rho = 1$ (not ranked). Ties: $t = 0.058$ (physical), $t = 0.047$ (embedding).

Model	Physical	Embedding	Rank
MP-20 (D = MP-20 train)			
D	1.000	1.000	—
MiAD	0.524	**0.632**	**1.5**
DiffCSP++	0.519	**0.669**	**1.5**
Crystalite	**0.618**	0.477	2.0
Wyformer-DiffCSP++	0.498	0.550	2.0
SymmCD	0.376	0.580	2.5
MatterGen	0.469	0.195	3.0
FlowMM	0.478	0.111	3.5
DiffCSP	0.419	0.217	3.5
CrystalDiT	0.211	0.487	3.5
OMatG	0.365	0.033	4.5
Alex-MP-20 (D = Alex-MP-20)			
D	1.000	1.000	—
MiAD	**0.490**	**0.545**	**1.0**
MatterGen	**0.485**	0.397	1.5
CustomDataset (D = Alex-MP-20)			
Chemeleon2	**0.000**	**0.043**	**1.0**
CrysJEPA	**0.000**	**0.010**	**1.0**

Table 2: Discovery ranking (GEM-disc $= \rho \cdot \mathrm{mean}_T(\mathrm{VS}_T/n)$) by training-data block. The D row scores dataset D as a generator with $\rho = 1$ (not ranked). Ties: $t = 0.004$ (physical), $t = 0.008$ (embedding).

Model	Physical	Embedding	Rank
MP-20 (D = MP-20 train)			
D	0.103	0.012	—
DiffCSP	0.107	0.051	**2.0**
MatterGen	0.106	0.053	**2.0**
DiffCSP++	**0.121**	0.015	2.5
FlowMM	0.087	**0.095**	2.5
SymmCD	0.107	0.015	3.0
OMatG	0.082	**0.091**	3.0
Wyformer-DiffCSP++	0.099	0.016	3.5
Crystalite	0.091	0.030	3.5
CrystalDiT	0.077	0.016	5.0
MiAD	0.065	0.016	5.5
Alex-MP-20 (D = Alex-MP-20)			
D	0.100	0.011	—
MatterGen	**0.095**	**0.020**	**1.0**
MiAD	0.074	**0.012**	1.5
CustomDataset (D = Alex-MP-20)			
Chemeleon2	**0.075**	**0.064**	**1.0**
CrysJEPA	0.042	0.006	2.0

Figure 2 (Appendix G) shows weak correlation between GEM-match and GEM-disc on the MP-20 block under either kernel: the best matching checkpoints are not the best on discovery, so the relevant column depends on the declared aim and on the kernel.

MiAD and DiffCSP++ share the best matching Rank (1.5), while DiffCSP and MatterGen share the best discovery Rank (2.0). FlowMM leads embedding GEM-disc (0.095) and DiffCSP++ leads physical GEM-disc (0.121). MiAD is last on discovery (Rank 5.5). Crystalite leads physical GEM-match (0.618) at matching Rank 2.0 and discovery Rank 3.5. CrystalDiT has the lowest physical GEM-match (0.211, ρ=0.615) because its on-hull accepted set disagrees with the dataset under κ_{phys}, while its embedding GEM-match (0.487) is Rank 3.5. Appendix D shows high panel yield lost to form disagreement or low Vendi. On Alex-MP-20, physical matching scores tie under t, MiAD leads embedding GEM-match, and MatterGen leads GEM-disc. In the CustomDataset block, physical GEM-match is degenerate, embedding matching scores tie under t, and Chemeleon2 leads GEM-disc.

4.4 Relation to S.U.N.

Table 3 reports CHGNet S.U.N. and M.S.U.N. on the same ten MP-20 generation batches from public model repositories, evaluated with the MatterGen S.U.N. protocol (Zeni et al., 2025). Same-batch MACE-small yields at those hull cuts (Appendix D, Table 12) keep the same top three (MiAD, CrystalDiT, Crystalite), so the GEM-disc reorder is not only an MLIP swap.

MiAD leads S.U.N. and is last on GEM-disc. CrystalDiT is second on S.U.N. but only Rank 5.0 on GEM-disc. High S.U.N. can be near-duplicates past a hull cut, whereas GEM-disc rewards distinct modes and GEM-match rewards dataset agreement, so S.U.N. is a hull-yield reference rather than either GEMBench aim. Appendix F keeps matching leaders near the top on hull cuts alone, but discovery becomes a multi-model tie that still excludes the S.U.N. leader.

4.5 Sampling variability

Rankings are informative only where score differences exceed resolution $t = 2\hat{\sigma}$ (Section 4.3). At a smaller protocol (N=1000, n=100, Table 10, Appendix C), physical GEM-match noise exceeds the leading gap and would collapse into one tie class. At N=10,000, n=1000, t lies at or below the

Table 3: CHGNet S.U.N. and M.S.U.N. on the same ten MP-20 generation batches as Table 1. S.U.N. uses $E_{\mathrm{hull}} < 0$ and M.S.U.N. uses $E_{\mathrm{hull}} < 0.1$ eV/atom under the published CHGNet protocol (GEMBench's stability panel uses MACE-small, Section 3.2). Last two columns copy Rank from Tables 1 and 2. Rows ordered by mean S.U.N./M.S.U.N. rank.

Model	S.U.N. (%)	M.S.U.N. (%)	GEM-match rank	GEM-disc rank
MiAD	**13.2**	55.9	**1.5**	5.5
Crystalite	11.0	54.3	2.0	3.5
CrystalDiT	12.3	44.5	3.5	5.0
MatterGen	8.3	**56.0**	3.0	**2.0**
DiffCSP++	9.0	43.2	**1.5**	2.5
DiffCSP	8.6	48.5	3.5	**2.0**
SymmCD	9.0	39.1	2.5	3.0
FlowMM	6.8	40.4	3.5	2.5
Wyformer-DiffCSP++	7.9	37.9	2.0	3.5
OMatG	4.0	38.9	4.5	3.0

top-five gaps in every column. GEM-disc is incomparable across n, and t is transferred from one generator, not a per-model standard error.

5 Conclusion

Our experiments show that distribution matching and candidate discovery can lead to different assessments of the same crystal generators even when the generation batch, novelty definition, stability protocol, and crystal-similarity measure are held fixed. On MP-20, GEM-match and GEM-disc rank the same published checkpoints differently, indicating that performance under one objective does not imply performance under the other.

The numerical results also depend on how similarity between crystals is represented: the physical and embedding kernels can produce different scores and per-kernel rankings. We therefore report the similarity representation explicitly rather than treating it as an implicit implementation choice.

Because GEM-match and GEM-disc represent different evaluation objectives, combining them into a single score would require an explicit, task-dependent trade-off that GEMBench does not assume. We therefore report the two objectives separately. The reported results characterize the evaluated checkpoints rather than the full capacity of the corresponding model families, since none of the models was selected or tuned for GEMBench. The conclusions are further conditional on the chosen stability probes, similarity kernels, and sampling budget.

5.1 Limitations

Stability is an MLIP screening proxy, not experimental or DFT proof, and the short NPT runs do not establish melting: the 50 K stage is an integrity screen (Appendix A.5). The collapse suite covers nine marginal families, not arbitrary joint collapse, and GEM-disc is comparable only at matched N, n, kernels, and panel. Reported uncertainty is evaluated-subset sampling plus a transferred t, not generator seed, MLIP, or MD noise. Evidence is the ten-model MP-20 block, with Alex-MP-20 a two-checkpoint case study.

AI USE STATEMENT

The text of the paper was polished for grammar and style using LLMs. We have reviewed all AI-assisted text and take responsibility for the final content of this work, including text, claims, and artifacts.

ETHICS STATEMENT

GEMBench is an evaluation protocol for crystal generative models. It does not release new datasets of hazardous materials, nor does it provide synthesis instructions. Benchmark scores rely on machine-learning interatomic potentials and short molecular-dynamics screens as computational proxies. They are not experimental stability claims. As with other generative materials tools, improved discovery metrics could accelerate identification of useful or harmful compounds if paired with experimental pipelines. We encourage responsible use and independent validation before any laboratory work.

REPRODUCIBILITY STATEMENT

Protocol details, stability-probe tuning, kernel calibration, ranking tie rules, and per-probe diagnostics are documented in the appendix. We will release code and evaluation configs to facilitate reproduction of the reported scores.

REFERENCES

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, et al. A foundation model for atomistic materials chemistry, 2023.

Siddharth Betala, Samuel P Gleason, Ali Ramlaoui, Andy Xu, Georgia Channing, Daniel Levy, Clémentine Fourrier, Nikita Kazeev, Chaitanya K Joshi, Sékou-Oumar Kaba, et al. LeMat-GenBench: A unified evaluation framework for crystal generative models. *arXiv preprint arXiv:2512.04562*, 2025.

Pierre-Paul De Breuck, Hai-Chen Wang, Gian-Marco Rignanese, Silvana Botti, and Miguel AL Marques. Generative AI for crystal structures: a review. *arXiv preprint arXiv:2509.02723*, 2025.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

Dan Friedman and Adji Bousso Dieng. The Vendi Score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Tin Hadži Veljković, Joshua Rosenthal, Ivor Lončarić, and Jan-Willem van de Meent. Crystalite: A lightweight transformer for efficient crystal modeling, 2026.

Xiao-Qi Han, Ze-Feng Gao, Wen-Kao Li, Peng-Jie Guo, and Zhong-Yi Lu. PhononBench: A large-scale phonon-based benchmark for dynamical stability in crystal generation. *arXiv preprint arXiv:2512.21227*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Philipp Höllmer, Thomas Egg, Maya Martirossyan, Eric Fuemmeler, Zeren Shui, Amit Gupta, Pawan Prakash, Adrian Roitberg, Mingjie Liu, George Karypis, Mark Transtrum, Richard Hennig, Ellad B. Tadmor, and Stefano Martiniani. Open materials generation with stochastic interpolants. In *Forty-second International Conference on Machine Learning*, 2025.

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):011002, 2013.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17464–17497, 2023.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *The Twelfth International Conference on Learning Representations*, 2024.

Nikita Kazeev, Wei Nong, Ignat Romanov, Ruiming Zhu, Andrey E Ustyuzhanin, Shuya Yamazaki, and Kedar Hippalgaonkar. Wyckoff transformer: Generation of symmetric crystals. In *Forty-second International Conference on Machine Learning*, 2025.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Kin Long Kelvin Lee, Mikhail Galkin, Santiago Miret, and Siamak Ravanbakhsh. Symmcd: Symmetry-preserving crystal generation with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Nian Liu, Nikita Kazeev, Stephen Gregory Dale, Artem Maevskiy, Yuwei Zeng, Ryoji Kubo, Pengru Huang, Thomas Laurent, Yann LeCun, Kostya S. Novoselov, and Xavier Bresson. Crys-JEPA: Accelerating crystal discovery via embedding screening and generative refinement, 2026.

Antoine Loew, Dewen Sun, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Universal machine learning interatomic potentials are ready for phonons. *npj Computational Materials*, 11 (1):178, 2025.

Maya M Martirossyan, Thomas Egg, Philipp Höllmer, George Karypis, Mark Transtrum, Adrian Roitberg, Mingjie Liu, Richard G Hennig, Ellad B Tadmor, and Stefano Martiniani. All that structure matches does not glitter. In *Advances in Neural Information Processing Systems*, 2025.

Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching. In *International Conference on Machine Learning*, pp. 35664–35686. PMLR, 2024.

Masahiro Negishi, Hyunsoo Park, Kinga Oliwia Mastej, and Aron Walsh. Continuous SUN (stable, unique, and novel) metric for generative modeling of inorganic crystals. *Machine Learning: Science and Technology*, 7(3):035064, 2026.

Andrey Okhotin, Maksim Nakhodnov, Nikita Kazeev, Andrey E Ustyuzhanin, and Dmitry Vetrov. MiAD: Mirage atom diffusion for de novo crystal generation. *arXiv preprint arXiv:2511.14426*, 2025.

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, 2013.

Hyunsoo Park and Aron Walsh. Guiding generative models to uncover diverse and novel crystals via reinforcement learning. *Nature Machine Intelligence*, 2026. doi: 10.1038/s42256-026-01262-4.

Hyunsoo Park, Zhenzhu Li, and Aron Walsh. Has generative artificial intelligence solved inverse materials design? *Matter*, 7(7):2355–2367, 2024.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, 2018.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, 2016.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2021.

Xiaohan Yi, Guikun Xu, Zhong Zhang, Liu Liu, Yatao Bian, Xi Xiao, and Peilin Zhao. CrystalDiT: Simple diffusion transformers for crystal generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1462–1470, 2026.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

A STABILITY PROBES: TUNING AND VALIDATION

Section 3.2 states the six probe pass criteria and the clash gate. This appendix states the full pass/fail logic and numeric thresholds used in GEMBench. Implementation backends (integrators, batching, randomness) are omitted. Every bound below is frozen in the benchmark protocol.

A.1 PROTOCOL OVERVIEW

Each crystal is scored by six binary probes after a shared geometry preparation step. A separate static *clash gate* is not a member of the panel $\mathcal{T}$: a failed gate skips all remaining probes and zeros every flag, so overlapping atoms cannot be relaxed into a passing structure during dynamics.

The default panel is

$$\{\text{Phonon},\ \text{Stable}_{50},\ \text{Stable}_{300},\ \text{Stable}_{600},\ \text{EaH}_0,\ \text{EaH}_{0.1}\}.$$

Only probes with dataset pass fraction $\beta_T > 0.05$ enter $\mathcal{T}$. Otherwise rate and form comparisons are too noisy at the evaluation subset sizes used in the paper. Each probe defines a distinct material-quality class. The benchmark does not prioritize one class over another, so both scores use an unweighted mean over $\mathcal{T}$. The probes are partly nested: the MD stages form a temperature ladder (failure at a lower temperature stops higher temperatures), and $\text{EaH}_{0.1}$ is implied by EaH_0.

All probes use the same frozen MACE-small checkpoint for forces and energies (Section 3.2). Crystals whose chemistry lies outside the MLIP element table fail every probe.

A.2 SHARED RELAXATION AND THE CLASH GATE

Every probe except phonon starts from a common *prerelaxation*: the input cell is relaxed with a FIRE optimizer on MACE-small until the maximum atomic force falls below 0.05 eV/Å (cap 500 steps), relaxing the cell volume and shape whenever the structure is a bulk 3D crystal. Phonon stability applies an additional, tighter relaxation to 0.005 eV/Å before computing force constants.

Immediately after prerelaxation, a deterministic static clash gate runs on both the raw input and the relaxed geometry:

- **Minimum internuclear distance** below 0.7 Å on the raw input or on the relaxed structure.
- **Maximum relaxed force** above 1.0 eV/Å.

Either condition fails the gate. Non-finite positions, cell parameters, or volume after prerelaxation also fail the gate. A gate failure assigns $S_T = 0$ for every probe and terminates the panel for that crystal.

A.3 PHONON STABILITY

Phonon stability asks whether the relaxed crystal has a real vibrational spectrum. Force constants are computed by finite displacements (0.01 Å) on a supercell of size $3 \times 3 \times 3$ for bulk crystals or $5 \times 5 \times 1$ for vacuum-padded 2D cells. Frequencies are evaluated on a Brillouin-zone mesh of $30 \times 30 \times 30$ points (3D) or $50 \times 50 \times 1$ (2D).

The crystal **passes** if the minimum mesh frequency is at or above -10^{-3} THz, equivalently, no mesh mode lies more than 10^{-3} THz below zero. This threshold was not retuned on the known-solid catalogs in Section A.7. It is the conventional imaginary-mode cutoff used throughout the benchmark.

A.4 ENERGY ABOVE HULL

Thermodynamic stability is measured relative to a frozen DFT Materials Project convex hull (Jain et al., 2013). The MLIP relaxed energy per atom after the shared prerelaxation is compared to the hull energy of the same composition:

$$\text{EaH} = E_{\text{MLIP}}/N_{\text{atoms}} - E_{\text{hull}}^{\text{DFT}}.$$

Two cuts define the official probes:

- EaH_0: pass if EaH < 0 eV/atom (on or below the hull).
- $\text{EaH}_{0.1}$: pass if EaH < 0.1 eV/atom (metastable window).

Non-finite energies or missing hull entries fail both probes. Hull thresholds were not retuned on the known-solid catalogs.

A.5 FINITE-TEMPERATURE MD

A.5.1 SETUP AND LADDER

Each MD probe runs NPT dynamics at approximately 1 atm on an integer-replicated supercell sized to 60–128 atoms. The integration timestep is 1 fs. Trajectories are sampled every 1 ps. Three temperatures form a kinetic ladder (50, 300, and 600 K), and failure at a lower temperature stops the ladder. The screen is not a melting-point solver.

Per-temperature protocol lengths and equilibration grace periods (during which only hard integrity applies) are:

Temperature	Duration	Grace	Soft melt votes	$D_{\min}$ (Å^2/ps)
50 K	6 ps	2 ps	off	—
300 K	12 ps	3 ps	on	0.15
600 K	20 ps	5 ps	on	0.05

Here $D_{\min}$ is the diffusivity threshold used in the soft melt rule below. At 300 K the diffusivity cut is relaxed relative to the 600 K default so that soft metals are not rejected by a single persistent diffusivity vote on sub-melting packings. True melts at 300 K still exceed this bound.

A.5.2 HARD INTEGRITY

Hard integrity guards are active at every sample, including during the equilibration grace. They detect numerical or packing collapse, not melting:

- **Internuclear clash:** minimum interatomic distance below 0.50 Å at any sample (distinct from the static 0.7 Å gate).
- **Non-finite state:** non-finite atomic positions, cell, volume, or potential energy.
- **Catastrophic volume runaway:** relative volume change $|V - V_0|/V_0 > 5.0$ versus the reference cell at MD start, including during warmup when defined.
- **Post-equilibration volume explosion:** after the grace period, $|V - V_{\text{warmup}}|/V_{\text{warmup}} > 0.50$, where V_{warmup} is the mean cell volume over the grace window.
- **Energy excursion:** $|\Delta E|/N_{\text{atoms}} > 5.0$ eV/atom relative to the initial MD energy.

Any hard failure ends the trajectory at that temperature and marks $\text{Stable}_T = 0$.

At 50 K only hard integrity is evaluated. Soft melt votes at 50 K were dropped because they reshuffled on marginal packings under resampling while the reproducible 50 K failures were already caught by hard guards. The 50 K stage is therefore an integrity screen, with kinetic melt verdicts deferred to 300 K and 600 K.

A.5.3 SOFT MELT VOTES

After the equilibration grace at 300 K and 600 K, three soft criteria vote for melting. Each is evaluated on post-grace samples only:

1. **Wendt–Abraham order parameter.** Let d_{nn} be the mean nearest-neighbor distance in the reference frame at MD start. In the radial shell from $0.7\,d_{\text{nn}}$ to $1.4\,d_{\text{nn}}$, compute $R = g(r_{\text{first min}})/g(r_{\text{first max}})$ on the instantaneous $g(r)$. A **vote fires** if $R > 0.50$ (liquid-like loss of long-range order).

2. **Diffusivity.** Atomic displacements are accumulated with minimum-image conventions and corrected for homogeneous cell deformation under NPT. The mean-squared displacement $\mathrm{MSD}(t)$ is fit linearly for t at or after the grace time. The diffusivity is $D = \mathrm{slope}(\mathrm{MSD})/6$. A **vote fires** if $D > D_{\min}$, with $D_{\min}$ from the table above (0.15 at 300 K, 0.05 at 600 K).
3. **Volume drift.** Relative to the post-warmup reference volume V_{warmup}, a **vote fires** if $|V - V_{\mathrm{warmup}}|/V_{\mathrm{warmup}} > 0.08$.

The trajectory **fails** the soft melt rule if either condition holds:

- at least **two** of the three criteria fire on the same 1 ps sample, or
- **one** criterion fires on **three consecutive** 1 ps samples (persistence over ≈ 3 ps).

The default vote quorum is two simultaneous votes. The persistence rule catches slow melts that activate only one signal at a time.

For structures without a spanning covalent or ionic framework (molecular crystals, vdW solids, and other multi-cluster packings), the volume-drift and diffusivity thresholds may be scaled upward by a configurable factor. The frozen benchmark uses scale 1.0 (no adjustment).

A.5.4 DIAGNOSTICS

Structural diagnostics (RMSD, bond retention, coordination numbers, and thermal drift metrics) are recorded along each trajectory but *do not* enter pass/fail. They support offline audits only.

A.6 FROM ACCURATE TO TRACTABLE

The starting MD procedure used a heavier MLIP, higher-precision arithmetic, larger supercells, and longer trajectories and grace periods than the frozen screen. That setup separated melting-temperature bands well but was too slow for scoring thousands of generated crystals under the proposed budget. We therefore retuned compute while holding the scientific criterion fixed: keep separating solids that fail at 50 K, pass 50 K and fail 300 K, pass 300 K and fail 600 K, or pass all three. The locked lightweight screen uses MACE-small, reduced-precision force evaluation, smaller supercells (order 60–128 atoms), and the shortened NPT durations and equilibration grace periods in the table above. Taken together these choices are about $25\times$ faster than the procedure we started from, at the cost of accepting residual errors concentrated on cryogens and soft molecular solids (Section A.7). Phonon meshes and hull cuts were left at their conventional thresholds. Only the MD kinetic screen was retuned for speed.

A.7 CATALOGS AND ACCURACY

Known elemental and binary solids from different structural families were grouped by melting-temperature band (fail at 50 K, pass 50 K and fail 300 K, pass 300 K and fail 600 K, pass all three) and split into two disjoint 100-crystal catalogs: one for tuning criteria and compute budget, and one as hold-out validation of the frozen screen. At each temperature the ground-truth label is whether the solid is expected to survive that stage of the melting ladder, and the screen prediction is whether the frozen NPT criteria pass. Tables 4 and 5 report the corresponding confusion counts (TP, TN, FP, FN) on the tuning and hold-out catalogs. Overall accuracies are 0.93 / 0.90 / 0.99 (tuning) and 0.84 / 0.88 / 0.97 (hold-out) at 50 / 300 / 600 K.

Table 4: Confusion counts for the NPT screen on the tuning catalog (100 known solids). TP/TN: screen agrees with the melting-band label (survive / fail). FP: screen passes a solid expected to fail. FN: screen fails a solid expected to survive.

Temperature	TP	TN	FP	FN
50 K	70	23	2	5
300 K	41	49	1	9
600 K	25	74	1	0

Table 5: Confusion counts for the NPT screen on the hold-out catalog (100 known solids, disjoint from Table 4). Columns as in Table 4.

Temperature	TP	TN	FP	FN
50 K	66	18	7	9
300 K	39	49	1	11
600 K	24	73	2	1

Errors concentrate on cryogens (Ne, Ar, H_2, He) and on soft molecular/vdW solids (S, I_2, Cl_2, Br_2), while refractory networks (diamond-like, rocksalt carbides/nitrides, close-packed metals) are recovered. Superheating near $T_{\mathrm{m}} \approx 600$ K limits the 600 K stage and is accepted rather than repaired by a longer trajectory. The screen is a lightweight filter for the accepted set, not a melting-point predictor.

B CRYSTAL KERNELS: CALIBRATION

Section 3.3 names the two kernels and the collapse-calibration idea. Section 3.4 defines MMD and the form readout (7) used as the matching form term. This appendix freezes the suite that implements that test, the estimators (including bagged log-ridge / π^2 least squares), and the fitted parameters. None of these quantities is refit on generator outputs.

B.1 FROZEN SUITE

The pool is 2500 crystals drawn from MP-20 train (Xie et al., 2021). Each collapse test compares two index sets of size $|P| = |Q| = 200$. The bagged calibration suites used below contain 5 null-collapse tests and 180 collapse tests: nine ladders at four grades, with continuous cores replicated $n_{\text{rep}} = 5$ times inside each suite.

On that pool, each collapse test mixes the pool law F with an extreme core L,

$$P_\pi = (1 - \pi)F + \pi L, \tag{11}$$

so that for any kernel $\mathrm{MMD}^2(P_\pi, F) = \pi^2\,\mathrm{MMD}^2(L, F)$: the mixture weight sets a definite severity target. A collapse test for ladder ℓ and grade $f \in \{0.75, 0.50, 0.25, 0.05\}$ uses (11) with locked core L_ℓ: Q is a pool subsample (a draw from F), and P is a finite mixture with $\mathrm{round}(200\,\pi)$ draws from L_ℓ and the rest from the pool. Because $\mathrm{MMD}^2(P_\pi, F) = \pi^2 C_\ell$ with $C_\ell = \mathrm{MMD}^2(L_\ell, F)$ for any kernel, the mixture weight is set to

$$\pi^2 = \frac{1 - f}{0.95}. \tag{12}$$

Once channel weights (or the embedding scale) normalize C_ℓ to 0.95, every grade of that ladder has a definite form target. The cores themselves are extreme subsets of the pool:

- **Discrete ladders** (atom count, element count, crystal system): L is the modal value of that coordinate on the pool.
- **Scalar packing ladders** (density, cell anisotropy, minimum internuclear distance, packing fraction): L is a central sliver $|x - c| \le 0.015\,\mathrm{IQR}(x)$, with c drawn in the inter-quantile range $[q_{20}, q_{80}]$ (redrawn per replication).
- **Vector ladders** (cell metric, composition statistics): L is a k-nearest-neighbor ball ($k = 25$) in the pool-standardized feature (redrawn per replication). The ball centre is chosen among random candidates to minimize total-variation shift of the three discrete coordinates, so the ball isolates the vector channel rather than confounding a discrete ladder.

Joint (dependence) locks are not in the suite: at pool size 2500 the conditional groups are too small for a stable target.

B.2 ESTIMATOR AND FORM ERROR

MMD^2 is the unbiased U-statistic (Gretton et al., 2012). The clip in (7) absorbs the small negatives that a U-statistic can take under the null. The diagnostic reported for a frozen fit is the mean absolute form error (MAE) $\mathrm{mean}\,|f - \mathrm{target}|$ over collapse tests. It is not the selection objective for either default kernel (Sections B.3 and B.4).

B.3 PHYSICAL KERNEL

Each channel is a Gaussian on a pool-standardized encoding, as in (3) for κ_{phys}. Crystal system is encoded as the $\log$ of the lattice point-group order $(2, 4, 8, 16, 12, 24, 48)$ for triclinic through cubic. Atom count (and the other geometry channels) is taken on the *primitive* cell passed into the kernel: generated and reference crystals are represented as primitives before feature extraction, so the discrete atom-count ladder is not confounded by conventional or supercell representations. Lengthscales stay at the pool median heuristic (no discrete bandwidth grid). Because MMD^2 is linear in the channel weights, the fit targets the π^2 grade profile per ladder: a through-origin regression yields a sensitivity row $\widehat{C}_\ell$, and nonnegative weights are solved by log-ridge toward equal log-weights (ridge fraction 0.1

of $\mathrm{mean}\,\mathrm{diag}(\widehat{C}^\top \widehat{C})$), which keeps every channel strictly positive. To reduce dependence on a single core draw, each continuous ladder is replicated with $n_{\mathrm{rep}} = 5$ independent cores inside a suite, and the final weight vector is the average of $n_{\mathrm{bag}} = 5$ such log-ridge fits on suite redraws. Element-count receives the largest weight, as required by the π^2 scaling (12) of that ladder.

On the five bag-member suites the bagged κ_{phys} has mean MAE 0.184 (Table 6). Bagging was chosen for ranking stability of GEM-match under suite redraws.

B.4 Embedding kernel

Frozen MACE checkpoints. The stability panel uses one frozen MLIP for every probe (prerelaxation, phonon forces, NPT MD, and hull energies) so all pass/fail readouts refer to the same potential. That checkpoint is MACE-small: the NPT screen was retuned for tractability at fixed scientific criteria (Section A.6), and hold-out catalog accuracies remain 0.84–0.97 per stage (Section A.7). The embedding kernel is a separate design choice: MACE-MP-medium maps each crystal to node features under a widely used MACE representation, then κ_{embed} is calibrated on the collapse suites below. Reported benchmark times cover the object-level panel only (Table 10, Appendix C), so kernel evaluation is negligible relative to stability screening and we do not downscale the embedding backbone for speed. A future protocol revision could swap the embedding network or recalibrate kernels without revisiting the frozen stability thresholds.

The embedding is the mean of final-layer node features of that MACE-MP-medium network (Batatia et al., 2023). Features are pool z-scored (no ℓ_2 normalization). The Gaussian kernel

$$\kappa_{\mathrm{embed}}(x, x') = w\,\exp\!\big(-\|z(x) - z(x')\|^2/(2\ell^2)\big)$$

has lengthscale ℓ frozen at the pool median heuristic and a single free scale w, with MMD^2 linear in w. That scale is frozen by bagged π^2 least squares on the same replicated collapse suites used for κ_{phys} ($n_{\mathrm{rep}} = 5$, $n_{\mathrm{bag}} = 5$): a scalar fit, not the multi-channel log-ridge used for physical weights.

At $\ell \approx 22.27$ and $w \approx 15.92$ the bagged κ_{embed} has mean MAE 0.361 (Table 6). The Gaussian lengthscale ℓ is a bandwidth parameter on pool z-scored embeddings. It is distinct from the discovery Gram sensitivity σ in Appendix B.6, which is $\sigma = 0$ (identity) for κ_{embed}. It follows continuous geometric ladders more than discrete or compositional collapse: atom-count, element-count, crystal-system, and composition-statistic collapse tests remain poorly calibrated even after scaling, and composition-ball collapse still saturates MMD on many grades. That failure mode is why Section 3.3 introduces κ_{embed} first and then adds κ_{phys} as a complementary collapse meter rather than replacing MACE. The two kernels are therefore not interchangeable.

Table 6: Collapse-suite calibration on the frozen MP-20 pool (2500 crystals). MAE is mean $|\mathrm{form} - \mathrm{target}|$ over collapse tests on bag-member suites (180 collapse tests per member, $n_{\mathrm{bag}} = 5$, $n_{\mathrm{rep}} = 5$).

Kernel	MAE
Physical, bagged	0.184
Embedding, bagged ($w \approx 15.92$)	0.361

B.5 Stability under alternate collapse constructions

The official suite fixes one way to realize mild→severe collapse (mixture onto a locked core with forms $0.75/0.50/0.25/0.05$). To check that GEM-match rankings are not an artefact of that choice, the same bagged log-ridge procedure is refit on severity-matched constructions that change the *geometry* of the collapse while keeping a comparable punishment ladder: softer or harsher top-k mode unions, cores that never reach a single-mode Dirac, and a shifted mixture form grid. Table 7 reports dense ranks under $t = 2\hat{\sigma}$ for each construction. Crystalite remains rank 1 throughout. The leading group (DiffCSP++, MiAD, FlowMM, Wyformer) stays contiguous. Across the five severity-matched constructions (excluding a deliberately non-matched top+tail variant), only $2/100$ model×construction pairs produce disjoint position intervals, and the mean dense-rank spread is 0.60.

Table 7: Physical GEM-match ranks under severity-matched collapse constructions on the bagged log-ridge fit. "Original" is the official mixture suite. Other columns vary how mild-to-severe collapse is realized at comparable punishment.

Model	Original	Soft top-k	Harsh top-k	No-Dirac	Mix shifted
Crystalite	1.0	1.0	1.0	1.0	1.0
MiAD	2.0	3.0	3.0	2.0	2.0
DiffCSP++	2.0	2.0	2.0	2.0	2.0
Wyformer-DiffCSP++	2.0	2.0	2.0	2.0	2.0
FlowMM	2.0	2.0	3.0	2.0	3.0
MatterGen	3.0	3.0	4.0	3.0	3.0
DiffCSP	3.0	4.0	4.0	4.0	4.0
SymmCD	4.0	4.0	4.0	4.0	4.0
OMatG	4.0	4.0	4.0	3.0	5.0
CrystalDiT	5.0	5.0	5.0	4.0	6.0

B.6 Discovery Gram

Matching and discovery share the same calibrated kernel (κ_{embed} or κ_{phys}), then read its Gram differently (Section 3.5 states only the two frozen steps). This subsection fixes their parameters. The discovery kernel is a frozen transformation of the matching kernel κ, applied identically to every generator:

1. **Move to** $[0, 1]$**.** Divide by the self-similarity,

$$\kappa_0(x, y) = \frac{\kappa(x, y)}{\kappa(x, x)},$$

so that $\kappa_0(x, x) = 1$ and off-diagonal similarities lie in $[0, 1]$. For both kernels this divisor is a single constant shared by all crystals ($\kappa_{\text{embed}}(x, x) = w$ and $\kappa_{\text{phys}}(x, x) = \sum_j w_j$), so this step is a global rescaling of similarities, not a per-crystal one.

2. **Change the overall contrast.** Apply the optional sharpening

$$\kappa_F(x, y) = \exp\big(\sigma(\kappa_0(x, y) - 1)\big), \tag{13}$$

which controls the size of a typical mode under the e^{-1} cutoff $\kappa_F \geq e^{-1}$. The value $\sigma = 0$ recovers κ_0.

Both steps use only the already-frozen matching kernel and reference-pool geometry, so every model is scored with the same discovery kernel.

Why a mode-size map is needed. After bagged calibration, κ_{embed} already separates modes on the identity map ($\sigma = 0$): off-diagonal similarities are small, and Vendi of accepted sets is informative without further sharpening. κ_{phys}, by contrast, remains soft on the same pool: a typical crystal's e^{-1}-neighborhood covers about 97% of a calibration subsample at $\sigma = 0$, so nearly every pair looks "near" and Vendi collapses. The switch between kernels should then be a change of *which* geometric information is measured, not a change of the implicit mode-size convention.

Unified rule. We therefore fit σ by a mode-size cap that is kernel-agnostic: a typical pool crystal must be far from at least a fraction p of the pool, i.e. the median over x of

$$\frac{\#\{y \neq x : \kappa_F(x, y) \geq e^{-1}\}}{n - 1}$$

must be at most $1 - p$. For each of ten random pool subsamples of size 400, binary search finds the smallest such σ. The packaged value is the average over those draws. far95% ($p = 0.95$, median neighborhood $\leq 5\%$) is the default: σ is fit exclusively from reference-pool geometry and frozen before any generator is scored ($\sigma \approx 12.37$ on the MP-20 pack). far90% and far99% are sensitivity checks (Table 8): absolute scores move, but the physical-only order is stable (DiffCSP++ first, MiAD last). κ_{embed} keeps $\sigma = 0$ because the mode-size rule is already satisfied. κ_{phys} uses the far95% pack.

Alex-MP-20 packs. Block 2 of Section 4.2 scores models trained on Alex-MP-20, so the kernels are refit on that training set: the same bagged log-ridge / bagged π^2 / far95% protocol on a fresh draw of 2500 crystals from Alex-MP-20 train (Zeni et al., 2025) (separate packs, and the MP-20 fits above are not reused). Bag-member mean MAE is 0.147 for the physical kernel and 0.379 for the embedding kernel at $\ell \approx 21.89$, $w \approx 19.36$. The Alex far95% map gives $\sigma \approx 7.00$ for κ_{phys} only (κ_{embed} keeps $\sigma = 0$). Data-side spacegroups for Alex-MP-20 are computed with the same `SpacegroupAnalyzer` path as generated crystals and stored in the data caches (not read from CSV).

A median-heuristic bandwidth $\sigma = 1/(1 - m)$ (m the median off-diagonal of κ_0) was evaluated earlier and rejected for this soft physical Gram: it is too weak to meet any of the $10\%/5\%/1\%$ caps.

Table 8: Physical GEM-disc under three far-fraction mode-size maps on the MP-20 block. far95% (Appendix B.6) is the default. Order is stable across caps with DiffCSP++ leading and MiAD last.

	far90% (σ=8.98)		far95% (σ=12.37)		far99% (σ=23.47)	
Model	GEM-disc	Rank	GEM-disc	Rank	GEM-disc	Rank
DiffCSP++	0.078	1.0	0.121	1.0	0.250	1.0
SymmCD	0.071	2.0	0.107	2.0	0.212	4.0
DiffCSP	0.068	2.0	0.107	2.0	0.230	3.0
MatterGen	0.066	3.0	0.106	2.0	0.239	2.0
Wyformer-DiffCSP++	0.063	3.0	0.099	3.0	0.211	4.0
Crystalite	0.057	4.0	0.091	4.0	0.203	5.0
FlowMM	0.054	4.0	0.087	4.0	0.199	5.0
OMatG	0.051	5.0	0.082	5.0	0.185	6.0
CrystalDiT	0.049	5.0	0.077	6.0	0.162	7.0
MiAD	0.040	6.0	0.065	7.0	0.153	8.0

C REPRODUCIBILITY DETAILS

This appendix collects the protocol details behind the rankings (Section 4.3) and the variability study (Section 4.5).

Evaluation subsets. Each ranking starts from one evaluated subset per model: up to $n = 1000$ crystals drawn from the novel-eligible pool of the $N = 10{,}000$-crystal generation batch, then prerelaxed with MACE-small (Appendix A.2). Novelty versus D is assessed twice: on the emitted geometry (ρ_{pre}) and again on the prerelaxed geometry within this subset (ρ_{post}). The headline factor is $\rho = \rho_{\mathrm{pre}} \cdot \rho_{\mathrm{post}}$. Crystals that match D only after prerelaxation are dropped from the evaluated subset before stability rates, form, and Vendi are computed, so the final evaluated size can be strictly less than 1000 (e.g. CrystalDiT retains 884). Table 9 lists ρ_{pre}, ρ_{post}, and ρ for every ranked model. CrystalDiT is the outlier ($\rho_{\mathrm{post}} = 0.884$, 116/1000 leaks). Dataset D is subsampled at the same sizes, and the MP-20 data cache is drawn consistently with that subset.

Table 9: Novelty factor ρ by model: pre-relaxation rate ρ_{pre}, post-relaxation retention ρ_{post}, and product $\rho = \rho_{\mathrm{pre}} \cdot \rho_{\mathrm{post}}$. CrystalDiT loses 11.6% of pre-novel eval crystals after prerelax (116/1000).

Model	ρ_{pre}	ρ_{post}	ρ
MP-20			
Crystalite	0.822	0.999	0.821
DiffCSP++	0.974	1.000	0.974
MiAD	0.742	0.998	0.740
MatterGen	0.936	1.000	0.936
DiffCSP	0.925	0.999	0.924
FlowMM	0.909	0.982	0.893
CrystalDiT	0.696	<u>0.884</u>	0.615
Wyformer-DiffCSP++	0.973	1.000	0.973
SymmCD	0.952	0.994	0.946
OMatG	0.824	0.962	0.793
Alex-MP-20			
MiAD	0.681	1.000	0.681
MatterGen	0.730	0.999	0.729

Tie scale. The ranking resolution $t = 2\hat{\sigma}$ uses $\hat{\sigma}$ estimated on MatterGen: five evaluated subsets of the same 10,000-crystal generation batch, reusing stability labels and holding the default MP-20 data cache fixed. Under the bagged kernels and the far95% physical discovery map, the sample standard deviations are $\hat{\sigma} = 0.0288$ (physical GEM-match), 0.0234 (embedding GEM-match), 0.0022 (physical GEM-disc), and 0.0041 (embedding GEM-disc), giving $t = 0.058$ and 0.047 for GEM-match and $t = 0.004$ and 0.008 for GEM-disc. The scale is estimated once on one generator and transferred to all models.

Table 10: Sampling variability of MatterGen scores at two evaluation budgets (D = MP-20 train). Smaller protocol uses N=1000, n=100. Proposed protocol uses N=10,000, n=1000.

Setup	Benchmark time	GEM-match		GEM-disc	
	(h, 1 A100)	Physical	Embedding	Physical	Embedding
Smaller subsets (N=1000, n=100)					
MatterGen	≈6 h	0.426±0.140	0.163±0.105	0.367±0.008	0.191±0.016
Top-5 mean adjacent gap		0.033	0.265	0.016	0.124
Proposed protocol (N=10,000, n=1000)					
MatterGen	≈60 h	0.462±0.029	0.177±0.023	0.108±0.002	0.057±0.004
Top-5 mean adjacent gap		0.087	0.083	0.018	0.050

Variability table. MatterGen GEM-match in Table 10 is the mean and sample standard deviation over ten (evaluated-subset, data-cache) pairs sampled from the five subsets and the available caches

(five in the smaller block, three in the proposed one), with each subset and each cache appearing at least once. GEM-disc does not use the data cache. Its spread is over the five evaluated subsets. The smaller block uses N=1000 and n=100. The proposed block uses N=10,000 and n=1000. Both blocks use the bagged κ_{embed} and κ_{phys} kernels. Physical GEM-disc uses the far95% mode-size map in both blocks. The neighbor-gap row is the mean absolute adjacent difference among the top five MP-20 models of Tables 1 and 2 (table order), under the smaller protocol in the first block and taken from the ranking tables in the second. Reported benchmark times ($\approx$6 h and $\approx$60 h on one A100) cover the object-level panel at the two scales.

Novelty matcher. Novelty versus D (and uniqueness diagnostics) use pymatgen's `StructureMatcher` (Ong et al., 2013) with the library defaults `ltol`=0.2, `stol`=0.3, `angle_tol`=5°, `primitive_cell`=$True$, `scale`=$True$, and with `attempt_supercell`=$False$ and `symmetric`=$False$ as frozen GEMBench overrides. The first pass runs on generated geometries as emitted against the CIFs of dataset D. Invalid or unreadable outputs are scored non-novel. The second pass reruns the same matcher on the prerelaxed geometries of the evaluated subset. The product $\rho = \rho_{\mathrm{pre}} \cdot \rho_{\mathrm{post}}$ enters both scores, and structures failing the second pass are excluded from the stability and kernel readouts.

D PER-PROBE DIAGNOSTICS

The ranking tables report only the means over $\mathcal{T}$. Table conventions follow Section 4.3. This appendix records the six per-probe terms that enter those means, for every model in Tables 1 and 2, under the same generation batches (evaluated subset after the post-relaxation novelty filter, Table 9) and the same bagged kernels. In every table below, higher is better. In the MP-20 block, Tables 11, 12, 15, and 16 bold the column maximum among models (ties included, and the D row is excluded). Tables 13 and 14 underline the column minimum among models (ties included), because form failures are the clearer diagnostic when many f_T values sit near one.

Table 11 gives the kernel-independent rates γ_T (without ρ) and the dataset rates β_T. Table 12 multiplies the same γ_T by $\rho = \rho_{\mathrm{pre}} \cdot \rho_{\mathrm{post}}$ (Table 9). The EaH_0 and $\mathrm{EaH}_{0.1}$ columns of Table 12 are therefore novel-stable yields under GEMBench's MACE-small hull cuts on the same generation batches as Table 3 (without within-batch uniqueness), and Section 4.4 uses them as a same-potential check against CHGNet S.U.N. Tables 13 and 14 give the form agreement f_T that multiplies s_T in $\mathrm{GEM\text{-}match}$ (null and size-matched form checks are in Appendix E). Tables 15 and 16 give the per-probe discovery terms $\rho \cdot \mathrm{VS}_T/n$ whose mean is $\mathrm{GEM\text{-}disc}$. Under a collapse-sensitive kernel these are a diversity readout in the same spirit as counting unique stable novel crystals, but with soft mode identity instead of a hard uniqueness cut. The MP-20 and Alex-MP-20 blocks include a D row: dataset self-scores with $\rho = 1$ on form tables, and the reference-dataset discovery readout on the discovery tables (as in Table 2).

High novel-stable yield with low form or low Vendi. A large $\rho \cdot \gamma_T$ means many generated samples clear the object-level gate, but that mass can still fail matching through low f_T or fail discovery through low VS_T relative to the accepted size m_T. To compare cells fairly we use absolute losses against the yield scale, not form or Vendi alone. For matching, $L_T^{\mathrm{form}} = \rho \cdot \gamma_T\,(1 - f_T)$ is the novel-stable mass nullified by form disagreement under κ (Tables 12 and 13). For discovery, an all-distinct accepted set would give $\mathrm{VS}_T = m_T$ and hence $\rho \cdot \mathrm{VS}_T/n = \rho \cdot \gamma_T$ on the evaluated subset, so $L_T^{\mathrm{vendi}} = \rho \cdot \gamma_T - \rho \cdot \mathrm{VS}_T/n = \rho \cdot \gamma_T\,(1 - \mathrm{VS}_T/m_T)$ is the gap between that yield upper bound and the reported discovery term (Tables 12 and 15–16, physical VS uses the far95% map). We highlight six MP-20 cells that are severe under these losses while spanning probes, models, and the two score factors (physical kernel unless noted):

1. **SymmCD, $\mathbf{Stable_{50}}$, form:** L^{form}=0.527 ($\rho \cdot \gamma_T$=0.876, f_T=0.398): the largest form loss in the block.
2. **DiffCSP, $\mathbf{Stable_{300}}$, form:** L^{form}=0.404 (0.822, f_T=0.508): the largest form loss at 300 K.
3. **CrystalDiT, $\mathbf{EaH_{0.1}}$, form:** L^{form}=0.269 (0.271, f_T=0.006): the largest form loss on either hull cut. On EaH_0 under κ_{phys}, f_T=0.000 with $\mathrm{MMD}^2 \approx 1.28$ after the post-relaxation filter. The embedding column gives f_T=0.929 on the same EaH_0 cell, so this hull-form signal is physical-kernel specific.
4. **Wyformer-DiffCSP++, $\mathbf{Stable_{50}}$, Vendi:** L^{vendi}=0.795 physically (0.966, discovery term 0.171) and 0.941 under embedding: the largest Vendi losses under either kernel.
5. **MiAD, Phonon, Vendi:** L^{vendi}=0.390 (0.460, discovery term 0.070, $\mathrm{VS}_T/m_T \approx 0.15$): the largest Phonon Vendi loss. MiAD likewise leads $\mathbf{EaH_{0.1}}$ yield with a large Vendi gap.
6. **MatterGen, $\mathbf{Stable_{600}}$, Vendi:** L^{vendi}=0.538 (0.662, discovery term 0.124): the largest Vendi loss at 600 K.

These absolute losses are why means over $\mathcal{T}$ can disagree with a hull- or MD-only yield: form and Vendi remove substantial novel-stable mass across phonon, hull, and temperature probes, not only on a single MD gate.

Table 11: **Stability rates.** Per-probe stable fractions $\gamma_T = \mathrm{Pr}_G(S_T{=}1 \mid V{=}N_D{=}1)$ and corpus rates $\beta_T = \mathrm{Pr}_D(S_T{=}1)$ (kernel-independent; γ_T enters s_T in GEM-match). Higher is better. In the MP-20 block, bold marks the column maximum among models (ties included; the D row is excluded).

Model	Phonon	EaH_0	$\mathrm{EaH}_{0.1}$	Stable_{50}	Stable_{300}	Stable_{600}
MP-20 (D = MP-20 train)						
D (β_T)	0.684	0.137	0.478	0.999	0.971	0.884
Crystalite	0.585	0.103	0.353	0.987	0.935	0.780
MiAD	0.621	0.139	**0.446**	0.985	**0.942**	0.819
DiffCSP++	0.449	0.081	0.267	0.989	0.863	0.664
MatterGen	0.485	0.053	0.292	0.988	0.906	0.707
FlowMM	0.448	0.056	0.236	0.949	0.866	0.659
DiffCSP	0.468	0.073	0.290	0.992	0.889	0.654
Wyformer-DiffCSP++	0.389	0.054	0.189	**0.993**	0.796	0.544
SymmCD	0.423	0.078	0.256	0.926	0.759	0.550
CrystalDiT	**0.658**	**0.154**	0.440	0.955	0.919	**0.826**
OMatG	0.465	0.023	0.189	0.979	0.891	0.648
Alex-MP-20 (D = Alex-MP-20)						
D (β_T)	0.613	0.153	0.622	0.999	0.965	0.813
MiAD	0.584	0.080	0.521	0.970	0.938	0.767
MatterGen	0.561	0.075	0.509	0.996	0.930	0.725
CustomDataset (D = Alex-MP-20)						
Chemeleon2	0.731	0.168	0.775	1.000	0.981	0.876
CrysJEPA	0.685	0.020	0.895	0.988	0.947	0.639

Table 12: **SN: stability and novelty rates.** Per-probe novel-stable rates $\rho \cdot \gamma_T$ with $\rho = \rho_{\mathrm{pre}} \cdot \rho_{\mathrm{post}}$ (Table 9); β_T is shown for reference only. Higher is better. In the MP-20 block, bold marks the column maximum among models (ties included; the D row is excluded).

Model	Phonon	EaH_0	$\mathrm{EaH}_{0.1}$	Stable_{50}	Stable_{300}	Stable_{600}
MP-20 (D = MP-20 train)						
D (β_T)	0.684	0.137	0.478	0.999	0.971	0.884
Crystalite	**0.480**	0.085	0.290	0.811	0.768	0.641
MiAD	0.460	**0.103**	**0.330**	0.729	0.697	0.606
DiffCSP++	0.437	0.079	0.260	0.963	0.841	0.647
MatterGen	0.454	0.050	0.273	0.925	**0.848**	**0.662**
FlowMM	0.400	0.050	0.211	0.847	0.773	0.588
DiffCSP	0.433	0.068	0.268	0.917	0.822	0.604
Wyformer-DiffCSP++	0.379	0.053	0.184	**0.966**	0.775	0.529
SymmCD	0.400	0.074	0.242	0.876	0.718	0.521
CrystalDiT	0.405	0.095	0.271	0.587	0.565	0.508
OMatG	0.368	0.018	0.150	0.776	0.706	0.513
Alex-MP-20 (D = Alex-MP-20)						
D (β_T)	0.613	0.153	0.622	0.999	0.965	0.813
MiAD	0.398	0.054	0.355	0.660	0.639	0.522
MatterGen	0.409	0.055	0.371	0.726	0.678	0.529
CustomDataset (D = Alex-MP-20)						
Chemeleon2	0.707	0.162	0.750	0.968	0.949	0.848
CrysJEPA	0.614	0.018	0.801	0.884	0.848	0.572

Table 13: **Form agreement.** Per-probe form f_T under κ_{phys} (form term of GEM-match). Empty accepted sets score 0. Higher is better. In the MP-20 block, underline marks the column minimum among models (ties included; the D row is excluded).

Model	Phonon	EaH_0	$\mathrm{EaH}_{0.1}$	Stable_{50}	Stable_{300}	Stable_{600}
MP-20 (D = MP-20 train)						
D	1.000	1.000	1.000	1.000	1.000	1.000
Crystalite	0.862	0.893	0.854	0.865	0.871	0.887
MiAD	0.834	0.524	0.679	0.824	0.815	0.799
DiffCSP++	0.748	0.945	0.898	0.599	0.610	0.683
MatterGen	0.736	1.000	0.783	0.557	0.580	0.658
FlowMM	0.767	0.965	0.885	0.696	0.721	0.773
DiffCSP	0.614	0.843	0.691	0.494	0.508	0.640
Wyformer-DiffCSP++	0.747	0.975	0.887	0.745	0.783	0.856
SymmCD	0.490	0.840	0.785	0.398	0.506	0.665
CrystalDiT	0.415	0.000	0.006	0.622	0.583	0.544
OMatG	0.754	0.985	0.709	0.648	0.678	0.747
Alex-MP-20 (D = Alex-MP-20)						
D	1.000	1.000	1.000	1.000	1.000	1.000
MiAD	0.864	0.830	0.776	0.828	0.831	0.847
MatterGen	0.765	0.725	0.786	0.792	0.815	0.809
CustomDataset (D = Alex-MP-20)						
Chemeleon2	0.000	0.000	0.000	0.000	0.000	0.000
CrysJEPA	0.000	0.000	0.000	0.000	0.000	0.000

Table 14: **Form agreement.** Per-probe form f_T under κ_{embed} (form term of GEM-match). Empty accepted sets score 0. Higher is better. In the MP-20 block, underline marks the column minimum among models (ties included; the D row is excluded).

Model	Phonon	EaH_0	$\mathrm{EaH}_{0.1}$	Stable_{50}	Stable_{300}	Stable_{600}
MP-20 (D = MP-20 train)						
D	1.000	1.000	1.000	1.000	1.000	1.000
Crystalite	0.612	0.999	0.952	0.461	0.500	0.641
MiAD	0.918	0.955	0.986	0.810	0.838	0.885
DiffCSP++	0.921	0.990	0.974	0.881	0.909	0.942
MatterGen	0.330	0.955	0.719	0.000	0.056	0.196
FlowMM	0.000	0.941	0.726	0.000	0.000	0.000
DiffCSP	0.254	0.908	0.883	0.000	0.000	0.290
Wyformer-DiffCSP++	0.915	0.940	0.950	0.825	0.891	0.927
SymmCD	0.909	1.000	0.950	0.827	0.888	0.927
CrystalDiT	0.860	0.929	0.928	0.766	0.770	0.848
OMatG	0.000	0.424	0.458	0.000	0.000	0.000
Alex-MP-20 (D = Alex-MP-20)						
D	1.000	1.000	1.000	1.000	1.000	1.000
MiAD	0.940	0.943	0.977	0.879	0.886	0.939
MatterGen	0.628	0.707	0.735	0.567	0.628	0.645
CustomDataset (D = Alex-MP-20)						
Chemeleon2	0.000	0.293	0.000	0.000	0.000	0.000
CrysJEPA	0.000	0.526	0.000	0.000	0.000	0.000

Table 15: **GEM-disc per stability probe (physical).** Terms $\rho \cdot \mathrm{VS}_T/n$ under κ_{phys} with the far95% mode-size map (Appendix B.6); their unweighted mean is GEM-disc in Table 2. Higher is better. In the MP-20 block, bold marks the column maximum among models (ties included; the D row is excluded).

Model	Phonon	EaH_0	$\mathrm{EaH}_{0.1}$	Stable_{50}	Stable_{300}	Stable_{600}
MP-20 (D = MP-20 train)						
D	0.112	0.040	0.061	0.144	0.138	0.125
Crystalite	0.099	0.029	0.053	0.132	0.125	0.106
MiAD	0.070	0.018	0.037	0.098	0.092	0.078
DiffCSP++	**0.123**	0.024	0.052	**0.206**	**0.182**	**0.141**
MatterGen	0.114	0.021	**0.059**	0.165	0.151	0.124
FlowMM	0.087	0.022	0.048	0.138	0.126	0.103
DiffCSP	0.111	0.027	0.058	0.169	0.155	0.123
Wyformer-DiffCSP++	0.109	0.025	0.040	0.171	0.142	0.106
SymmCD	0.114	0.025	0.046	0.192	0.157	0.111
CrystalDiT	0.084	**0.032**	0.047	0.106	0.101	0.091
OMatG	0.086	0.014	0.045	0.130	0.122	0.098
Alex-MP-20 (D = Alex-MP-20)						
D	0.105	0.045	0.085	0.132	0.127	0.108
MiAD	0.079	0.025	0.057	0.102	0.098	0.082
MatterGen	0.098	0.029	0.075	0.136	0.126	0.105
CustomDataset (D = Alex-MP-20)						
Chemeleon2	0.078	0.035	0.075	0.093	0.090	0.079
CrysJEPA	0.043	0.012	0.051	0.057	0.054	0.036

Table 16: **GEM-disc per stability probe (embedding).** Terms $\rho \cdot \mathrm{VS}_T/n$ under κ_{embed} with the far95% mode-size map (Appendix B.6); their unweighted mean is GEM-disc in Table 2. Higher is better. In the MP-20 block, bold marks the column maximum among models (ties included; the D row is excluded).

Model	Phonon	EaH_0	$\mathrm{EaH}_{0.1}$	Stable_{50}	Stable_{300}	Stable_{600}
MP-20 (D = MP-20 train)						
D	0.013	0.007	0.010	0.014	0.014	0.014
Crystalite	0.033	**0.007**	0.011	0.050	0.046	0.035
MiAD	0.017	0.005	0.009	0.025	0.023	0.019
DiffCSP++	0.018	0.005	0.007	0.024	0.021	0.017
MatterGen	0.051	**0.007**	0.020	0.094	0.083	0.065
FlowMM	**0.092**	**0.007**	0.018	**0.182**	**0.157**	**0.114**
DiffCSP	0.051	0.005	0.011	0.104	0.084	0.053
Wyformer-DiffCSP++	0.018	0.006	0.007	0.025	0.021	0.018
SymmCD	0.016	0.006	0.007	0.024	0.020	0.014
CrystalDiT	0.017	**0.007**	0.010	0.024	0.023	0.019
OMatG	0.081	**0.007**	**0.022**	0.175	0.151	0.109
Alex-MP-20 (D = Alex-MP-20)						
D	0.012	0.007	0.010	0.013	0.013	0.012
MiAD	0.013	0.005	0.008	0.018	0.017	0.014
MatterGen	0.022	0.008	0.014	0.029	0.026	0.023
CustomDataset (D = Alex-MP-20)						
Chemeleon2	0.067	0.016	0.063	0.080	0.078	0.078
CrysJEPA	0.005	0.005	0.007	0.008	0.008	0.005

E FORM READOUT AND ACCEPTED-SET SIZE

The form term uses the same clipped unbiased MMD^2 readout as the collapse calibration suite (Appendix B): at small accepted-set sizes the U-statistic can take small negative values under a true null, and the one-sided clip maps that variance into upward bias in f_T. On thermodynamic hull probes the generator accepted counts in Table 13 span roughly 22–139 crystals (dataset-side pools: m=137 on $\mathrm{EaH_0}$, 478 on $\mathrm{EaH_{0.1}}$), whereas Phonon accepted sets are an order of magnitude larger, so any size-driven clip effect concentrates on the hull columns practitioners weight most.

Table 17 estimates the null form distribution by splitting the dataset-side accepted pool into two random size-m subsets (40 redraws, physical kernel only). Mean null f rises toward one as m grows, and a majority of null draws have $\mathrm{MMD}^2 < 0$ at $m \leq 60$ on $\mathrm{EaH_0}$. Accordingly, f_T=1.000 for MatterGen (m_T=53) on $\mathrm{EaH_0}$ in Table 13 is consistent with null clip mass at those sizes rather than strong evidence of perfect form match to the dataset. OMatG (m_T=22, f_T=0.985) is similarly clip-prone. By contrast, CrystalDiT's $\mathrm{EaH_0}$ form failure (f_T=0.000) survives common-m subsampling (Table 18). Under the embedding kernel, $\mathrm{EaH_0}$ shows no inverse yield/form relation across models, so the monotone pattern between γ_T and f_T on hull probes in Table 13 should not be read as a kernel-independent law.

Table 17: Null distribution of physical form f under dataset-vs-dataset splits at fixed accepted-set size m (40 redraws). On $\mathrm{EaH_0}$ and Phonon the accepted pools have 137 and 684 crystals. Small m inflates f through one-sided clipping of negative MMD^2 (Eq. (7)).

	$\mathrm{EaH_0}$ (m_{pool}=137)			Phonon (m_{pool}=684)		
m	mean f	$\mathrm{P}(\mathrm{MMD}^2<0)$	$\mathrm{P}(f\geq 0.99)$	mean f	$\mathrm{P}(\mathrm{MMD}^2<0)$	$\mathrm{P}(f\geq 0.99)$
10	0.784	0.65	0.68	—	—	—
15	0.775	0.57	0.57	—	—	—
20	0.878	0.65	0.65	—	—	—
25	0.934	0.65	0.70	0.849	0.55	0.57
30	0.917	0.57	0.60	—	—	—
40	0.941	0.65	0.68	—	—	—
50	0.955	0.72	0.78	0.963	0.62	0.65
60	0.987	0.65	0.70	—	—	—
100	—	—	—	0.977	0.62	0.68
150	—	—	—	0.985	0.62	0.68
200	—	—	—	0.990	0.68	0.72
300	—	—	—	0.992	0.62	0.82

Table 18: Physical-kernel $\mathrm{EaH_0}$ form after subsampling accepted sets to common size m (30 redraws). Full-set column reproduces Table 13. OMatG (m_T=22) is omitted at m=25 and m=50.

Model	m_T	full f_T	m=25	m=50
Crystalite	103	0.893	0.772±0.303	0.837±0.151
MiAD	139	0.524	0.628±0.300	0.587±0.201
DiffCSP++	81	0.945	0.913±0.137	0.933±0.072
MatterGen	53	1.000	0.967±0.068	0.997±0.010
FlowMM	55	0.965	0.883±0.193	0.963±0.066
DiffCSP	73	0.843	0.852±0.174	0.852±0.080
Wyformer-DiffCSP++	54	0.975	0.953±0.080	0.956±0.048
SymmCD	78	0.840	0.795±0.252	0.853±0.098
CrystalDiT	136	0.000	0.249±0.294	0.077±0.147
OMatG	22	0.985	—	—

F HULL-ONLY SCORES (S.U.N. / M.S.U.N. TESTS)

The main rankings average over the full panel $\mathcal{T}$ (phonons, NPT at 50/300/600 K, and both hull cuts). S.U.N. and M.S.U.N. use only energy-above-hull gates ($E_{\mathrm{hull}}<0$ and <0.1 eV/atom). This appendix asks what happens if GEMBench keeps the same combination rules, kernels, generation batches, and $\rho = \rho_{\mathrm{pre}} \cdot \rho_{\mathrm{post}}$, but restricts the mean to the two hull probes $\mathcal{T} = \{\mathrm{EaH}_0, \mathrm{EaH}_{0.1}\}$ (MACE-small cuts on the same generation batches as Section 4.4, still without within-batch uniqueness). Phonon and MD flags are ignored. Clash gating and the post-relaxation novelty filter are unchanged. Tables 19 and 20 report the resulting GEM-match and GEM-disc. Dense ranks and the Rank column use the same transferred MatterGen $t = 2\hat{\sigma}$ as the full-panel tables (Appendix C), not a hull-only re-estimate of $\hat{\sigma}$.

Table 19: **GEM-match with hull-only stability panel.** Matching ranking with $\mathcal{T} = \{\mathrm{EaH}_0, \mathrm{EaH}_{0.1}\}$ (MACE-small hull cuts), same dumps and kernels as Tables 1–2.

Model	Physical	Embedding	Rank
MP-20 (D = MP-20 train)			
D	1.000	1.000	—
Crystalite	**0.535**	0.598	**1.5**
MiAD	0.425	**0.688**	**1.5**
DiffCSP++	**0.516**	0.550	2.0
SymmCD	0.426	0.511	2.5
DiffCSP	0.402	0.472	3.0
MatterGen	0.405	0.379	3.5
FlowMM	0.371	0.332	3.5
Wyformer-DiffCSP++	0.358	0.363	4.0
CrystalDiT	0.002	0.513	4.0
OMatG	0.176	0.100	5.0
Alex-MP-20 (D = Alex-MP-20)			
D	1.000	1.000	—
MiAD	**0.369**	**0.446**	**1.0**
MatterGen	**0.364**	0.346	1.5
CustomDataset (D = Alex-MP-20)			
Chemeleon2	**0.000**	**0.128**	**1.0**
CrysJEPA	**0.000**	0.030	1.5

Table 20: **GEM-disc with hull-only stability panel.** Discovery ranking with $\mathcal{T} = \{\mathrm{EaH}_0, \mathrm{EaH}_{0.1}\}$.

Model	Physical	Embedding	Rank
MP-20 (D = MP-20 train)			
D	0.051	0.009	—
DiffCSP	**0.042**	**0.008**	**1.0**
Crystalite	**0.041**	**0.009**	**1.0**
MatterGen	**0.040**	**0.013**	**1.0**
CrystalDiT	**0.039**	**0.008**	**1.0**
FlowMM	0.035	**0.013**	1.5
DiffCSP++	0.038	0.006	2.0
SymmCD	0.035	0.006	2.0
OMatG	0.029	**0.014**	2.0
Wyformer-DiffCSP++	0.032	0.006	2.5
MiAD	0.027	**0.007**	2.5
Alex-MP-20 (D = Alex-MP-20)			
D	0.065	0.008	—
MatterGen	**0.052**	**0.011**	**1.0**
MiAD	0.041	**0.006**	1.5
CustomDataset (D = Alex-MP-20)			
Chemeleon2	**0.055**	**0.040**	**1.0**
CrysJEPA	0.032	0.006	2.0

On matching, Crystalite and MiAD share the best Rank (1.5): Crystalite leads the physical column and MiAD the embedding column, so a hull-only panel does not overturn the matching leader set. CrystalDiT's physical GEM-match collapses to $\approx$0.002 (form failure on EaH_0, Appendix D), while its embedding score remains mid-table. On discovery the absolute scores compress (physical values lie in a $\approx$0.015 band), and under the transferred t four models share Rank 1.0 (DiffCSP, Crystalite, MatterGen, CrystalDiT). MiAD remains outside that top class despite leading CHGNet S.U.N. and MACE $\rho \cdot \gamma_{\mathrm{EaH}_0}$ (Table 12): the hull-only GEM-disc still multiplies yield by Vendi$/n$, so high novel-stable mass that collapses in the Gram does not lead. Relative to Tables 1–2, dropping phonon and MD therefore preserves the matching/S.U.N. disagreement in spirit, but softens discovery separation and pulls CrystalDiT into the hull-only discovery tie: evidence that the full panel's temperature and dynamical probes, not only the hull cuts, drive the main-text discovery order. On the CustomDataset block (Chemeleon2, CrysJEPA), hull-only matching collapses to the embedding column because both models score 0 on EaH_0 under κ_{phys} (Table 13); hull-only discovery still separates them (Chemeleon2 leads on both kernels in Table 20).

G SCORE SCATTER PLOT

Figure 2 plots GEM-match against GEM-disc for the ten MP-20 checkpoints under each kernel. The two scores induce different orderings under either kernel, and the kernels disagree most on the discovery axis (Section 4.3).

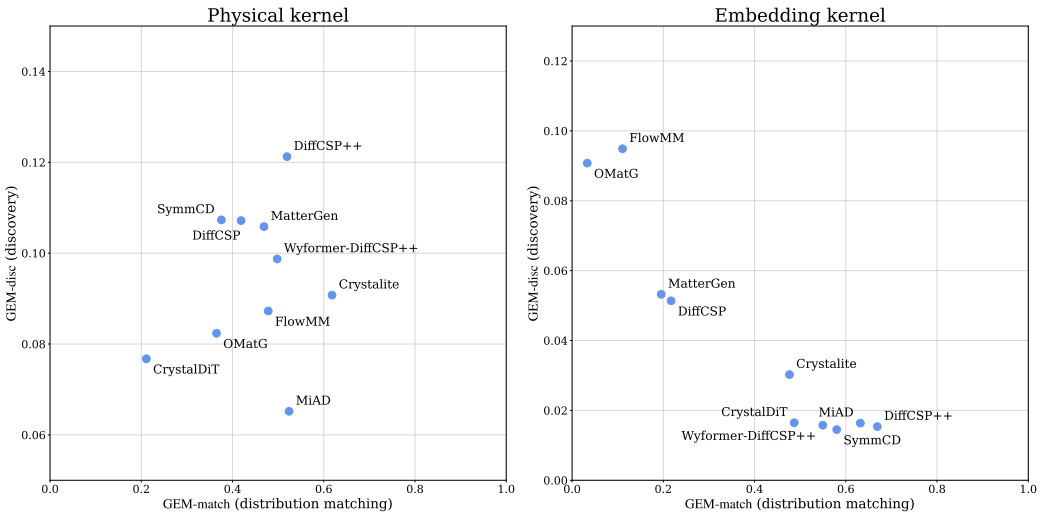

Figure 2: Matching versus discovery on the MP-20 block (physical kernel, left; embedding kernel, right). Point labels name the checkpoint; axis scales follow Tables 1 and 2.