THE NNP SCORE COMPLETENESS BOUNDARY: PHYSICS DECOMPOSITION FOR CRYSTAL STRUCTURE PREDICTION WITH NEURAL NETWORK POTENTIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Crystal structure prediction (CSP) faces a trade-off between data-driven generative models, which need substantial training data, and physics-based sampling, which is computationally expensive. Pre-trained neural network interatomic potentials (NNPs) offer a middle ground: their energy gradients define conservative scores, so they can in principle sample the Boltzmann distribution without training a generative model. We show that using them directly exposes a fundamental limitation. NNPs are not uniformly complete: they capture long-range chemistry and equilibrium forces accurately, but their training data contain no deep-overlap configurations, leaving short-range Pauli repulsion unconstrained and driving noise-dependent trajectories into an out-of-distribution (OOD) regime. We introduce a physics decomposition that pairs learned long-range chemistry with analytically specified short-range universality. The augmented score is exactly conservative and needs no generative training. Experiments show that the augmented score holds the induced rate to at most a few percent across the entire scan. The frontier where the score stops being complete—the NNP score completeness boundary—is a property of the training data that any machine-learned potential inherits, not an architectural defect. Supplying the missing short-range physics analytically is what makes a pre-trained potential a usable generative score. Our code is available at `https://anonymous.4open.science/r/MaterialGen_anoy-3A8B`.

1 INTRODUCTION

Generating crystal structures with a target composition is the central primitive of computational materials discovery. Data-driven diffusion and flow models (Xie et al., 2022; Jiao et al., 2023; Zeni et al., 2025; Miller et al., 2024) learn a score from tens of thousands of relaxed structures but are bounded by their training distribution and must be retrained for new chemistries. Physics-based sampling—molecular-dynamics annealing, basin hopping (Wales & Doye, 1997), metadynamics (Laio & Parrinello, 2002), nested sampling (Skilling, 2006; Pártay et al., 2010)—makes no distributional assumption and applies to any composition, but spends 10^5–10^7 force evaluations per trajectory because basin escape is thermally activated. Pre-trained neural network potentials (NNPs) (Chen & Ong, 2022; Batatia et al., 2025; Deng et al., 2023; Barroso-Luque et al., 2026) suggest a third route: an NNP is an energy function whose negative gradient is exactly the score of a Boltzmann distribution (Elijošius et al., 2025), $s_{\text{NNP}}(x) = -\beta\nabla_x U_{\text{NNP}}(x), \beta = 1/(k_B T)$, so crystal generation becomes sampling from $p_{\text{B}}(x) \propto e^{-\beta U_{\text{NNP}}(x)}$ without training a generative model at all.

This identity is exact, but its usefulness depends on where x lies relative to the potential's training distribution. Modern NNPs are trained on DFT-relaxed structures and their mild distortions, whose interatomic distances end in a hard lower edge rather than a fading tail; deep-overlap configurations are absent from every such dataset, their energies diverging as $r_{ij} \to 0$ (Fig. 1). The score is therefore *chemically complete*—faithful for bond stretching, angle bending, torsion, and electrostatics within the descriptor cutoff—but *sterically incomplete* where Pauli exclusion dominates. The first term is scoped to the chemistry the potential was trained on: a structure can be free of any overlap and still be out of distribution through an oxidation state, coordination environment, or composition the

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.

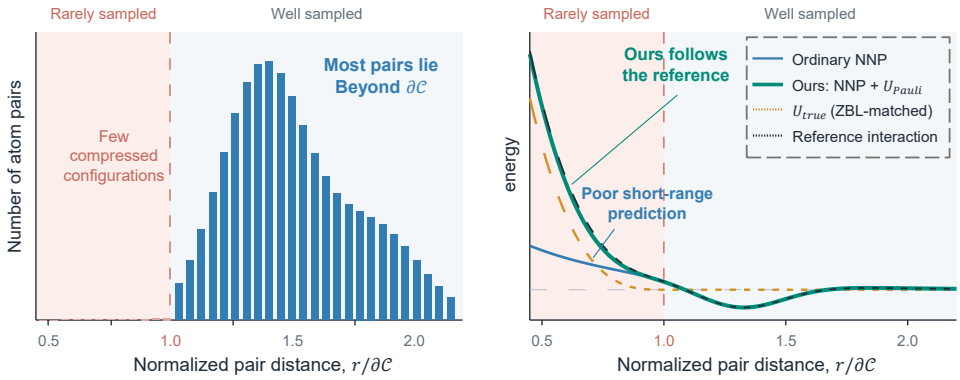


Figure 1: r is the interatomic pair distance and $\partial\mathcal{C}$ the *completeness boundary*; the abscissa of both panels is r in units of $\partial\mathcal{C}$. (a) Training-data coverage: pair counts fall steeply below $\partial\mathcal{C}$, so compressed configurations are rarely sampled, while most pairs lie beyond it. (b) Restoring short-range repulsion: below $\partial\mathcal{C}$ an ordinary NNP makes a poor short-range prediction, whereas the NNP augmented with the ZBL-matched U_{Pauli} follows the reference interaction. Schematic illustration; curves and counts are not measured data. Each of the three quantities drawn schematically—the corpus edge of (a), the potential's own short-range force spread, and the sampler collapse scale—is measured separately in Appendix A.9, which also states what remains unmeasured.

training data do not cover, which is a second boundary this study does not map (§6.1). The boundary we characterize is a short-range geometric one and is universal in a way that boundary is not, because every training set built from relaxed structures shares the same missing region. We formalize this as the **NNP score completeness boundary**. $\partial\mathcal{C} = \{x : \exists(i,j) \text{ s.t. } r_{ij}(x) = f_{\text{typ}}\,(R_i^{\text{cov}} + R_j^{\text{cov}})\}$, Inside it the score is reliable; outside it the score extrapolates, and using it in a sampler produces unphysical configurations. This is not a defect of a particular architecture but a coverage gap in the training data, which any machine-learned potential inherits—including potentials trained on non-equilibrium *dynamics*, since no dynamics at ordinary temperature visits configurations hundreds of eV above the ground state. The gap is not benign: the noise scale controls how often a trajectory crosses the boundary, and in a crystal the crossing couples to the cell degree of freedom, producing collapse and expansion modes with no molecular analogue.

The boundary also reveals how to repair it: a decomposition of "physics" for generative purposes. Long-range chemistry is system-specific and learned from DFT data; short-range sterics is universal for a given element pair and analytically expressible. We therefore augment the potential's energy with a Pauli repulsion term acting only below the completeness threshold, $s^{\text{aug}}(x) = -\beta\nabla_x\big[U_{\text{NNP}}(x) + U_{\text{Pauli}}(x)\big]$, where $U_{\text{Pauli}}(x)$ is an analytic pair potential whose parameters are matched to the Ziegler–Biersack–Littmark screening function (Ziegler et al., 1985) at the threshold radius. The augmentation is zero-training, exactly conservative, universal, and vanishes at equilibrium. Combined with a hard reflecting constraint on the non-overlap manifold—which acts on the cell as well as on individual pairs—it forms a defense architecture whose stationary distribution is the Boltzmann distribution of the augmented energy, at the effective temperature set by the noise schedule (Theorem 1).

Our contributions can be summarized as follows. First, we show that a pre-trained NNP cannot serve directly as a generative score—its training data contain no deep-overlap configurations, leaving short-range Pauli repulsion unconstrained—and define the completeness boundary $\partial\mathcal{C}$ that delimits where the score is complete. Second, we supply the universal short-range sterics analytically as a Pauli wall activated only below the boundary, and show that the augmented score is conservative by construction, that its error decomposes into a controllable smoothness term and a model-mismatch term (Theorem 3), and that it is stationary under the reflected dynamics (Theorem 4). Third, we show that a bare score is driven across the boundary as the noise scale grows and fails predominantly through atom overlap, that the collapse scale is composition-dependent rather than geometrically determined, and that the wall combined with a hard constraint holds the failure rate low at no measurable energy cost.

2 Related Work

Crystal structure prediction (CSP) is conventionally a two-stage pipeline: a search stage proposes candidate structures—by *ab initio* random structure searching (Pickard & Needs, 2011), evolutionary algorithms (Glass et al., 2006), or particle-swarm optimisation (Wang et al., 2012)—and a relaxation stage locally minimises each candidate on a potential-energy surface before the candidates are ranked by energy. Learned generative models now occupy the first stage: variational autoencoders (Xie et al., 2022), equivariant diffusion on joint position–lattice representations (Jiao et al., 2023; 2024), flow matching (Miller et al., 2024), and the large-scale MatterGen (Zeni et al., 2025), with language models as an orthogonal route (Gruver et al., 2024). The methodological difference is which stage each supplies: those models learn a distribution over relaxed structures and draw from it—DiffCSP's CSP variant is conditioned on composition and atom count, and generates coordinates and the cell from noise—so their competence is bounded by their database, and their scoring is under no constraint to be conservative, whereas parameterizing the field as a gradient restores conservative forces and the sampling guarantees that follow (Plainer et al., 2025). We take the relaxation stage instead and use a pre-trained potential as its score directly: no generative training, no database bound, conservative by construction, applicable to any composition. What must then be characterised is not how good the candidates are but how far from equilibrium one may be before the potential fails to relax it—exactly the regime the search stage produces, since a proposed candidate is by construction far from its relaxed geometry. Denoising diffusion (Sohl-Dickstein et al., 2015; Ho et al., 2020) and its continuous-time formulation (Song & Ermon, 2019; Song et al., 2021) supply the sampler; our score comes from physics rather than from data.

Universal potentials trained on large DFT corpora—M3GNet (Chen & Ong, 2022), CHGNet (Deng et al., 2023), MACE-MP-0 (Batatia et al., 2025), and eSEN on OMat24 (Barroso-Luque et al., 2026)—reproduce equilibrium energies and forces near DFT accuracy across the periodic table, and their use as Boltzmann generators for thermodynamics is established (Plainer et al., 2025). Their use as *generative* scores is newer and, as we show, blocked by a training-data gap rather than by accuracy.

Analytic repulsion is standard in classical force fields, and the ZBL screening function (Ziegler et al., 1985) is spliced into machine-learned potentials to correct them at short range. This is direct prior art, which we do not claim to have introduced: MACE builds in ZBL-style pair repulsion behind a smooth distance transform with an inner cutoff at the covalent radii (Batatia et al., 2025), and the same construction serves collision cascades and radiation damage, where the dynamics are short-range and high-energy rather than near-equilibrium (Byggmästar et al., 2019). That literature corrects the *potential*, validated against the dynamics it was fitted to. Our object is different: we use the potential as a *score*, measure where that score stops being usable, and show that the analytic wall's activation threshold is a measurable property of the training distribution rather than a numerical convenience. Two consequences follow. The gain we report is measured on top of whatever short-range physics the released model already carries, so the threshold is empirical (§5.2) rather than a design parameter. And the repulsion-off comparison is not in this study (Appendix A.9): our ablation establishes the effect of the protection *bundle* on the potential as released, not the marginal value of the analytic form over an architecture's own term. Elijošius et al. (2025) proposed a time-dependent "middle potential" for molecules; the mechanism it guards against—irreversible bond dissociation with free composition—does not apply to fixed-composition crystals, where the target class and constraint manifold are well defined, so the time-independent score suffices once the wall is restored (§4.1).

3 Preliminaries

A variance-exploding forward process, $dx = g(t)\, dW_t$ with $g(t) = \sqrt{d\sigma^2/dt}$, transports data to Gaussian noise; the reverse process (Anderson, 1982) requires the score $s_{\text{true}}(x, \sigma) = \nabla_x \log p_\sigma(x)$, which denoising score matching makes learnable (Vincent, 2011; Hyvärinen, 2005) and which the Tweedie identity relates to the posterior mean, $s_{\text{true}}(x, \sigma) = (\mathbb{E}_{x_0|x,\sigma}[x_0] - x)/\sigma^2$. Removing the stochastic term yields the probability-flow ODE (Song et al., 2021), whose deterministic trajectories preserve the marginals of the reverse SDE. At inverse temperature β the equilibrium distribution is $p_{\text{B}}(x) \propto e^{-\beta U(x)}$, whose score is exactly $-\beta \nabla U(x)$; a pre-trained NNP supplies U_{NNP} and, by

automatic differentiation, its energy gradient and—through the virial—its stress. We use annealed Langevin dynamics (ALD) and the probability-flow ODE (PF-ODE) as the two sampling paradigms.

A structure is $(L, \boldsymbol{f})$ with lattice matrix L (rows as lattice vectors) and fractional coordinates $\boldsymbol{f} \in [0,1)^{N\times 3}$; periodicity makes the domain a torus and the configuration-space dimension is $d = 3N + 9$ including the cell. Two operators carry a field between the Cartesian and the fractional chart and must not be conflated: a *displacement* transforms by the reciprocal of the lattice, $\Delta\boldsymbol{f} = \Delta\boldsymbol{r}L^{-1}$, whereas a *gradient* transforms by the chain rule, $\partial U/\partial\boldsymbol{f} = (\partial U/\partial\boldsymbol{r})L^{\top}$. Our score is a displacement generator rather than a fractional-chart gradient, so the update is expressed in Cartesian coordinates:

$$s^{\text{frac}} = L^{-\top} s^{\text{cart}}, \qquad \Delta\boldsymbol{x}^{\text{cart}} = \alpha\,(s^{\text{frac}} L), \tag{1}$$

that is, each step is converted back before it is taken, so the transformations compose to the identity and the update actually applied is $\Delta\boldsymbol{x}^{\text{cart}} = \alpha s^{\text{cart}}$. The distinction is load-bearing twice: the stationary law of Appendix I and the conservative structure of Theorem 2 are both statements about the Cartesian update, the one the Einstein relation applies to, where the score is a genuine gradient of $U(x)$ rather than a transformed field. The naive transpose $L^{\top}\boldsymbol{F}$ over-amplifies the drift by $|L|^2$ and is incorrect. Cell volume is $|\det L|$, and $r_{\min}$ is the minimum interatomic distance under the minimum-image convention.

4 METHOD

Because the composition is fixed and the cell bounded, the target is the Boltzmann distribution on a well-defined constrained manifold, so this score is a coherent time-independent target. Its error vanishes as $\sigma \to 0$—pointwise where the data density is smooth at scale σ, and in distribution for the near-δ modes of a relaxed-structure dataset, where what converges is the sampled law rather than the field (Theorem 3)—and its smoothness term is controlled while $\sigma \ll R_{\text{basin}}/\sqrt{d}$, beyond which the posterior spans several basins (Corollary 1). Four results frame everything below; all are proved in Appendix I. The score is the gradient of a scalar, so its Jacobian is everywhere symmetric and the field curl-free—a property no training choice can break (Theorem 2)—and the fixed-noise annealed chain therefore has the Boltzmann distribution of the augmented energy as its steady state, the noise schedule acting as the temperature schedule in the implemented convention (Theorem 1). The score error splits into a smoothness term that vanishes as $\sigma \to 0$ and a term that does not: it combines the potential's force error at the posterior mean with the mismatch between the data distribution and the Boltzmann law the score targets, and vanishes only where both do (Theorem 3, Remark 1); the reliability condition $\sigma \ll R_{\text{basin}}/\sqrt{d}$ bounds only the first. Reflection leaves the stationary law a truncated Boltzmann distribution of the augmented energy (Theorem 4), and the deterministic sampler propagates score errors without injecting a noise term of its own (Proposition 1).

4.1 THE COMPLETENESS BOUNDARY AND PHYSICS DECOMPOSITION

Where the score is complete, and where it is not. Above the boundary $\partial\mathcal{C}$ defined in §1 the score faithfully carries bond stretching, angle bending, torsion, and cutoff-limited electrostatics, with forces agreeing with DFT to within tens of meV/Å. Below it the potential extrapolates: its force can be wrong by orders of magnitude or point in the attractive direction while the true force is strongly repulsive, $F \propto e^{-r/a_B}$, with a_B the Bohr radius, an atomic length scale rather than a chemical one. A milder incompleteness sits at the opposite extreme: beyond the descriptor cutoff (∼5–6 Å) interactions are truncated, discarding dispersion and unscreened electrostatics—bounded and benign for stability at fixed cell, but it biases the sampled distribution and removes the restoring attraction that would counteract a stress-driven cell inflation (Remark 2).

Three quantities, not one. $\partial\mathcal{C}$ is where the protection threshold is *placed*: a fixed multiple $f_{\text{typ}} = 0.75$ of the covalent sum, calibrated against the measured edge of the training distribution rather than derived. It is not the radius at which the potential's own forces begin to deviate from the true ones, nor the noise level σ_c at which a given sampler collapses, which depends on the initial condition, step size and budget as well as on the potential (§5.2). This study measures the third directly, calibrates the first against the data, and measures the second on the potential's own side only; Appendix A.9 states which of the three each number here is, and which cannot be resolved.

Periodicity sharpens the boundary. In a crystal, crossing it has two collective modes with no molecular analogue: *lattice collapse*, the cell shrinking until every atom overlaps its periodic images, and *lattice explosion*, inflating past cohesion. They need different diagnostics—collapse by $r_{\min}$, explosion by cell volume, and $r_{\min}$ alone cannot detect both, since it grows without bound during an explosion—so the defenses below act on the cell as well as on pairs.

The physics decomposition. The boundary partitions the score into long-range, system-specific chemistry (bonding, electrostatics, dispersion, charge transfer—what potentials learn from DFT) and short-range, universal sterics (Pauli exclusion and sub-Å nuclear repulsion—functions of atomic numbers and distances alone). The second needs no learning, which also explains a known empirical fact: Elijošius et al. (2025) found that a learned energy-based score acts as a generic restorative potential early in generation and only becomes chemistry-specific late—what one expects if much of what early denoising requires is the universal wall.

4.2 Analytic Pauli Repulsion

We close the gap with

$$U_{\text{Pauli}}(x) = \sum_{i<j} A_{Z_i Z_j}\, e^{-B_{Z_i Z_j} r_{ij}}\, \Theta\big(f_{\text{typ}}(R_i^{\text{cov}} + R_j^{\text{cov}}) - r_{ij}\big)\,, \tag{2}$$

with Θ a smooth step of width w centered on the completeness threshold and element-pair parameters obtained by matching the energy *and force* of the ZBL screening function (Ziegler et al., 1985) just inside the activation radius, plus a per-pair entry-force calibration against the potential's own learned short-range force (Appendix A.4). Four properties follow. It is *zero-training* and *universal*: analytic, with tabulated parameters depending only on atomic numbers. It is *exactly conservative*: $\nabla \times s^{\text{aug}} \equiv 0$, both terms being gradients of scalars. It *vanishes at equilibrium*: the step is centred on the threshold, $\Theta(f_{\text{typ}}(R_i^{\text{cov}} + R_j^{\text{cov}})) = \frac{1}{2}$, and decays on the scale $w = 0.05$ Å, so a pair more than $5w$ outside the boundary has $\Theta < 5 \times 10^{-5}$ and a residual force below 2×10^{-2} eV/Å, and smaller still at a real bond distance (Ti–O at 1.97 Å is $7.4w$ outside, $\Theta < 10^{-6}$). It is therefore a safety net inside the boundary rather than a modification of equilibrium physics—a distinction that matters, since a wall at the full covalent sum would fire on every step. And it is *cheap*: an $\mathcal{O}(N\, n_{\text{neigh}})$ pair sum, under 1% of one potential evaluation. Because a steep exponential can still overshoot a compressed pair into its neighbours within one finite step, each atom's drift is throttled relative to its nearest-neighbour pair, so a deep pair climbs out at bounded speed; the two caps and their vanishing points are specified in Appendix A.4. They act on the discrete step only; s^{aug} is unchanged.

4.3 Reflecting Boundary Conditions

A repulsive gradient is a soft constraint: in $d = 3N + 9$ dimensions, isotropic noise has non-negligible probability of producing an overlap in a single step. We therefore impose a hard geometric constraint on the manifold

$$\mathcal{M} = \big\{x = (\boldsymbol{f}, L) : r_{ij}(x) \geq f_{\text{typ}}(R_i^{\text{cov}} + R_j^{\text{cov}})\ \forall i < j,\ \ |\det L| \leq f_{\text{out}} |\det L_0|\big\}, \tag{3}$$

where the inner wall sits on the completeness boundary and the outer wall caps the cell volume at f_{out} times the trajectory's initial volume. The latter does not distort the target: a cell inflated past cohesion is a dilute atomic gas, absent from both the data distribution and the potential's training set. The reflected SDE

$$dx = \sigma^2 s^{\text{aug}}(x)\, dt + \sqrt{2}\, \sigma\, dW_t + dk_t \tag{4}$$

uses the boundary local time k_t (Skorokhod, 1961), which pushes the trajectory along the inward normal only when it touches $\partial\mathcal{M}$. In discrete form this is a projection along the separation direction (Algorithm 1): penetration of a pair is split equally along it, the inward normal, and a cell exceeding the volume cap is isotropically rescaled. The projection is a finite iterated sweep rather than a closed-form guarantee: its pass cap, tolerance, cell-recovery and failure handling are specified in Appendix A.7, and because the convergence flag is not carried into the trajectory records, "hard" means enforced at every step whose sweep converges, not verified at every step. so it is the discrete analogue of the local-time term, not a specular reflection. Reflection costs no force evaluations, modifies only violating pairs, and is a safety net for finite-step discretization only: the wall's activation window makes the Pauli barrier at $\partial\mathcal{M}$ finite (1–5 eV per violating pair, Table 4), so the density

there is suppressed by $e^{-\beta U_{\text{Pauli}}(\partial\mathcal{M})/\sigma_k^2}$ and the stationary law is the truncated Boltzmann distribution of Theorem 4 at the effective temperature of Theorem 1. Without the wall the same constraint surface would distort the distribution; the two are a gradient/constraint pair, not substitutes. *Which process is being simulated.* Eq. (4) carries the noise level in the score, $\sigma^2 s^{\text{aug}}$, and its fixed-noise steady state is $e^{-\beta U_{\text{tot}}}$ at every σ. The implemented step carries it in the noise coefficient instead, $x_{t+1} = x_t + \alpha_k s^{\text{aug}}(x_t) + \sqrt{2\alpha_k}\,\sigma_k \boldsymbol{\epsilon}$. The two share a drift-to-diffusion ratio only when $\alpha_k = \sigma_k^2$, which the schedule does not satisfy, so the implemented chain is a separate process with a separate target, $e^{-\beta U_{\text{tot}}/\sigma_k^2}$ (Theorem 1); Eq. (4) is the idealized reference for the structural results below, and Remark 4 lists what else the discrete chain changes.

4.4 Defense Architecture and Sampling

The components above are organized into three preventive layers acting at successive stages of each step (Table 5, Appendix A.7): **L1**, the Pauli-augmented score, supplies the correct short-range gradient before the step; **L2**, density-adaptive noise, reduces the noise amplitude per atom by local density under the Einstein relation, $\sigma_i = \sigma_{\text{base}}/(1 + \gamma\rho_i)$; **L3**, the reflecting boundary, enforces the constraint during the step. All three cost no additional force evaluations. L1 is a *bundle*: it switches on the analytic wall and the wall-relative drift caps of §4.2 together, so a bare-to-L1 contrast on ALD measures the bundle, not the wall alone—no caps-only arm was run, the caps being ALD-only, so the same contrast on PF-ODE isolates the wall, where its independent effect is measured (Appendix C.2).

4.5 Samplers and Noise Schedules

ALD runs a short Langevin chain at each noise level, $d\tilde{x} = \sigma^2 s^{\text{aug}} dt + \sqrt{2}\sigma\, dW_t$ followed by reflection; annealing σ from $\sigma_{\max}$ to $\sigma_{\min}$ is therefore a temperature schedule *in the implemented convention*—where the chain's steady state is $e^{-\beta U_{\text{tot}}/\sigma_k^2}$—that moves the chain from exploration to exploitation and quenches it into the zero-temperature limit (Theorem 1). PF-ODE removes the stochastic term and integrates $dx/d\sigma = -\sigma s^{\text{aug}}$ deterministically, descending the augmented energy monotonically; injecting no fresh noise, it adds no noise term of its own to the score-error budget (Proposition 1). It is our default, though its advantage is empirical and relative rather than absolute (Appendix C). Lattice updates follow the same flow with the stress-derived score $s_L = -\beta V \boldsymbol{\sigma} L^{-\top}$, gated to the near-equilibrium tail, and $\sigma_{\max}$ is capped per composition at the measured collision threshold (§5.2). Ensemble-disagreement uncertainty quantification and its adaptive scheduler are documented in Appendix F; their threefold cost is why they are not adopted.

5 Experiments

5.1 Setup

Task, in CSP terms. The sampler here is the *relaxation stage* of the pipeline of §2, not its search stage: it receives a composition and a candidate structure and returns a relaxed one, with the reference coordinates standing in for the search stage's output. The scan below therefore measures that stage's capture radius—the quantity a pipeline needs and which no learned generator reports about the potential it is paired with. Match rate and coverage are recovery metrics, and §6.1 carries the composition-only scope.

We use eSEN-30M-MPTrj (Barroso-Luque et al., 2026) as the primary potential and MACE-MP-0 (Batatia et al., 2025) for the large noise scan, all inference-only. The scan's potential was chosen by a pre-registered speed calibration; the conclusions are re-verified with eSEN on five compositions (Appendix D), where the bare score is half again as fragile as MACE-MP-0's and the four-layer protected arm is again perfect on validity, though not energy-neutral at the highest noise levels. All results are on MP-20, whose 20 test compositions are stratified by chemistry and each initialized from that composition's lowest-formation-energy test structure; the generation-quality protocol adds Perov-5 and Carbon-24 (Appendix G). Stability is the energy above the convex hull, E_{hull} (pymatgen (Ong et al., 2013)), and because the sampler is training-free the released splits serve as reference sets rather than training data; Appendix A documents the split accounting—including why the splits are not composition-disjoint—with the remaining tables.

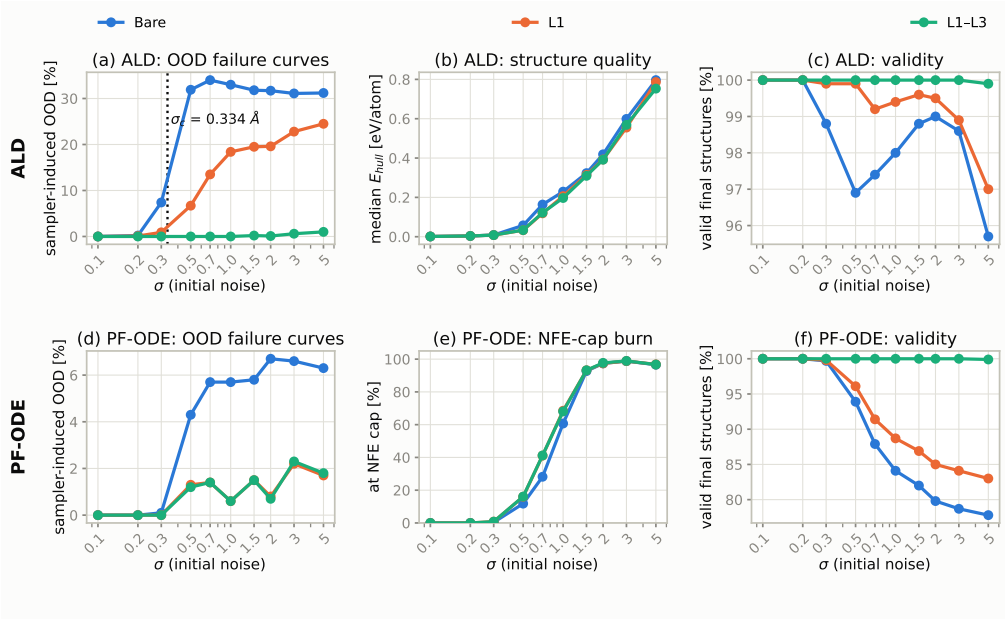


Figure 2: The completeness boundary and the effect of protection on MACE-MP-0, over 60,000 trajectories at ten noise levels. ALD (top row) and PF-ODE (bottom row) are on separate scales, their induced rates differing by roughly $5\times$. (a) Sampler-induced OOD rate versus $\sigma_{\max}$ for the bare score, Layer 1, and L1–L3; the dotted line marks the bare-ALD fitted collapse scale $\sigma_c = 0.334$ Å. Fits exclude the start-OOD floor. (b) Median E_{hull} and (c) terminal validity: both protected arms match the bare score at the two lowest noise levels and beat it from $\sigma \geq 0.3$ upward. (d) Induced rate for PF-ODE. (e) Fraction of PF-ODE trajectories exhausting the 1000-evaluation cap; above $\sigma = 2$ essentially all are truncated. (f) PF-ODE terminal validity, which no arm holds at 100%.

OOD accounting. We separate *start-OOD*—initial conditions already violating the constraint, a floor that is empty at low noise and grows past two-fifths of pooled starts at the top of the scan—from *sampler-induced* OOD, an event at step ≥ 1 from a clean start. Reported rates use the clean-start numerator over all trajectories; the event criteria and integrator convention are in Appendix A.

5.2 Measuring the Completeness Boundary

The boundary is real, steric, and self-recovering. Below a threshold noise the bare ALD sampler induces no violations at all; above it about a third of clean-start trajectories fail at least once; and the rate then *saturates* rather than continuing to grow—the signature of a boundary crossed rather than of a gradual loss of accuracy (Fig. 2(a); rates in Table 15). The fitted collapse scale, $\sigma_c = 0.334$ Å, places that crossing inside the recommended operating range rather than at its edge. Part of the plateau is a denominator effect rather than a change in risk, the rate counting clean starts in its numerator but every trajectory in its denominator; the convention, however, sets the level and not the comparison—conditioning on the clean pool leaves the arm ordering, the layer hierarchy, and the near-elimination of the curve by L1–L3 unchanged, though the rate itself keeps climbing rather than flattening (Table 14). What the boundary does *not* do is collapse the final state—terminal validity never falls below 95%, so a failure is a bounded transient that the trajectory recovers from, and that is why the induced-rate metric rather than a terminal one locates it.

Failures are early-step events, not an initial-condition artifact. Collisions concentrate in the first few steps of a trajectory whatever its start, and clean starts supply a quarter to a third of them, so the sequence rather than the initial condition is doing the work. Re-sampling starts to enforce a minimum pair distance was pre-registered to cut the $\sigma = 5$ Å collision rate by at least a factor of two; it rejects a quarter to a third of starts and delivers 13%, so hardening the initial condition is not the mechanism.

The collapse scale is composition-dependent, and its magnitude is estimator-dependent. Reading σ_c as the noise at which the per-step collision rate reaches 0.005, the crossing moves by more than an order of magnitude across the 20 test compositions, from below the scan's first resolved point to past its edge (Table 16); its exact value is not resolved, the statistic being size-confounded and floor-sensitive, so the table orders compositions rather than fixing their scales (Appendix H). The conclusion that matters survives every estimator tried: the spread is far wider than the nearly flat range basin geometry alone predicts (Corollary 1), so what sets where the boundary sits is the potential's learned short-range extrapolation, not the basin being sampled. In practice an a-priori force probe costing minutes per composition and requiring no sampling predicts the crossing to within a factor of two—enough to choose $\sigma_{\max}$ without a scan.

Protection eliminates the failure curve rather than shifting it, and Layer 1 is free on the stochastic arm. Layer 1 alone lowers the induced rate everywhere the bare sampler is not already safe, most sharply just above the threshold, and the three-layer stack holds it at or below one percent across the entire scan (Fig. 2(a)). On the ODE arm the wall is nearly inert on the failure rate but not on validity, which it raises to essentially full by ejecting the pairs that start out already violating the constraint (Fig. 2(f)). Protection is not bought with quality: Layer 1's median E_{hull} matches the unprotected arm where the bare sampler is safe and sits at or below it where it is not (Fig. 2(b)). That energy neutrality is a property of the noising dynamics rather than of the wall, however: it holds only on the stochastic arm, whereas the same layer sits *above* the bare score at every noise level where the ODE arm is unprotected (Table 15). The balance therefore favours $\sigma_{\max} \leq 1$ Å with the three-layer stack.

Layer hierarchy, sampler choice, and parameter sensitivity. A grid over the defense-layer subsets (Table 8) ranks the reflecting constraint first and nearly sufficient on its own, and the analytic wall a strong second where no constraint is present but largely redundant once one is. Both rankings survive statistics that threshold nothing—which the constraint, acting on the same distance the OOD metric tests, cannot reach by construction (Appendix C.2, which also carries the marginals and robustness checks for all three ablations). Unprotected PF-ODE fails five to seven times less often than unprotected ALD, but it is not OOD-free, and its advantage must be read beside its budget: above $\sigma_{\max} = 0.5$ Å the share of trajectories exhausting the evaluation cap climbs from a minority to essentially all of them, and the truncated group—carrying almost all of its remaining events at barely-annealed structures—never reached the low-noise regime at all. Proposition 1 establishes only that determinism contributes no noise term of its own; the operating advantage is measured rather than derived. The wall's zero-calibration claim is qualified at the screening-adjacent $B = 3\,\text{Å}^{-1}$, where metal–oxygen pairs are non-monotone.

5.3 Evaluation Protocols and Executed Result Slots

Three questions the measurements above do not settle are how generation quality compares with trained baselines; how much of what a learned score acquires from data is the wall we supply analytically; and whether the conclusions transfer to DFT. The protocols that measure them—generation-quality benchmarking, score-field comparison and curl violation, efficiency against molecular dynamics and basin hopping, cross-potential validation, and four-tier DFT validation—are specified in Appendix G, along with those that remain reserved. The generation-quality benchmark is the one since executed on its sampler side, reported here in Table 1 and in full in Appendix G.2; its trained-baseline half was not run.

Protection is complete on terminal validity, and it is not paid for in quality. Every protected arm terminates at full validity—both layer sets, both samplers—against two-thirds to nine-tenths for the bare arms, so the claim rests on no particular choice of aggregate. The quality columns move the other way or not at all: protection lowers both the median E_{hull} and the angular KL, leaves coverage at the operating noise level at least as high, and roughly doubles it where the bare arm is destructive (Table 1). One asymmetry reproduces on a generation benchmark the layer hierarchy the noise scan established: at the operating noise level the constraint buys nothing that Layer 1 does not, and the wall earns its place at higher noise instead, where it is the difference between a largely invalid arm and a fully valid one.

What the secondary datasets add. Both reproduce the headline results: protection again takes validity to full on every protected arm and the layer ordering is unchanged, and Perov-5's shortfalls track

Table 1: Generation-quality benchmark on MP-20, eSEN, 1,000 trajectories per cell (20 test compositions × 5 seeds × 10 candidates). Rows 1–6 are ALD and rows 7–9 PF-ODE; the two samplers are never pooled. C5 is the fixed-cell configuration and is the primary lattice ablation against C8, which changes both the cell treatment and the layer set and so is not a clean contrast.

Cell	$\sigma_{\max}$	Layers	Valid	Stable	S.U.N.	Match	Cover.	AMSD	R-KL	E_{hull}	Cap
C1	0.5	—	70.6	36.6	14.0	58.4	74.5	1.0	5.80	125	—
C3	1.5	—	66.2	3.0	2.1	11.6	10.0	1.3	8.97	598	—
C5†	0.5	—	70.3	35.5	13.3	58.5	67.0	1.0	5.82	127	—
C8	0.5	L1–L3	100.0	72.1	20.5	72.5	72.8	0.8	4.36	67	—
C9	1.5	L1–L3	100.0	4.7	3.4	13.2	22.0	1.3	8.56	613	—
C11	0.5	L1	100.0	72.7	20.3	74.6	74.5	0.8	4.35	55	—
C5	0.5	—	90.1	74.0	21.3	82.6	71.2	0.6	4.16	46	10.0
C8	0.5	L1–L3	100.0	82.3	17.2	80.3	76.2	0.6	3.96	29	12.4
C11	0.5	L1	100.0	82.2	17.2	80.2	76.2	0.6	3.95	29	12.0

All rates in %; AMSD in Å; R-KL in nats against the reference angular distribution (lower is closer); E_{hull} in meV/atom, a median over cells. Stable is the two-sided band $-50 \leq E_{\text{hull}} < 100$ meV/atom (Appendix A), so a suspiciously low E_{hull} counts as a failure. Cap is the share of PF-ODE trajectories exhausting the 1,000-evaluation budget, undefined for ALD's fixed step budget. †C5 holds the cell fixed; every other row updates it.

its chemistry rather than any layer. Carbon-24 is sharper—stable and novel are disjoint there, so no candidate is both, an emptiness in the intersection and not in the marginals: its samplers reached the stability window only where they reproduced an allotrope the dataset already held.

Two qualifications carry over from the noise scan, both properties of the setup rather than of the sampler. Protection buys validity, not energy quality—at $\sigma_{\max} = 1.5$ Å both arms are destructive and the wall cannot recover the energy distribution—and coverage measures recovery of a known neighbourhood rather than search from a composition alone (§6.1).

6 Discussion and Conclusion

Our central finding is that a pre-trained potential's score has a completeness boundary—chemically complete but sterically incomplete, blind to Pauli repulsion because overlapping configurations are absent from every DFT-derived training set. It follows from the training distribution rather than from any architecture, and it is measurable. In the CSP pipeline of §2 that boundary sits exactly where a training-free potential looks most attractive—the relaxation stage—and it turns the distance at which this stage stops working from an assumption into a measurement.

The boundary points to a decomposition: because the score is the gradient of an energy rather than a generative objective, the repair costs no generative training—system-specific chemistry is what the potential learns, universal sterics can be supplied analytically. The wall lowers the failure rate where the bare sampler is at risk; the constraint removes the curve rather than shifting it. Energy neutrality does not travel—the wall is neutral on the stochastic arm and costly on the deterministic one—so what is neutral is the wall together with the noising dynamics that conceal it, not the wall alone.

6.1 Limitations

The analytic correction is universal but not parameter-free. Element-pair parameters transfer across chemical environments but must be fixed once per pair, and zero-calibration holds only in the steep-wall regime: for metal–oxygen pairs a wall as steep as the screening function is non-monotone in its prefactor (§C.4). *The cell is gated, not freely sampled*: stress tensors of highly perturbed structures are unreliable, so the cell is held through the high-noise phase and the method refines its initial cell rather than searching polymorphs. *Scope*: unit cells of at most tens of atoms, thermodynamic but not dynamical (phonon) stability, no property-conditioned generation, and only the steric boundary.

STATEMENT ON THE USE OF AI ASSISTANTS

Generative AI tools were used in three capacities, matching the AI-use declarations of our submission form: to aid and polish the writing of the manuscript; to draft portions of the manuscript text; and to assist with research ideation and execution, specifically exploratory discussion of experimental design, the implementation and debugging of the sampling and analysis code, and the root-cause analysis of anomalies in the generation pipeline that led to the corrected initial-condition generator described in §6.1. AI tools were *not* used to generate synthetic datasets, to formulate mathematical claims or to assist in the writing of proofs, to clean or reformat datasets, for translation, or for qualitative and thematic data analysis. Every AI-assisted artifact was reviewed by the authors: code was executed and its outputs checked against the underlying data, and every mathematical statement in Appendix I was derived and verified by the authors. No result, citation, or analysis appears in this paper without verification by the authors against the underlying data or the primary literature, and the authors take full responsibility for all content.

ETHICS STATEMENT

This work develops methods for computational materials discovery and uses only publicly available datasets and publicly released pre-trained models. No human subjects, personal data, or animal subjects are involved. The principal societal consideration is the general one for accelerated materials screening: generated candidates are surrogate-level proposals that require higher-level validation before any experimental or application claim, and we make the reliable domain of the surrogate an explicit and measurable object of study rather than an implicit assumption.

REPRODUCIBILITY STATEMENT

All methods are specified analytically: the score is the gradient of a pre-trained potential plus an analytic pair potential whose parameters derive from tabulated covalent radii and the universal screening potential, with no fitted components. The sampling algorithm is given as pseudocode (Algorithm 1) with all protocol constants reported in Appendices A and B; datasets, splits, and pre-trained checkpoints are all public. Theory assumptions are stated with each result in Appendix I. The trajectories underlying §5.2 have been regenerated under the corrected initial-condition generator described in §6.1, and every table and figure in this paper reports corrected cells; the data and the analysis code that produces every table and figure (including `regen_sigma_pooled.py` for Table 15 and `comp_sigma_c.py` for Table 16) are available in the anonymized repository at `https://anonymous.4open.science/r/MaterialGen_anoy-3A8B`.

REFERENCES

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. doi: 10.1016/0304-4149(82)90051-5.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, Matt Uyttendaele, C. Lawrence Zitnick, and Zachary W. Ulissi. The open materials 2024 (omat24) inorganic materials dataset and models. *Nature Computational Science*, 6(6):642–652, 2026. doi: 10.1038/s43588-026-00996-w. Preprint: arXiv:2410.12771 (2024).

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry

Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18):184110, 2025. doi: 10.1063/5.0297006. Preprint: arXiv:2401.00096 (2023/2024).

Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical Review Letters*, 97(17):170201, 2006. doi: 10.1103/PhysRevLett.97.170201. The FIRE optimizer.

J. Byggmästar, A. Hamedani, K. Nordlund, and F. Djurabekova. Machine-learning interatomic potential for radiation damage and defects in tungsten. *Physical Review B*, 100(14):144105, 2019. doi: 10.1103/PhysRevB.100.144105.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022. doi: 10.1038/s43588-022-00349-3.

Beatriz Cordero, Verónica Gómez, Ana E. Platero-Prats, Marc Revés, Jorge Echeverría, Eduard Cremades, Flavia Barragán, and Santiago Alvarez. Covalent radii revisited. *Dalton Transactions*, (21):2832–2838, 2008. doi: 10.1039/b801115j.

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. doi: 10.1038/s42256-023-00716-3.

J. R. Dormand and P. J. Prince. A family of embedded runge–kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980. doi: 10.1016/0771-050X(80)90013-3.

Rokas Elijošius, Fabian Zills, Ilyes Batatia, Sam Walton Norwood, Dávid Péter Kovács, Christian Holm, and Gábor Csányi. Zero shot molecular generation via similarity kernels. *Nature Communications*, 16(1):5991, 2025. doi: 10.1038/s41467-025-60963-3.

P. P. Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der Physik*, 369(3):253–287, 1921. doi: 10.1002/andp.19213690304.

Colin W. Glass, Artem R. Oganov, and Nikolaus Hansen. USPEX—evolutionary crystal structure prediction. *Computer Physics Communications*, 175(11-12):713–720, 2006. doi: 10.1016/j.cpc.2006.07.020.

Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h–pu. *The Journal of Chemical Physics*, 132(15):154104, 2010. doi: 10.1063/1.3382344.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. Preprint: arXiv:2402.04379.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pp. 6840–6851, 2020.

M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics — Simulation and Computation*, 19(2):433–450, 1990. doi: 10.1080/03610919008812866.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

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 36 (NeurIPS 2023)*, pp. 17464–17497, 2023. doi: 10.52202/075280-0767. Preprint: arXiv:2309.04475.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. Preprint: arXiv:2402.03992. Introduces DiffCSP++, the space-group-constrained successor to DiffCSP.

G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6(1):15–50, 1996. doi: 10.1016/0927-0256(96)00008-0.

G. Kresse and D. Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B*, 59(3):1758–1775, 1999. doi: 10.1103/PhysRevB.59.1758.

Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002. doi: 10.1073/pnas.202427399.

Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rasmus Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, Bjørk Hammer, Cory Hargus, Eric D. Hermes, Paul C. Jennings, Peter Bjerre Jensen, James Kermode, John R. Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Pedersen, Carsten Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S. Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W. Jacobsen. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017. doi: 10.1088/1361-648X/aa680e.

Thomas E. Markland and Michele Ceriotti. Nuclear quantum effects enter the mainstream. *Nature Reviews Chemistry*, 2(3):0109, 2018. doi: 10.1038/s41570-018-0109.

Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M. Wood. Flowmm: Generating materials with riemannian flow matching. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35664–35686, 2024. Preprint: arXiv:2406.04713.

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. doi: 10.1016/j.commatsci.2012.10.028.

Lívia B. Pártay, Albert P. Bartók, and Gábor Csányi. Efficient sampling of atomic configurational spaces. *The Journal of Physical Chemistry B*, 114(32):10502–10512, 2010. doi: 10.1021/jp1012973.

Chris J. Pickard and R. J. Needs. *Ab initio* random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201, 2011. doi: 10.1088/0953-8984/23/5/053201.

Michael Plainer, Hao Wu, Leon Klein, Stephan Günnemann, and Frank Noé. Consistent sampling and simulation: Molecular dynamics with energy-based diffusion models. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025. Preprint: arXiv:2506.17139.

John Skilling. Nested sampling for general bayesian computation. *Bayesian Analysis*, 1(4):833–859, 2006. doi: 10.1214/06-BA127.

A. V. Skorokhod. Stochastic equations for diffusion processes in a bounded region. *Theory of Probability & Its Applications*, 6(3):264–274, 1961. doi: 10.1137/1106035.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 11895–11907, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *The Ninth International Conference on Learning Representations (ICLR 2021)*, 2021. Preprint: arXiv:2011.13456.

Alexandre Tkatchenko, Robert A. DiStasio, Roberto Car, and Matthias Scheffler. Accurate and efficient method for many-body van der waals interactions. *Physical Review Letters*, 108(23): 236402, 2012. doi: 10.1103/PhysRevLett.108.236402.

Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Computational Materials*, 6(1):20, 2020. doi: 10.1038/s41524-020-0283-z.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17 (3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

David J. Wales and Jonathan P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1997. doi: 10.1021/jp970984n.

Yanchao Wang, Jian Lv, Li Zhu, and Yanming Ma. CALYPSO: A method for crystal structure prediction. *Computer Physics Communications*, 183(10):2063–2070, 2012. doi: 10.1016/j.cpc.2012.05.008.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *The Tenth International Conference on Learning Representations (ICLR 2022)*, 2022. Preprint: arXiv:2110.06197.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025. doi: 10.1038/s41586-025-08628-5.

J. F. Ziegler, J. P. Biersack, and U. Littmark. *The Stopping and Range of Ions in Solids*. Pergamon Press, New York, 1985. ISBN 978-0-08-021603-4. Volume 1 of the series The Stopping and Range of Ions in Matter.

A Experimental Setup and Protocol Details

We fix as many degrees of freedom as possible before experiments and report all constants, so that the protocol can be reproduced without contact with the authors. The *design* principles are: use community-standard splits and evaluation protocols wherever established; freeze the protocol before the runs; and report every hyperparameter, seed, and software version.

A.1 Hardware, Software, and Seeds

Generation experiments run on a single NVIDIA A100 (80 GB); DFT calculations are CPU-only. The Python environment is 3.11 with `torch` 2.4.1 (CUDA 12.4), `fairchem-core` for the eSEN calculator, `mace-torch` for MACE-MP-0, `pymatgen` (Ong et al., 2013) $\geq$ 2024.0.0 for structure handling and `StructureMatcher`, and `ase` (Larsen et al., 2017) $\geq$ 3.22.0 for the cell and I/O layer. The sampling experiments use the seeds $\{42, 123, 999, 2024, 7777\}$ and the classical baselines use $\{0, \ldots, 9\}$; a seed controls both the initial noise sample and the Langevin noise sequence.

A.2 Potential Models and Force Computation

Table 2: Potential models used throughout. All are inference-only, with no fine-tuning; forces are obtained by automatic differentiation of the energy with respect to positions. Reference errors are those reported by the model authors on equilibrium DFT datasets; accuracy on generated out-of-equilibrium structures is expected to differ, which is precisely what the DFT validation protocol of Appendix G is designed to measure.

Property	**eSEN**	**MACE-MP-0**
Checkpoint	`esen_30m_mptrj.pt`	`MACE-MP-0` (medium)
Package	`fairchem` OCPCalculator	`mace-torch`
Cutoff radius	6.0 Å	6.0 Å
Stress support	Yes (autodiff)	Yes
Precision	float32	float32
Force MAE (ref.)	$\sim$15 meV/Å	$\sim$25 meV/Å
Energy MAE (ref.)	$\sim$8 meV/atom	$\sim$12 meV/atom

Force, stress, and score assembly. For a structure with atomic numbers $\{Z_i\}$, Cartesian positions $\{\boldsymbol{r}_i\}$, and lattice matrix L (rows as lattice vectors), we build a periodic `Atoms` object, evaluate $U(x)$ and $\boldsymbol{F}_i = -\partial U/\partial \boldsymbol{r}_i$, and when lattice updates are enabled compute the stress $\boldsymbol{\sigma} = V^{-1}\,\partial U/\partial \boldsymbol{\epsilon}|_{\boldsymbol{\epsilon}=0}$ and convert it to a lattice score $s_L = -\beta V \boldsymbol{\sigma} L^{-T}$. Cartesian forces are pulled back to fractional coordinates as $\boldsymbol{f}_i^{\text{frac}} = L^{-T}\boldsymbol{F}_i$ (Eq. (1)), and the total score is $s(x) = \beta\,[\boldsymbol{f}_1^{\text{frac}}, \ldots, \boldsymbol{f}_N^{\text{frac}}, \text{vec}(\boldsymbol{f}^{\text{latt}})]^\top$.

Diagnostic separation. For the augmented score we record the potential and wall force components separately. Every OOD criterion and every Step-Monitor threshold in this paper refers to the potential's force alone; the analytic wall legitimately reaches 10^2–10^3 eV/Å inside the completeness boundary and must not trip any diagnostic.

A.3 Datasets and Splits

Table 3: Dataset statistics and splits. All follow the standard CDVAE/DiffCSP protocol, and the counts are those of the released split files. Because the sampler is training-free the splits are reference sets rather than training data; the portion each one uses, and why the split is not composition-disjoint, are given in the text.

Dataset	**Structures**	**Comps.**	**Train**	**Val**	**Test**
MP-20	45,229	37,217	27,136 (60.0%)	9,047 (20.0%)	9,046 (20.0%)
Carbon-24	10,153	1	6,091 (60.0%)	2,032 (20.0%)	2,030 (20.0%)
Perov-5	18,928	18,928	11,356 (60.0%)	3,787 (20.0%)	3,785 (20.0%)

MP-20 (Jiao et al., 2023) is a Materials Project subset restricted to at most 20 atoms per cell, spanning 37,217 distinct compositions over oxides, sulfides, fluorides, nitrides, and intermetallics. Carbon-24 (Xie et al., 2022) is a single-composition dataset (C) of carbon polymorphs of 6–24 atoms per cell at densities of 2.2–3.7 g/cm^3, generated by random structure search and DFT relaxation at 10 GPa. Perov-5 (Xie et al., 2022) comprises 18,928 five-atom ABX_3 perovskites, essentially one per composition and predominantly of chemistries (Ti–Os–O–F, Co–N–O–Tl) largely absent from the Materials Project; the five canonical compounds $BaTiO_3$, $CaTiO_3$, $SrTiO_3$, $BaZrO_3$ and $CaZrO_3$ occur once each.

What the splits are for. The sampler is training-free, so no split is fitted on and we do not present these as a train/test protocol in the machine-learning sense. The two portions play different roles in different parts of the study. The noise scan of §5.2 uses the test portion alone: each of its 20 cells is one composition's lowest-formation-energy test structure, and that structure's DFT formation energy anchors the cell's calibration. The generation-quality protocol of Appendix G uses both: the test portion supplies the target structures for match rate and coverage, and the training portion supplies the reference index for the novelty test. Both roles require only that the two portions be disjoint as *structures*, which they are; they are not disjoint by composition, and neither role needs them to be. 1,686 reduced compositions occur in both MP-20 portions (22.8% of MP-20 test rows) and 2,214 in both Perov-5 portions (58.5% of Perov-5 test rows); Carbon-24 has a single composition, so composition-disjointness is impossible there. On Perov-5 the overlap is deliberate and load-bearing: Perov-5 holds essentially one structure per composition, so a composition taken only from the test portion would have no training-set neighbour and every generation of it would be novel for free. Restricting the axis to dual-split compositions is what makes the novelty test non-trivial and match rate meaningful. We use the released splits unchanged rather than re-splitting by composition, so that validity, coverage and novelty remain comparable with the learned-baseline literature that uses the same files.

Cell axes. A *cell* is one starting structure, noised at $\sigma_{\max}$. MP-20 and Perov-5 cells are one per composition, so pooling per cell and per composition coincide. Carbon-24 has a single composition, so its 20 cells are density strata of the test split's volume-per-atom distribution and the uniqueness pool is per cell rather than per composition—pooling all 1000 structures of composition C together would measure the strata against each other instead of measuring the sampler. Because the lattice is held fixed there, lattice-update configurations coincide on Carbon-24 (C1 with C5, C8 with C10) and are reported as the identity they are.

For potential evaluation we convert the CIF records to `pymatgen Structure` objects as released, without primitive standardization or supercell expansion, so evaluations run on the published cells. Those cells are short: the median shortest perpendicular cell height is 4.2 Å in the MP-20 test portion (minimum 2.0 Å, 90% below the 6 Å cutoff), 4.1 Å for Perov-5 and 2.4 Å for Carbon-24. An atom's periodic image therefore falls inside its own descriptor range, and the potential evaluates an environment that does not correspond to an isolated bulk cell. This applies to the reference energies used for calibration and to the samplers and baselines alike, and we do not correct for it.

Convex-hull energies are precomputed with `pymatgen.analysis.phase_diagram`, using for each dataset the reference phases available for its chemistry. For MP-20 the entries are the dataset's own Materials Project formation energies over all splits (37,217 phases across 23,371 chemical systems, including the test portion, as is standard for hull-based stability evaluation); this is Materials Project data, but restricted to the MP-20 subset, so competing phases that exist in the full Materials Project but not here are missing. Removing reference phases can only raise the reference hull—a minimum over a subset is at least the minimum over the full set—so the reported E_{hull} is a *lower* bound and every stability rate and S.U.N. value in this paper is optimistic relative to the complete diagram, by an amount this study does not quantify. For Perov-5 the entries are the dataset's own `heat_all` formation energies; 83% of its chemical systems hold a single composition, so E_{hull} reduces to a formation energy rather than a distance to a competing phase, and the same collapse applies to the majority of the MP-20 systems (16,402 of 23,371, i.e. 70.2%, hold one composition). For Carbon-24 the only reference is elemental carbon, and we take the dataset's own lowest energy per atom as the zero; since these structures were relaxed at 10 GPa, an ambient-pressure carbon reference would be a pressure mismatch. Candidate energies are mapped onto each hull's DFT formation-energy scale by a per-composition constant shift, $e^{\text{NNP}}_{\text{form}}(x) = E_{\text{NNP}}(x)/N - [\,E_{\text{NNP}}(x_{\text{ref}})/N_{\text{ref}} - e^{\text{DFT}}_{\text{form}}(x_{\text{ref}})\,]$, with x_{ref} the cell's reference structure: the shift absorbs the potential's constant per-atom energy offset, which assumes that offset is structure-independent rather than that the potential is accurate. Stability is tested as a two-sided band, $-50 \le E_{\text{hull}} < 100$ meV/atom, not the one-sided published criterion: a diverged structure carries an enormous *negative* E_{hull} (a measured Carbon-24 case lands 58 eV/atom below its hull), and the one-sided test would count that as a stability success, so a candidate below the floor is counted as *not stable*—it stays in the stability denominator and fails—and is excluded only from the reported E_{hull} median, with the affected fraction, the median excursion and the minimum reported alongside as `n_unphysical`, `unphysical_rate` and `unphysical_e_hull_median`. A record whose composition has no calibrated formation energy has no defined stability value at all and drops out

of that denominator; that count is not surfaced as a separate field, so the stability denominator is the calibrated candidates rather than all candidates. E_{hull}, and with it the stability leg of S.U.N., is therefore comparable across methods *within* a dataset but not across datasets. All reported E_{hull} values are per-atom energies in eV/atom.

A.4 The Analytic Pauli Correction

Covalent radii. We use the Cordero et al. tabulation of single-bond covalent radii, which covers H–Cm; for the elements appearing in our datasets the radii range from 0.31 Å (H) to 2.58 Å (Cs).

Screening-potential matching. The primary parameterization derives from the Ziegler–Biersack–Littmark universal screening function

$$U_{\text{ZBL}}(r_{ij}) = \frac{1}{4\pi\varepsilon_0}\frac{Z_i Z_j e^2}{r_{ij}}\,\Phi\left(\frac{r_{ij}}{a_{\text{scr}}}\right), \tag{5}$$

with $a_{\text{scr}} = 0.8854\,a_0/(Z_i^{0.23} + Z_j^{0.23})$, $a_0 = 0.529\,\text{Å}$, and $\Phi(x) = 0.1818e^{-3.2x} + 0.5099e^{-0.9423x} + 0.2802e^{-0.4029x} + 0.02817e^{-0.2016x}$. Because the ZBL form carries explicit nuclear charges, which is inconvenient for compositional generalization, we match the exponential form of Eq. (2) to the ZBL energy *and force* at $r^* = f_{\text{typ}}(R_i^{\text{cov}} + R_j^{\text{cov}}) - 0.1\,\text{Å}$, just inside the activation threshold. For $U = Ae^{-Br}$, force matching gives $B = F_{\text{ZBL}}(r^*)/U_{\text{ZBL}}(r^*)$ and $A = U_{\text{ZBL}}(r^*)e^{Br^*}$ in closed form. The parameters are symmetric under exchange of the pair.

Table 4: Representative Pauli parameters produced by the matching procedure, for element pairs common in our datasets. The complete table for all tabulated elements is provided as supplementary data.

Pair	A (eV)	B (Å^{-1})	**Pair**	A (eV)	B (Å^{-1})
C–C	4.3×10^2	3.33	Ti–O	6.9×10^2	2.91
Si–Si	8.0×10^2	2.95	Ba–O	6.9×10^2	2.67
O–O	9.6×10^2	3.74	Ca–O	5.2×10^2	2.77
Si–O	8.5×10^2	3.25	Sr–O	6.7×10^2	2.74
Li–O	1.8×10^2	2.84	Na–Cl	4.4×10^2	2.62
Li–S	2.1×10^2	2.68	Zr–O	9.0×10^2	2.89

Entry-force probe calibration. The matched wall is additionally calibrated at its activation radius against the potential's own learned short-range force. For each element pair of the validation compositions, a two-atom probe places the pair at $r_{\text{wall}} = f_{\text{typ}}(R_i^{\text{cov}} + R_j^{\text{cov}})$ in a large cell and measures $F_{\text{NNP}}(r_{\text{wall}})$ by two-point finite differencing ($\sim$3 minutes per pair, no sampling). The tabulated amplitude is then scaled by a per-pair constant $\eta_{Z_iZ_j}$ chosen so that $F_{\text{Pauli}}(r_{\text{wall}}) \geq 2.5\,F_{\text{NNP}}(r_{\text{wall}})$: the wall must dominate the potential's force at the boundary, which may be attractive or extrapolated, or the augmentation inherits the extrapolation it is meant to repair. Over the 24 pairs of the five validation compositions the unscaled ratio $F_{\text{Pauli}}/F_{\text{NNP}}$ at r_{wall} spans 1.4 to 32.6 (median 12.4; ≥ 10 for 19 of the 24 pairs), so the requirement is already met at $\eta = 1$ in 23 of the 24 cases. The single exception is Li–Co, where $F_{\text{Pauli}}(r_{\text{wall}}) = 1.43\,F_{\text{NNP}}(r_{\text{wall}})$ and the loaded constant is the per-pair scale $\eta = 2.5/1.43 = 1.75$, which is what places that wall at $2.5\times$ the learned force rather than merely matching it—the ratio the text quotes for Li–Co is the scaling factor, not a measured force ratio. The calibration is a per-pair constant loaded at construction and adds no sampling-time cost.

Smooth step and neighbor list. The activation step is $\Theta(d) = \frac{1}{2}[1+\tanh(d/w)]$ with $w = 0.05$ Å and analytic gradient $\Theta'(d) = \frac{1}{2w}\,\text{sech}^2(d/w)$, so the wall is a smooth window centred on the threshold rather than a hard switch: it is $\Theta = \frac{1}{2}$ exactly at $r = r_{\text{wall}}$, and it does not vanish on either side of it. The tail is cut rather than zero, and the cut is placed where the residual is negligible. A neighbor list with element-adaptive cutoff $2\max_i R_i^{\text{cov}} + 5w + 0.3$ Å (up to $\sim$5.7 Å for the heaviest elements) is rebuilt each step; because $2\max_i R_i^{\text{cov}} \geq R_i^{\text{cov}} + R_j^{\text{cov}}$ for every pair in the list, that cutoff always reaches at least $5w + 0.3$ Å past the pair's own wall, where $\Theta = \frac{1}{2}[1 + \tanh(-5)] =$

4.5×10^{-5}, and the pair sum additionally skips any pair beyond $8w$, where $\Theta = 1.5\times10^{-6}$. Over the twelve tabulated pairs the force left in the window between $5w$ and $8w$ is at most 1.9×10^{-2} eV/Å (vs. wall forces of order 10^0–10^1 eV/Å at the boundary), falling below 3×10^{-5} eV/Å past $8w$, and the energy dropped by the cut is $< 10^{-4}$ eV per pair. The truncation is therefore not C^1 at the list radius, but the force discontinuity is at most $\sim 2 \times 10^{-2}$ eV/Å—two to three orders of magnitude below the forces acting on a compressed pair—and it occurs only on pairs that have already left the wall's window and are subject to the drift caps of §4.2. The pair sum costs $\mathcal{O}(N\, n_{\text{neigh}})$ with $n_{\text{neigh}} \ll N$ in typical crystals.

A.5 Sampling Protocol, Schedules, and Numerical Safeguards

Initial conditions. The default prior is isotropic Cartesian noise at $\sigma_{\max}$, applied per atom and converted to fractional coordinates by $\boldsymbol{f}_i = \boldsymbol{f}_i^0 + \boldsymbol{d}_i L^{-1}$. Fractional-coordinate initialization is not used: on short-axis cells ($\sim$4.5 Å) even $0.1\sigma_{\max}$ of fractional noise displaces atoms by $\sim$0.48 Å, driving $r_{\min}$ into the OOD regime at the lowest σ of the scan. The lattice is initialized from the reference structure's cell—the lowest-E_{hull} test-set structure of the composition—with strain-isotropic multiplicative noise in log-Gram space at $\sigma_{\text{latt}} = 0.1\sigma_{\max}$: isotropic Gaussian noise is added to $\log(LL^\top)$, symmetrized to a pure strain, and mapped back through a Cholesky factor. This parameterization is always positive-definite and strain-isotropic, whereas per-vector multiplicative noise can flip a lattice vector and destroy the cell. The default prior is therefore a *noised reference structure at the full* $\sigma_{\max}$: the correct test structure and its cell are supplied, displaced by Cartesian noise of one to five Å across the scan. Every primary result in this paper is accordingly a recovery measurement in the sense of §6.1, not a composition-only generation result. The two alternative priors of the A5 protocol in Appendix C—a composition-only prior (uniform fractional coordinates in a cell taken from a same-composition train/val structure—never the test structure—scaled to the reference's atom count, so the composition and the atom count are supplied but no coordinate is) and a *local* prior (the reference structure displaced at $0.1\sigma_{\max}$ rather than at $\sigma_{\max}$)—have not been run; the phrase "noised known structures" in that protocol means the local variant, and not the default. The atom count is fixed to the ground-truth composition's count, or to the most common stoichiometry in the test set for variable-stoichiometry compositions.

Schedules. ALD uses geometric spacing, $\sigma_k = \sigma_{\max}(\sigma_{\min}/\sigma_{\max})^{k/K}$ with $\sigma_{\min} = 0.01$ and total evaluations $K \times M$. PF-ODE integrates $dx/d\sigma = -\sigma\, s^{\text{aug}}$ backward from $\sigma_{\max}$ to $\sigma_{\min}$ with the Dormand–Prince 5(4) pair (Dormand & Prince, 1980) (`scipy solve_ivp` (Virtanen et al., 2020), method `"RK45"`), relative tolerance 10^{-4}, absolute tolerance 10^{-6} on positions, and a maximum of 1000 steps. With the raw Boltzmann score the ODE is stiff on trajectories that cross the Pauli wall, because the exponential repulsion collapses the adaptive step size; the cap bounds that cost, and capped trajectories are flagged non-converged and evaluated on their own merits. The flag is worth more than a caveat: splitting each cell by it shows the capped group carrying 97% of the sampler's induced OOD events, so it separates trajectories that never reached the low-noise regime from trajectories that did (§C.3, `scripts/analyze_pfode_truncation.py`). Measuring where a capped trajectory's state last moved shows it is barely annealed at all: the median sits at $\sigma = 0.74$–0.81 at $\sigma_{\max} = 1$ but at 4.95–4.99 at $\sigma_{\max} = 5$, i.e. 19–27% and $\leq 1\%$ of the schedule (§C.3, `scripts/probe_pfode_reached_sigma.py`). The cyclical schedule is $\sigma(t) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})|\sin(\pi t/T_{\text{cycle}})|$ with $T_{\text{cycle}} = 0.2\, T_{\text{total}}$, and the uncertainty-adaptive schedule holds σ constant whenever the ensemble disagreement exceeds its threshold and otherwise decays it at $\eta = 0.01$ per step (Appendix E).

Step size and drift caps. The ALD update is $x_{t+\Delta t} = x_t + \alpha_k s^{\text{aug}}(x_t) + \sqrt{2\alpha_k}\,\sigma_k \boldsymbol{\epsilon}$: the score enters un-normalized, $\boldsymbol{\epsilon}$ is drawn isotropic in the Cartesian chart and converted to fractional coordinates by L^{-1} for the step (and back by L for the evaluation), so the update applied carries drift $-\alpha_k\beta\nabla U$ and diffusion coefficient $D = \alpha_k\sigma_k^2$, whose Einstein-relation stationary law is the $e^{-\beta U/\sigma_k^2}$ of Theorem 1—the σ_k^2 in the exponent comes from the noise coefficient, not from a division of the score. The per-level step size is $\alpha_k = \alpha_0(\sigma_k/\sigma_{\max})$ with $\alpha_0 = 10^{-3}$, and a per-atom drift cap of $3\sigma_k$ bounds the unbounded raw force scale (compressed bonds yield $|F| \sim$ 10–10^3 eV/Å, whose uncapped drift reaches multi-Å steps that wrap the periodic cell and collide atoms). When Layer 1 is active the cap is tightened near the wall by the wall-relative closing and opening caps of §4.2: a compressed pair's approach is throttled at $(r_{\min} - 0.5)/4$ and its repulsive escape at

$(r_{\text{wall}} + 8w - r_{\min})/2$. A pre-registered exploration-revision proposal (decoupling the noise scale; adding a σ-floor tail) was tested and *rejected* by its gates: the premise that ALD collapses to $\sim$1 Å of travel was a measurement artifact, and unified re-measurement showed ALD travels at least as far as PF-ODE, so the schedule stands unchanged.

Lattice updates. When enabled, the lattice uses a separate step $\alpha_{\text{lattice}} = 0.01\,\alpha$ and is projected onto the strain subspace via the velocity-gradient symmetrization $A = \Delta L\, L^{-1}$, $\Delta L_{\text{sym}} = \frac{1}{2}(A + A^{\top})L$. The stress tensor is symmetric, so only six lattice degrees of freedom are physical, and the raw nine-component update drives a spurious rotation gauge whose accumulated drift can carry the cell out of the positive-definite domain. For both samplers the stress-driven update engages only below $\sigma_k < f_{\text{reopen}}\sigma_{\max}$ with $f_{\text{reopen}} = 0.1$: the stress tensor is trusted only near equilibrium, and an ungated update during the high-noise phase is exactly the mechanism that triggers lattice collapse.

Reflection, noise modulation, and monitoring. The reflection operator iterates over all violating pairs until convergence (typically 2–5 passes, capped at 20) because reflecting one pair can create new violations. Density-adaptive noise uses $\rho_i = \sum_{j \neq i} \exp(-(r_{ij}/r_{\text{cut}})^2)$ with $r_{\text{cut}} = 4.0$ Å and $\gamma = 0.5$. The Step-Monitor rejects a step if (a) $U > 10^4$ eV/atom, (b) $\max_i \|\boldsymbol{F}_i\| > 10^3$ eV/Å (potential-only), (c) $\min_{i<j} r_{ij} < 0.3$ Å, or (d) $|\det L|/|\det L_0| > 10$ against the trajectory's own initial cell, which detects cumulative inflation that a per-step ratio would miss. Its phase-aware decay uses $p = 2.0$, $\gamma_{\text{base}} = 0.5$.

A.6 Baselines

MD simulated annealing: ASE `Langevin` with the eSEN calculator, linear cooling from 2000 K to 100 K over $\{2000, 5000, 10^4, 2 \times 10^4, 4 \times 10^4\}$ steps, $\Delta t = 1.0$ fs, friction $\gamma = 0.01\,\text{fs}^{-1}$, ten random initializations per composition in a cell dilated to $1.5\times$ the equilibrium volume, the lowest-energy snapshot selected and further relaxed with FIRE (Bitzek et al., 2006). *Basin hopping*: Metropolis basin hopping at $T_{\text{BH}} = 1000$ K, where each step displaces all atoms by $\sigma_{\text{disp}} = 0.3$ Å and relaxes the result with FIRE (Bitzek et al., 2006), for $\{50, 100, 200, 500\}$ hops (Wales & Doye, 1997). *Learned baselines*: DiffCSP (Jiao et al., 2023), DiffCSP++ (Jiao et al., 2024), CDVAE (Xie et al., 2022), and MatterGen (Zeni et al., 2025) with official checkpoints where public, otherwise published numbers cited as such, plus a random-plus-relaxation reference (1000 random structures per composition relaxed with FIRE (Bitzek et al., 2006) on the same potential surface).

A.7 Defense Architecture and the Divergence Guard

The defense architecture consists of the three preventive layers that the ablation of §5.2 identifies as the method, plus one optional guard that is not part of any recommended configuration. L1 supplies the correct short-range gradient before the step; L2 reduces the noise amplitude by local density under the Einstein relation, $\sigma_i = \sigma_{\text{base}}/(1 + \gamma\rho_i)$, with ρ_i the neighbour count within a fixed radius; L3 projects any violating step back onto the constraint manifold. L4 is a post-step rejection test with a phase-aware noise decay; it targets numerical divergence — energy, force, or volume blow-up — not the completeness boundary, and the ablation finds it net-harmful at the operating point, so it is documented as an ablation. Table 5 summarizes what each component acts on and what it targets.

Why the guard is not adopted. Adding the divergence guard is what makes high noise expensive: its arm's median E_{hull} rises from 1.3 meV/atom at $\sigma = 0.1$ to 968 at $\sigma = 5$, with a p90 inflated to 9.1×10^4 meV/atom at $\sigma = 1$ and 3.4×10^5 at $\sigma = 5$ by an absorbing state of the guard's own making, and terminal validity falling to 90.9% at $\sigma = 1$ while the guard-free stack holds 99.9–100%. The guard's rejection test is evaluated on the pre-move structure, so a trajectory that crosses the force threshold cannot leave it: every later step is rejected, the previous state is restored, and the trajectory ends frozen in the diverged configuration ($\max_i \|\boldsymbol{F}_i^{\text{NNP}}\| \sim 10^5$ eV/Å, $r_{\min} < 0.5$ Å; the rejection counter is positive on every frozen trajectory of a guard-containing arm). The count of cell-expansion outliers matches the unprotected baseline.

Feasibility is bounded and diagnosed, not asserted. Algorithm 1 does not return a point that is on $\mathcal{M}$ by construction, and the theorem it approximates is not a statement about the pseudo-code as written. Three things stand between the projection and a membership guarantee. (a) The

Table 5: Defense architecture. L1–L3 are the preventive layers and define the recommended configuration; L4 is an optional divergence guard, evaluated as an ablation in §5.2.

Layer	Mechanism	Stage	Prevents	NFE cost
L1: Pauli-augmented score	long-range fidelity + short-range repulsion	pre-step (score)	drift into the OOD region	0
L2: density-adaptive noise	local-density-modulated σ_i	intra-step (noise)	noise-induced collisions	0
L3: reflecting boundary	hard geometric constraint	intra-step (normal projection)	inner and outer wall penetration	0
L4: divergence guard (optional)	rejection + phase-aware noise decay	post-step (accept)	numerical divergence, at a measured cost in recovery	0–1

Algorithm 1 Reflected Langevin step

$\tilde{x} \leftarrow x_t + \alpha\, s^{\text{aug}}(x_t) + \sqrt{2\alpha}\,\sigma\,\boldsymbol{\epsilon}$ ▷ un-normalized score; $\boldsymbol{\epsilon}$ isotropic Cartesian, mapped to fractional by L^{-1}; drift cap in units of σ_k; wall-relative caps if Layer 1 is active
$p \leftarrow 0$; conv $\leftarrow$ FALSE
while $p < 20$ **and not** conv **do** ▷ atomic sweep, tolerance $\varepsilon = 10^{-6}$ Å
  $\mathcal{V} \leftarrow \{(i,j) : r_{ij}(\tilde{x}) < f_{\text{typ}}(R_i^{\text{cov}} + R_j^{\text{cov}}) - \varepsilon\}$
  if $\mathcal{V} = \emptyset$ **then** conv $\leftarrow$ TRUE; **break**
  end if
  for each $(i,j) \in \mathcal{V}$ **do** ▷ one sequential pass; reflecting a pair can create new violations
    $\boldsymbol{n} \leftarrow (\boldsymbol{r}_i - \boldsymbol{r}_j)/\|\boldsymbol{r}_i - \boldsymbol{r}_j\|$; $\delta \leftarrow f_{\text{typ}}(R_i^{\text{cov}} + R_j^{\text{cov}}) - r_{ij}(\tilde{x})$
    $\tilde{\boldsymbol{r}}_i \mathrel{+}= \frac{1}{2}\delta\boldsymbol{n}$, $\tilde{\boldsymbol{r}}_j \mathrel{-}= \frac{1}{2}\delta\boldsymbol{n}$ ▷ projection onto the wall along the separation direction
  end for
  $p \leftarrow p + 1$
end while
if not conv **and** lattice updates enabled **then** ▷ cell collapsed below the minimum-image constraint
  for $e = 1$ **to** 3 **do** ▷ expansion clamped to the reference volume: $|\det \tilde{L}| \le |\det L_0|$
    $\tilde{L} \leftarrow c\,\tilde{L}$, c the smaller of the resolution factor and $(|\det L_0|/|\det \tilde{L}|)^{1/3}$; retry the atomic sweep
    if conv **then** **break**
    end if
  end for
end if
if $|\det \tilde{L}| > f_{\text{out}}|\det L_0|$ **then**
  $\tilde{L} \leftarrow \tilde{L}\,(f_{\text{out}}|\det L_0|/|\det \tilde{L}|)^{1/3}$ ▷ outer wall: isotropic cell rescale
  retry the atomic sweep once, overwriting conv ▷ the downscale shrinks min-image distances, so fresh inner violations are possible
end if
$x_{t+1} \leftarrow \tilde{x}$; **return** (conv, $|\mathcal{V}(\tilde{x})|$) ▷ membership is *verified and reported*, not guaranteed by construction: on FALSE the residual violation count is returned and Layer 4 backstops

reflected step must be an operator with a fixed point. Projecting one pair can push a third atom into another pair—the reflection changes the configuration, so the sweep must be iterated rather than performed once—and a finite sweep terminates at a fixed point only if one is reached within the pass cap. The implementation sweeps sequentially over violating pairs at tolerance $\varepsilon = 10^{-6}$ Å for at most 20 passes, and converges when no pair remains below the threshold minus that tolerance; 2–5 passes is typical. The cap is a budget, not a contraction argument, and it is not proven that a configuration always converges within it. (b) A cell that has collapsed below what the minimum-image constraint can accommodate has no fixed point at all in the atomic sweep. The operator then attempts up to three isotropic expansions, each clamped to the trajectory's reference volume

$|\det L_0|$; the clamp is not cosmetic, since an unbounded expansion retries the sweep at a larger cell, relaxes, clashes again, and inflates once more, up to the f_{out} cap with E_{hull} destroyed and the OOD counter never triggered. (c) The outer wall acts *after* the atomic sweep and is followed by a cell rescale, which shortens minimum-image distances and can create fresh inner violations; the atomic sweep is therefore retried once, and the retry overwrites the convergence flag. Degenerate cells (volume below 10^{-6} Å^3) bail out immediately with FALSE. In every failure path the operator computes its convergence flag together with the residual violation count after the last pass, so non-convergence is *detected*; it is not *propagated*. Both samplers consume the projected structure and neither reads the flag, so a non-converged step is carried forward like any other and the only layer that can reject it is the optional L4 monitor—which the ablation finds net-harmful and which the recommended configuration does not include. No per-step non-convergence statistic is therefore available in the released artifacts: the constraint's effectiveness is measured indirectly, through the induced-OOD and $r_{\min}$ distributions of §5.2, not through a reported pass rate. The claim in §4.3 that reflection is a hard constraint is accordingly a statement about the converged case—the case the sweep reaches in the overwhelming majority of steps, judged by the $r_{\min}$ distributions, but not a rate we record. Theorem 4 is correspondingly a statement about the idealized fixed-cell reflected process, not about this operator, and Remark 4 separates the two.

A.8 METRICS

Metrics are computed per composition and macro-averaged, with error bars of one standard deviation across the five seeds.

- **Validity**: composition matches the target and $r_{\min} \geq 0.5$ Å under the minimum-image convention, following DiffCSP. Diverged trajectories (non-finite positions or forces) count as invalid and contribute zero to all downstream metrics.
- **S.U.N. rate**: fraction of generated structures that are *stable* ($-50 \leq E_{\text{hull}} < 100$ meV/atom, the two-sided band of Appendix A, with the hull built per chemical system from the dataset's own phases—MP-20's Materials Project formation energies, Perov-5's `heat_all` table, elemental carbon for Carbon-24—and the candidate's NNP energy mapped onto that hull's DFT formation-energy scale by the per-composition constant shift defined there), *unique* (no other generated structure of the same composition shares space group, volume within 10%, and E_{hull} within 20 meV/atom), and *novel* (no match to the training set under `StructureMatcher` with `ltol` $= 0.2$, `stol` $= 0.3$, `angle_tol` $= 5.0$).
- **Match rate**: fraction of generated structures matching a ground-truth test structure under the same `StructureMatcher` tolerances.
- **Coverage**: fraction of test-set structures recovered by at least one of the generated candidates.
- **AMSD**: mean over ensemble members of the minimum Kabsch-aligned RMS distance to another member; higher is more diverse.
- **R-Angle KL**: KL divergence of the bond-angle histogram (36 bins of 5°, triplets within 3.0 Å) against the test-set reference.
- **NFE**: force evaluations per structure, counted as $K \times M$ for ALD and as RHS evaluations for adaptive PF-ODE; per successful sample, divided by the S.U.N. rate.
- **OOD spike rate**: the sampler-induced rate of §5.1. An *event* is any accepted step at step ≥ 1 from a clean start with $r_{\min} < 0.5$ Å, $\max_i \|\boldsymbol{F}_i^{\text{NNP}}\| > 500$ eV/Å, or $|\det L|/|\det L_0| > 10$; the rate is the clean-start numerator over all trajectories, and *start-OOD*—an initial condition that already violates the constraint, reported separately—is excluded from it. Integrator diagnostics are taken on the accepted-step path, since intermediate stages are extrapolation states rather than trajectory states. The thresholds are deliberately looser than the Type I completeness threshold—criterion (a) marks only deep, unambiguous overlap, so the spike rate is a conservative measure and the two diagnostic scales are not interchangeable. The electronic-collapse criterion ($U < -50$ eV/atom) is retained as a diagnostic bit and is never counted toward the rate.
- ΔE: potential energy after relaxation minus that of the original structure relaxed on the same surface; negative means the sampler found a lower-energy configuration than the known structure.
- **Curl violation** and **angular deviation** ϕ_{ang}: defined in Appendix G.

Statistical protocol. Every experiment is repeated with five seeds. Pairwise comparisons (ALD versus PF-ODE; with versus without augmentation) use a two-sided Wilcoxon signed-rank test over per-composition values paired by composition, at $p < 0.05$ with Bonferroni correction within each experiment; multi-configuration comparisons use the Benjamini–Hochberg procedure at FDR $q = 0.05$. Confidence intervals are 95% bootstrap intervals (10,000 resamples, stratified by composition). Every rate is computed over all candidates for which the quantity is defined, and nothing is dropped for being low-energy or invalid: diverged trajectories count as invalid and contribute zero to every downstream metric, and a candidate whose E_{hull} falls below the -50 meV/atom floor is counted as *not* stable rather than removed, with the affected fraction reported beside the rate it would otherwise inflate. Two quantities are defined on a restricted set, and each restriction is labelled where it appears: the E_{hull} median is taken over valid, above-floor candidate energies within each cell, with a cell holding no such value dropping out of the per-cell average rather than contributing a zero (on Perov-5 that leaves 12–16 of the 20 cells, Appendix G.2); and the OOD spike rate is defined on clean-start trajectories only, with the start-OOD share reported separately (§5.1). The fraction of diverged trajectories is reported separately as well.

A.9 The Boundary, the Threshold, and the Sampler Scale

Three quantities are distinguished in this paper, and conflating them would overstate what is measured. They are set out here once, with what is known about each and what would be needed to close the remaining gaps.

Table 6: Three quantities that are separate objects. Only the first is a free choice, only the third is a sampler property, and only the third is measured as a scale in this work.

	What it is	**How it is fixed here**	**What would establish it**
(i) The protection threshold	The radius at which the analytic wall switches on	*Chosen*: $f_{\text{typ}} = 0.75$, calibrated against the training-corpus edge below	A calibration target other than the corpus edge; the value is a design parameter, not a measurement
(ii) Where the potential's forces begin to deviate from the true forces	A property of the learned model and of the reference it is compared against	*Not measured.* No DFT force curve across the threshold exists in this study; §A.4 probes the potential's own pair force, not its error	A two-atom or low-coordination DFT force scan over r across the threshold, for several element pairs, against the same functional the training data used
(iii) The noise level at which a given sampler collapses	A joint property of the potential, the sampler, its step size, its evaluation budget and the initial conditions	*Measured*, as σ_c of Table 16, with the estimator dependence of Appendix H	Nothing further; but it is a statement about a sampler, so it moves when the sampler does

The training-corpus edge, measured. Fig. 1(a) is a schematic, so the quantity it illustrates is measured directly here. Reading the MP-20 CIF corpus, forming every minimum-image pair distance and normalising it by the sum of single-bond covalent radii, $u = d_{ij}/(R_i^{\text{cov}} + R_j^{\text{cov}})$, gives a lower edge at $u = 0.709$ over 1,186,442 pairs in the 17,286 training structures: no pair in the corpus lies below $u = 0.70$, and 0.014% lie below the protection threshold of $f_{\text{typ}} = 0.75$. The validation and test portions behave the same way (773,062 pairs, minimum $u = 0.717$), as they must, since the three portions are drawn from the same relaxation pipelines. The statement is stronger at the level of structures than at the level of pairs, which is the level at which a sampler meets it: over the whole corpus of 28,689 structures and 1,959,504 pairs, only 0.43% of structures contain even a single pair below $f_{\text{typ}} = 0.75$, the median structure's closest pair sits at $u = 0.936$, and its first percentile is 0.764. A configuration is thus outside the training distribution in the short-range sense not when its contacts are merely atypical but when its closest one is compressed past a floor that all but a fraction of a percent of the corpus respects. Structures containing the elements for which no covalent radius is tabulated are skipped and counted (16,430 such structures; the lanthanides). Two

consequences are worth stating precisely and one is not. It is supported that the protection threshold sits at the empirical edge of the data rather than in the bulk of it, and that the edge is a property of the corpus rather than of a particular architecture, since the corpus is a union over many relaxation and calculation settings. It is *not* established by this statistic that the potential's forces are wrong at $u < 0.75$: the corpus edge is a statement about where data are, not about where the model errs, and the two would coincide only if the model's short-range behaviour were determined by the sample count alone. The gap between (i) and (ii) is exactly that step, and this study does not take it.

The potential's own short-range behaviour, measured. What is measured on the potential's side is its own pair force, in two probes. The entry-force probe of §A.4 places an isolated pair at the protection threshold and reads $F_{\mathrm{NNP}}(r_{\mathrm{wall}})$ by finite differences for the 24 element pairs of the five validation compositions; over those pairs the analytic wall is 1.4 to 32.6 times the learned force at the same radius, and the learned force itself is what the calibration constant is set against. A second probe compresses the shortest bond of each of the 20 scan compositions along $d = \{0.5, \ldots, 1.05\}\, r_{\mathrm{NN}}$ with all other atoms fixed and records the potential's energy and pair force as functions of that coordinate (`scripts/probe_pair_potential.py`). Its result is the one that matters for the universality claim: at $d = 0.5\, r_{\mathrm{NN}}$—a compression of 50% on bonds whose equilibrium length is 1.8–2.6 Å, i.e. 0.9–1.3 Å absolute—the potential's pair force spans 0.3 to 931 eV/Å across the twenty compositions—a factor of 2,900 for the same nominal deformation. The weakest are LiF (0.32 eV/Å), AlCo (3.9), CoF_3 (4.0) and $LiCoO_2$ (4.9); the strongest are GaN (931), MoS_2 (428), $BaTiO_3$ (378) and ZrO_2 (298). A universal repulsion would not depend on the chemistry like this, which is the premise of supplying it analytically rather than expecting it to be learned; and the same probe at $0.75\, r_{\mathrm{NN}}$ is the a-priori σ_c predictor of Table 16, which under-predicts σ_c by 1.5–2.2× on the three compositions it resolves, so the probe is a lower bound and a safe basis for choosing $\sigma_{\max}$. What this does not do is locate (ii): a force curve probed on a single displaced bond of a relaxed structure is a statement about the potential, not about the potential's error, and no reference (DFT) curve at those distances is computed here.

What would move the boundary from a definition to a measured location. Three experiments would, and none is in this study: a DFT force scan across the threshold on a few element pairs in the same functional as the training data, which would fill the (ii) row directly; a held-out-composition test in which f_{typ} is calibrated on some chemistries and evaluated on others, which would test whether the corpus edge transfers; and a repulsion-off arm at matched drift caps, which would separate the analytic wall's effect from the wall-relative caps' on a single sampler (Appendix C.2, where the PF-ODE contrast supplies the wall's isolated effect instead). Until the first of these exists, the honest description of $f_{\mathrm{typ}} = 0.75$ is an *effective short-range protection threshold* placed at the measured edge of the training distribution, rather than a boundary whose location has been verified against the potential's true error. What is claimed for it as a completeness boundary is the definitional part—the score is complete where the data support it, and the corpus edge is where that support ends—together with the fact that a threshold placed at the edge protects the sampler, which is what §5.2 measures.

Position relative to repulsion-corrected potentials. Two prior works are closely related and are not overclaimed against. MACE-MP-0 itself applies a ZBL-based repulsion with a distance-based switch calibrated to covalent radii in its random-structure-search applications; that is the same mechanism this paper supplies analytically, and the difference is not the presence of a repulsive term but which stage of the pipeline is modified—a learned model fitted to include short-range data, versus a training-free post hoc augmentation of a frozen score with a conservativity guarantee and a stated steady state. Byggmåstar et al. (Byggmästar et al., 2019) build a machine-learned potential for radiation damage and collision cascades, i.e. for exactly the high-energy short-range regime, and citing that line as an equilibrium-region correction was an error in an earlier version of this work; the correct contrast is that they train on the short-range regime while this paper supplies it analytically. What is new here relative to both is the diagnostic: treating the boundary as a measurable object, measuring the noise scale at which a sampler driven by a pre-trained score fails, and showing that an analytic wall plus a hard geometric constraint holds that failure rate down without additional force evaluations. Whether the analytic form is preferable to including short-range data is an empirical question that this study does not answer—it is the repulsion-off comparison above.

B PARAMETER CONFIGURATIONS

The configurations below form a controlled-variable study over the scheduler parameters, with each configuration differing from the baseline in one or two settings, plus four protection configurations that enable the preventive layers. The benchmark protocol of Appendix G evaluates the streamlined subset C1, C3, C5, C8, C9, C11; the full design space is documented here for completeness. All ALD configurations use geometric noise spacing with $\sigma_{\min} = 0.01$ and no Metropolis rejection. The tabulated step sizes are to be read with the normalization of Appendix A: the effective base step is $\alpha_0 = 10^{-3}$ with a per-atom drift cap of $3\sigma_k$; the relative ordering of the configurations (C2 conservative, C3 exploratory, and so on) is preserved. Protection configurations add Layers 1–3 at zero extra evaluations (C11 adds Layer 1 alone); the optional Layer 4 guard is not part of them.

Table 7: Eleven parameter configurations. C1–C7 vary scheduler parameters around the baseline; C8–C11 add the preventive layers. All configurations use $\sigma_{\min} = 0.01$ and, where lattice updates are enabled, the $\sigma < 0.1\sigma_{\max}$ gate of Appendix A. Every PF-ODE configuration uses the same adaptive integrator — DOPRI5(4) (Dormand & Prince, 1980), i.e. `scipy solve_ivp` (Virtanen et al., 2020) method `"RK45"`, at rtol 10^{-4} and atol 10^{-6} on positions with a uniform cap of $N_{\max} = 1000$ evaluations — so no configuration specifies a step size; the cap is inert below $\sigma_{\max} = 0.5$, already burns 11.8–16.1% of trajectories there, and binds from $\sigma_{\max} \approx 1$ upward (61–68%), where the stiff wall-crossing trajectories burn their budget and stall near where they started (median last evaluation at $\sigma = 4.95$ of 5).

#	Name	ALD parameters	PF-ODE parameters	NFE (ALD)	Key variable
C1	baseline	$\sigma_{\max}$=0.5, β=38.68, K=100, M=2, α=0.1, α_{lat}=0.001, upd_lat	$\sigma_{\max}$=0.5, β=38.68, DOPRI5(4), rtol 10^{-4}, atol 10^{-6}, $N_{\max}$=1000, upd_lat	200	reference point
C2	conservative	$\sigma_{\max}$=0.2, β=38.68, K=200, M=2, α=0.03, α_{lat}=0.0005, upd_lat	$\sigma_{\max}$=0.2, β=38.68, DOPRI5(4), rtol 10^{-4}, atol 10^{-6}, $N_{\max}$=1000, upd_lat	400	low-noise limits
C3	exploratory	$\sigma_{\max}$=1.5, β=11.60, K=100, M=2, α=0.12, α_{lat}=0.002, upd_lat	$\sigma_{\max}$=1.5, β=11.60, DOPRI5(4), rtol 10^{-4}, atol 10^{-6}, $N_{\max}$=1000, upd_lat	200	high-noise exploration
C4	fast	$\sigma_{\max}$=0.3, β=38.68, K=30, M=1, α=0.15, α_{lat}=0.001, upd_lat	$\sigma_{\max}$=0.3, β=38.68, DOPRI5(4), rtol 10^{-4}, atol 10^{-6}, $N_{\max}$=1000, upd_lat	30	minimal budget
C5	fixed_cell	C1, no lattice update	C1, no lattice update	200	lattice updates off
C6	high_temp	$\sigma_{\max}$=0.5, β=11.60 (others as C1)	$\sigma_{\max}$=0.5, β=11.60 (others as C1)	200	temperature only
C7	large_lattice_step	C1 with α_{lat}=0.005	C1 with α_{lat}=0.005	200	lattice step only

#	Name	ALD parameters	PF-ODE parameters	NFE (ALD)	Key variable
C8	protected_baseline	C1 + Layers 1–3	C1 + Layers 1–3	200	protection at baseline
C9	protected_exploratory	C3 + Layers 1–3	C3 + Layers 1–3	200	protection at high $\sigma_{\max}$
C10	protected_fixed_cell	C5 + Layers 1–3	C5 + Layers 1–3	200	protection, fixed cell
C11	pauli_only	C1 + Layer 1	C1 + Layer 1	200	augmented score alone

Motivation. C2 tests whether extreme caution compensates for OOD risk. C3 pushes $\sigma_{\max}$ to 1.5 Å, well into the region where the reliability condition becomes marginal, and raises the temperature to 1000 K; it is the configuration that directly probes the completeness boundary. C4 asks what 30 evaluations can achieve. C5 disables lattice updates, the primary ablation for whether stress-driven updates are usable on perturbed structures. C6 varies temperature alone, controlling for $\sigma_{\max}$, and C7 varies the lattice step alone, probing the stability boundary of the cell update. C8–C10 test whether the preventive physics of Layers 1–3 changes the outcome in the standard configuration, at high noise, and on the fixed-cell production setting; C11 isolates the augmented score. The reactive guard (Layer 4) is not part of any configuration here: it is evaluated as an ablation, where it costs recovery (§5.2).

C Supplementary Ablations

Two secondary design questions are answered here rather than in the main text, since neither bears on the completeness boundary. Neither experiment has been executed in the present study: the protocols are given so that they can be reproduced, and no result in this paper depends on them.

- **Temperature (A2).** How does temperature affect the exploration–exploitation balance? Protocol: $\beta \in \{10, 20, 38.68, 50, 100\}$ (corresponding to 1160 K down to 116 K), PF-ODE on a ten-composition MP-20 subset with the C5 configuration; metrics: coverage versus match rate, and AMSD versus β.
- **Prior distribution (A5).** Does the initial distribution matter for generation quality? Protocol: three priors on a ten-composition MP-20 subset—the default (reference structure displaced by isotropic Cartesian noise at $\sigma_{\max}$), a *composition-only* prior (uniform fractional coordinates on the cell of a same-composition train/val structure, scaled to the reference atom count—the composition is supplied, no coordinate of the target is), and a *local* prior (the same reference structure displaced at $0.1\sigma_{\max}$); metrics: coverage, AMSD, S.U.N. rate. The composition-only arm is the one that separates generation from recovery, because the default prior already supplies the reference structure's coordinates and cell; it is also the arm that would answer whether the match rate of §5.3 survives when the target structure is not handed to the sampler, and it has not been run.

The remaining ablations are hosted by the experiments that inform them rather than run as separate studies: the $\sigma_{\max}$ truncation ablation *is* the noise-level scan of §5.2; the defense-layer ablation is the layer grid of Appendix C.2; the Pauli-parameter ablation is the scan of Appendix C.4; the evaluation-budget ablation would be the knee of the energy-versus-evaluation curve in the efficiency protocol; and the potential-selection ablation is the cross-potential protocol, both in Appendix G.

C.1 Defense-Layer Grid

C.2 Layer Marginals in Full

A 2^4 grid at $\sigma_{\max} = 1.0$ Å—the empty set, all singletons, pairs, triples, and the full stack, 1000 trajectories per cell, 12 of the 16 subsets (Table 8)—measures each layer's marginal contribution in several contexts rather than along a single chain, which separates "add L2 to L1" from "add L1 to L2". The design is unbalanced, so averaging over the twelve cell means on each side of an

Table 8: Layer-grid dissection at $\sigma_{\max} = 1.0\,\text{Å}$ (ALD, 200 evaluations, 1000 trajectories per cell, 12 of the 16 layer subsets), from the corrected re-run. *Induced*: sampler-induced OOD rate. *Dwell/deep*: share of accepted steps with $r_{\min} < 1.3\,\text{Å}$ / $< 0.5\,\text{Å}$. E_{hull}: median over valid final structures. L4-attached conditions show the monitor's pinning signature.

Condition	Layers	Induced	Validity	Dwell	Deep	E_{hull} med.
bare	$\varnothing$	32.8%	98.6%	7.1%	1.86%	235 meV
L1	$\{1\}$	18.5%	99.6%	3.9%	0.74%	209 meV
L2	$\{2\}$	32.3%	97.6%	7.9%	2.60%	230 meV
L3	$\{3\}$	0.10%	100.0%	1.2%	0.05%	208 meV
L4	$\{4\}$	33.0%	81.0%	64.8%	18.68%	14.06 eV†
L1+L2	$\{1,2\}$	16.7%	99.7%	3.3%	0.52%	204 meV
L1+L3	$\{1,3\}$	0.10%	100.0%	1.4%	0.05%	212 meV
L1+L2+L3	$\{1,2,3\}$	0.00%	100.0%	1.4%	0.05%	197 meV
L1+L2+L4	$\{1,2,4\}$	16.8%	86.6%	48.3%	13.33%	361 meV†
L1+L3+L4	$\{1,3,4\}$	0.10%	90.9%	29.9%	9.10%	247 meV†
L2+L3+L4	$\{2,3,4\}$	0.10%	90.9%	29.9%	9.10%	244 meV†
L1–L4 (full)	$\{1,2,3,4\}$	0.00%	90.9%	30.0%	9.10%	238 meV†

Notes: every L4-attached condition shows the monitor pinning signature (dwell/deep up, validity down) — full stack: 293/1000 trajectories stay OOD for $\geq$190 of 200 steps, against 10/1000 under bare and 0/1000 under $\{1,2,3\}$; the reflecting wall alone rescues start-OOD pairs, the monitor freezes them. †the L4-alone E_{hull} median (14.06 eV/atom) reflects trajectories collapsed by unbounded drift; values for $\{1,2,4\}, \{1,3,4\}, \{2,3,4\}, \{1,2,3,4\}$ carry the pinning penalty over their L4-free counterparts (204 meV for $\{1,2\}$, 212 for $\{1,3\}$, 208 for $\{3\}$ alone and 197 meV for $\{1,2,3\}$). Condition names `bare`, `l1`, `l2`, `l3`, `l4`, `l112`, `l113`, `l11213`, `l11214`, `l11314`, `l21314`, `l114` refer to the runner's layer subsets; `l114` denotes the full stack, the prefix-ladder name being retained from an earlier convention.

on/off split is *not* a controlled comparison: the two sides differ in their other layers as well, and the imbalance is severe (for L4, three of the five L4-on cells also carry L3 but only three of the seven L4-off cells do). Every layer does, however, admit four pairs inside the same grid that differ in that layer and nothing else, and the controlled marginal is the average of those paired differences. Table 9 gives both, and pairing changes the ranking not at all but the magnitudes a great deal.

The controlled marginals read as follows. L3 dominates at -21.15 pp, and it is the one layer whose effect is large in every context (-32.7 with nothing else on, -16.7 to -18.4 with L1 and/or L2 present). L2 is negligible in every context, at most -1.80 pp and on average -0.62. L4 has no induced-rate marginal at all once pairing is enforced: $+0.20$, $+0.10$, 0.00, 0.00, a mean of $+0.08$ pp, so the unpaired -4.4 pp is entirely a confound with L3 rather than a benefit of the monitor. L4's validity column is another matter, and pairing makes it *larger* and more consistent than the unpaired design suggested: -17.6, -13.1, -9.1, -9.1, a mean of -12.2 pp, negative in all four pairs. Whatever else the monitor does, it costs recovery wherever it is attached, and the two facts together—no induced-rate benefit, a uniform recovery penalty—are what the recommendation against L4 rests on. The L1 marginal is heterogeneous in a way that is substantive rather than noisy: -14.3 and -15.6 pp in the two pairs where L3 is absent, 0.00 and -0.10 pp in the two where it is present. L1 is a strong second contributor where there is no constraint and redundant where there is one—the expected division of labour between a gradient and a constraint, and also the reason the recommended configuration can be argued from either direction. That reading is the strongest its ablation can support on this sampler, because L1 is a bundle: on ALD it switches the analytic wall and the wall-relative drift caps on together (Appendix A.7), so the -14 to -16 pp above is the bundle's marginal. The wall alone is isolated on the deterministic path instead, where no capping machinery exists at all: bare-to-L1 on PF-ODE moves the induced rate from 5.70% to 0.60% at $\sigma_{\max} = 1$ and from 6.30% to 1.70% at $\sigma_{\max} = 5$, with terminal validity rising $84.1 \to 88.7\%$ and $77.8 \to 83.0\%$ over the same two levels. We therefore report the ALD contrast as the bundle's effect and the PF-ODE contrast as the wall's, and we do not have the two-factor decomposition that would separate them on a single sampler: neither a caps-only arm nor a wall-on/off contrast at matched caps was run, and the interaction between wall and caps is therefore not measured here.

Table 9: Paired layer marginals at $\sigma_{\max} = 1.0$ Å (ALD, from the same cells as Table 8). Each row switches one layer with all others held fixed; entries are differences in percentage points, on-off. The final row of each block is the mean of the four paired differences, and the last column is the unbalanced on/off difference over the twelve cell means, shown for comparison. Sets are written explicitly because the runner's condition names are not compositional: `l1l4` denotes $\{1, 2, 3, 4\}$.

Layer	Pair (off → on)	Δ induced	Δ validity	Unbalanced
L1	$\varnothing \to \{1\}$	-14.30	$+1.00$	-12.20
	$\{2\} \to \{1, 2\}$	-15.60	$+2.10$	
	$\{3\} \to \{1, 3\}$	0.00	0.00	
	$\{2, 3, 4\} \to \{1, 2, 3, 4\}$	-0.10	0.00	
	paired mean	$\mathbf{-7.50}$	$+0.78$	
L2	$\varnothing \to \{2\}$	-0.50	-1.00	-3.12
	$\{1\} \to \{1, 2\}$	-1.80	$+0.10$	
	$\{1, 3\} \to \{1, 2, 3\}$	-0.10	0.00	
	$\{1, 3, 4\} \to \{1, 2, 3, 4\}$	-0.10	0.00	
	paired mean	$\mathbf{-0.62}$	-0.23	
L3	$\varnothing \to \{3\}$	-32.70	$+1.40$	-24.95
	$\{1\} \to \{1, 3\}$	-18.40	$+0.40$	
	$\{1, 2\} \to \{1, 2, 3\}$	-16.70	$+0.30$	
	$\{1, 2, 4\} \to \{1, 2, 3, 4\}$	-16.80	$+4.30$	
	paired mean	$\mathbf{-21.15}$	$+1.60$	
L4	$\varnothing \to \{4\}$	$+0.20$	-17.60	-4.36
	$\{1, 2\} \to \{1, 2, 4\}$	$+0.10$	-13.10	
	$\{1, 3\} \to \{1, 3, 4\}$	0.00	-9.10	
	$\{1, 2, 3\} \to \{1, 2, 3, 4\}$	0.00	-9.10	
	paired mean	$\mathbf{+0.08}$	$\mathbf{-12.22}$	

The ordering does not come from the OOD definition. Every headline metric of the scan is defined by a distance threshold, and L3 enforces that same distance by construction, so part of its margin could be the definition returning. Ranking the arms instead by statistics that threshold nothing—the median nearest-neighbour distance over all accepted steps of a cell, the median over trajectories of the closest approach, and the median and 90th percentile of $|F_{\text{NNP}}|$ over final structures—reproduces the induced-rate ordering within a sampler on 85.4% of the pairs the median distance does not tie (48 rankable pairs over the nine noise levels with a non-degenerate ordering), 90.9% for the closest approach, 75.0% for the force tail and 52.4% for the force median. The two statistics that do *not* reproduce it are informative rather than damaging. The median final force ties across the ALD arms, because the unprotected arm's *median* final structure is a relaxed one at every noise level (0.2–0.7 eV/Å median $|F|$); the unprotected arm's damage is entirely in the tail, which is where the 90th percentile separates it. And every cross-sampler comparison inverts: PF-ODE's median nearest-neighbour distance is *smaller* than bare ALD's at all ten noise levels (1.72–1.96 Å against 1.79–2.07 Å), although its induced rate is 4.7–$7.4\times$ lower over $\sigma_{\max} \geq 0.5$ Å. The three lowest levels are the ones where the ODE arm completes essentially every trajectory (Fig. 2(e): the cap takes 0.0% at $\sigma_{\max} \leq 0.2$ and 0.4–0.8% at 0.3), so the inversion is established where truncation is absent; at $\sigma_{\max} = 0.5$ the cap already takes 11.8–16.1%, and it is the capped group alone that carries induced events there, so that one level should be read with the truncation beside it. This is the same asymmetry Proposition 1 leaves to measurement: a trajectory that injects no noise neither crosses the boundary on a large step nor pushes its atoms apart, so the induced rate locates escape from the safe set and not the quality of the geometry sampled, and the two are reported separately here for that reason (`scripts/analyze_threshold_free.py`).

The step monitor is net-harmful at this operating point. It carries no induced-rate marginal and degrades every other axis: with L4 attached, the phase-aware noise decay freezes the start-OOD fraction in the collision state—293 of 1000 trajectories remain OOD for $\geq$190 of 200 steps, versus 10 under bare and 0 under $\{1, 2, 3\}$—cutting validity from 100% to 90.9% and raising the E_{hull} median from 197 to 238 meV/atom. The reflecting wall *ejects* a start-OOD pair, whereas rejection without bounded dynamics lets a trajectory wander into a deep-collapse state that the monitor then

pins. The hierarchy L3 $>$ L1 $\gg$ L2 $\approx$ L4 holds across the scan, and the monitor costs 1.2–1.3$\times$ the L1-alone E_{hull} wherever it is active at $\sigma \geq 1.5$.

C.3 PF-ODE Robustness Is Relative

Proposition 1 shows that determinism contributes no noise term of its own to the score-error budget and leaves the operating contrast to measurement. The data confirm the relative part and refute the absolute part. Unprotected PF-ODE holds at 4.30–6.70% across $\sigma \in [0.5, 5]$ (pooled mean 5.9%, per-composition up to 34% for the worst single composition), which is 4.7–7.4$\times$ safer than unprotected ALD across $\sigma \in [0.5, 5]$ (5.70% versus 33.00% at $\sigma = 1$) but nowhere near OOD-free: determinism removes the noise-injection path, not the large deterministic drift steps that alone can drive a pair across the boundary. Layer 1 closes most of the remainder, bringing PF-ODE to 0.0–2.2%. The same runs bound the high-noise ODE conclusions: the fraction of trajectories exhausting the 1000-evaluation cap is zero at $\sigma \leq 0.2$ (0.4% for the bare score at $\sigma = 0.3$), 12–16% at $\sigma = 0.5$, 61–68% at $\sigma = 1$ and 92.7–98.9% at $\sigma \geq 1.5$, so those results rest on truncated trajectories, flagged non-converged. Splitting each cell by that flag says where the events come from rather than just that they are suspect: pooled over the nine cells that contain both groups, the capped group carries 97.0% of all induced events (584 of 15,119 runs, 3.86%, against 18 of 8,881, 0.20%). At $\sigma_{\max} \in [0.3, 1]$ the completed group sits at 0.00–0.24% induced under all three arms while the capped group runs 0.0–36.4% (the bare score at $\sigma_{\max} = 0.5$ is the extreme; the protected capped groups stay within 6.9%), so the arm's high-noise numbers describe trajectories that never annealed rather than a sampler that is unsafe in the bulk. How little they annealed is measurable. The solver integrates in σ, and once the cap is spent the state stops moving while σ continues to decrease, so a capped terminal structure belongs to the noise level of its last real evaluation rather than to $\sigma_{\min}$; recording that level (`scripts/probe_pfode_reached_sigma.py`, 40 runs per cell: 20 compositions $\times$ 2 seeds) shows the cap all but cancels the schedule. The median capped trajectory's last evaluation sits at $\sigma = 0.74$ (bare) and 0.81 (`l1l4`) at $\sigma_{\max} = 1$—19–27% of the range—falling to 1.86 and 2.00 at $\sigma_{\max} = 2$ ($\leq 7\%$) and 4.95 and 4.99 at $\sigma_{\max} = 5$ ($\leq 1\%$). At $\sigma_{\max} \geq 2$, therefore, a capped structure is to good approximation an un-annealed initial draw: the ODE arm's behaviour there is a statement about the prior and the first few integrator steps rather than about a partially denoised sample. Above $\sigma_{\max} = 2$ the flag stops separating anything, because both groups are compromised: 92.7–98.9% of runs are capped, and the minority that terminate early are collapsed rather than completed structures—median terminal $r_{\min} \approx 0.03$ Å, 100% conditional OOD and terminal validity 0–17% at $\sigma_{\max} = 5$. The integrator's `RK45` method is the standard Dormand–Prince 5(4) pair, so the truncation reflects genuine stiffness of the flow near the Pauli wall rather than a solver mismatch.

C.4 Sensitivity to the Analytic Parameters

Scanning the wall's prefactor over four orders of magnitude and its steepness over a factor of ten for five element pairs qualifies the universality claim rather than confirming it outright: flatness in the prefactor is a steep-wall property. At $B = 10\,\text{Å}^{-1}$ the induced rate varies by at most 13 pp over $A \in [10^2, 10^6]$ eV (at most 5 pp for four of five pairs), so the wall is self-correcting there. At the screening-adjacent $B = 3\,\text{Å}^{-1}$ the metal–oxygen pairs are both sensitive and non-monotone (Ti–O 9–50%, Ba–O 19–49%, Li–O 35–61%): a weak wall leaves the trajectory unprotected while a strong one drives repulsive overshoot into neighbouring atoms. Carbon and silicon chemistry is insensitive throughout. The screening-derived default is mid-range—adequate everywhere, optimal nowhere—so the zero-calibration claim carries the qualification $B \gtrsim 10\,\text{Å}^{-1}$ for metal–oxygen pairs.

D Cross-Potential Recheck on eSEN

The σ scan of §5.2 uses MACE-MP-0, selected by a pre-registered speed calibration. The primary potential of this study is eSEN-30M-MPTrj, so the central claims are re-measured on it: five compositions stratified by chemistry ($SrTiO_3$, ZnS, LiF, GaN, $FeNi_3$) at four noise levels, five seeds and ten candidates each—250 trajectories per cell, 4000 in total (Table 10).

Table 10: Cross-potential recheck: sampler-induced OOD rate and terminal validity on eSEN-30M-MPTrj, the study's primary potential, under the same accounting as Table 15 (250 trajectories per cell). The recheck's four arms are bare ALD, bare PF-ODE, L1 ALD and the full four-layer ALD stack (the runner's `l1l4` arm, which carries the L4 guard on this path); no protected ODE arm was run here, and the protected row is therefore a *superset* of the recommended L1–L3 stack rather than a measurement of it. The distinction is not cosmetic: the guard was net-harmful at the operating point under MACE (Table 8, validity 90.9% versus 100% for L1–L3), whereas on eSEN this arm holds 100.0% validity at every noise level—so the recheck shows the guard's penalty is potential-dependent, while leaving open whether L1–L3 without it would do still better here. The E_{hull} medians quoted in the text (not tabulated) are over valid final structures only, and at high σ the unprotected medians rest on the minority of trajectories that survived, so they are not comparable with the protected arms' across the full cell.

	$\sigma = 0.5$		$\sigma = 1.0$		$\sigma = 2.0$		$\sigma = 5.0$	
Condition	Ind. %	Val. %	Ind. %	Val. %	Ind. %	Val. %	Ind. %	Val. %
Bare, ALD	34.0	71.6	50.8	51.6	51.2	48.0	49.2	55.6
Bare, PF-ODE	7.2	90.8	23.6	69.6	26.4	66.0	28.4	63.2
L1, ALD	0.0	100.0	9.2	100.0	16.4	100.0	21.6	100.0
Protected, ALD	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0

Note: the protected ALD arm is L1–L4 and the bare ODE arm is unprotected PF-ODE; the start-OOD floor is 2.0, 6.8, 7.6 and 9.2% at the four noise levels and is excluded from the induced rates.

The recheck reproduces the mechanism and strengthens the safety result. eSEN is the more fragile of the two potentials on the bare score—its induced rate reaches 50.8% at $\sigma = 1.0$ where MACE-MP-0 measures 33.0%, and its unprotected terminal validity falls to 48.0% against 99.0%, so the failure is not a MACE artifact but is if anything understated by the main scan. The analytic wall alone removes the induced rate at $\sigma = 0.5$ and cuts it by 2.3–5.5$\times$ thereafter; the constraint takes it to zero at every noise level measured, with 100% validity, against MACE-MP-0's 1.00% at $\sigma = 5.0$. Two differences are worth recording rather than smoothing over. First, the boundary's location moves: eSEN collapses at lower noise than MACE-MP-0, consistent with the composition-dependence the main scan measures and further evidence that the collapse scale is a property of the potential's learned extrapolation. Second, the energy quality is no longer free at high noise on this potential—the protected E_{hull} median sits 10–21% *above* the unprotected one at $\sigma \geq 2.0$ (531.6 versus 439.8 meV/atom at $\sigma = 2.0$; 1042.7 versus 947.2 at $\sigma = 5.0$), the reverse of the MACE-MP-0 ordering on the stochastic arm reported in §5.2 (MACE's own deterministic arm shows the same direction as eSEN here, so the reversal is a property of the sampler, not of the potential). The comparison is confounded in the direction that matters for a safety claim: at those noise levels the unprotected median is taken over the 48–56% of trajectories that remained valid, so the two arms are not medians over the same population, and the protected figure is the honest one.

E Secondary Sampling Paradigms

Three further samplers are documented for completeness. None is used in the main experiments, and their experimental evaluation would follow the protocol of Appendix G.

Underdamped Langevin dynamics introduces a momentum variable,

$$dx = v\,dt, \qquad dv = -\gamma v\,dt - \beta \nabla U(x)\,dt + \sqrt{2\gamma}\,dW_t, \tag{6}$$

which helps trajectories cross shallow barriers but can also carry atoms into the short-range regime between score evaluations, since momentum is not damped by the potential's restoring force.

Cyclical or reheating schedules set a non-monotonic noise level, $\sigma(t) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})|\sin(2\pi t/T_{\text{cycle}})|$. Each re-entry into the high-noise phase re-introduces the score error the annealing was designed to leave behind, and the errors accumulate across cycles.

Uncertainty-adaptive scheduling adjusts the noise level from a real-time uncertainty estimate,

$$\frac{d\sigma}{dt} = -\eta\,\mathbf{1}[u(x(t)) < u_{\text{threshold}}], \tag{7}$$

which pauses the decay precisely where the potential is unreliable. The in-principle OOD avoidance is strong, but the uncertainty estimate requires an ensemble and therefore a threefold cost in force evaluations, which is why the paradigm is not adopted: the Layer-1-protected PF-ODE already reaches $\leq 2.2\%$ induced OOD across the full scan without any uncertainty computation (Appendix C.3).

F Uncertainty and Ensemble Disagreement

The uncertainty estimate used by the adaptive scheduler and by the analysis below is the per-atom force standard deviation across the three potentials of Table 2, which requires three evaluations per step. The designed analyses are: (H1d) whether ensemble uncertainty spikes at the same noise level where the sampler-induced OOD rate rises, computed as the Pearson correlation between $\max_i u_i(x_t)$ and the per-trajectory OOD indicator over the bare-score noise scan on a reduced subset; and whether the uncertainty-adaptive schedule reduces the fraction of high-uncertainty visits relative to a fixed schedule at matched evaluations. Neither analysis has been executed. Independently of its outcome, the ensemble overhead is decisive for the practical recommendation, since the protected deterministic sampler achieves comparable OOD suppression at one third of the cost per step.

G Designed Evaluation Protocols

This section specifies the evaluation programme that measures what the main text does not: generation quality against trained baselines, the commensurability of the potential's score with learned scores, and transfer of the conclusions from the potential's surface to DFT. The generation-quality benchmark has since been executed on its sampler side—its six configurations on all three datasets are reported in §5.3 and in the subsection below—while its trained-baseline half, and every other protocol here, remains unexecuted: no result for the score-field comparison, curl violation, score mixing, efficiency, cross-potential, or DFT tiers is claimed anywhere in this paper. They are recorded here because the measurement design is itself a contribution—the reframing of Contribution 2 depends on the fact that these quantities are measurable—and because the protocols, including their thresholds and failure criteria, were fixed before execution. Table 11 is the reserved list.

Table 11: Protocols whose results remain reserved: each row's protocol is specified in this appendix and would report the metrics of Appendix A on MP-20, Carbon-24, and Perov-5. The generation-quality row of the earlier version of this table is no longer here—its sampler arms were executed and are reported in §5.3 and Appendix G.2; its trained-baseline half remains unrun. No number for any row below is claimed in the present version.

Measurement	**Quantity to be reported**
Score mixing	quality versus λ; bare and augmented $\lambda = 0$ overlay
Curl violation	$\mathbb{E}\|\nabla \times s\|_F$ for potential-based versus learned scores
Efficiency	evaluations to $E_{\text{hull}} < 100$ meV/atom versus MD and basin hopping
Cross-potential	per-potential validity, S.U.N., E_{hull}; rank correlation
DFT transfer	rank correlation, relaxation displacement, stability retention

G.1 Generation-Quality Benchmark

Two samplers are run over the six configurations C1, C3, C5, C8, C9, C11 of Appendix B on MP-20, Carbon-24, and Perov-5, at 10 candidates per test composition per seed, pooled over five seeds to 50 per composition. (The symbol K is reserved throughout for the number of ALD noise levels, $K = 100$; the candidate count is unrelated to it.) The pairing is not a full cross: the deterministic PF-ODE (recommended default, Proposition 1) is run on the three configurations that carry the exploration–safety axis, C5, C8 and C11, while stochastic ALD is run on all six, giving nine configuration–sampler cells per dataset rather than twelve. The three high-noise configurations that ALD already resolves (C1, C3, C9) are not duplicated on the ODE path, whose own failure mode is the evaluation cap rather than the collision. Structures are evaluated with the full metric suite

of Appendix A, including the flat-cell, C5-style and protected variants. The baseline half of this protocol—evaluating DiffCSP, DiffCSP++, CDVAE, MatterGen, and a random-plus-relaxation reference on the identical test compositions and metric suite—has not been run in the present study. Its one scored row, in §5.3, is DiffCSP's unconditional arm rather than the conditional `diffcsp_csp` arm this protocol designates, and so does not meet the identical-composition condition; it is reported there with that difference, and the potential mismatch, stated. Because learned models do not share the potential's energy surface, this comparison is made on generation-quality metrics only—validity, match rate, coverage, S.U.N., AMSD, and R-angle KL—and never on evaluations-to-E_{hull}, which is reserved for same-surface comparisons.

A measured constraint on this benchmark is the exploration budget: at $\sigma_{\max} \leq 1$ Å both samplers move less than one bond length end to end over their full budget, so the trajectory refines the basin of its noised starting structure rather than sampling across basins. Coverage below a learned baseline's level must therefore be reported against this measured floor, not against the nominal coverage of the noised prior, and diversity is carried by the candidate pool—which makes the candidate count and the prior design the primary coverage knobs. The high-noise configurations C3 and C9 exist precisely to map the coverage–safety trade-off that this creates; the executed grid's coverage numbers, read against this floor, are in the subsection below—on MP-20 they are 67–76% across the seven cells at $\sigma_{\max} = 0.5$ Å against 10.0% and 22.0% for the two at 1.5 Å, and the low coverage on the other two datasets is a property of their reference sets rather than of the sampler.

G.2 Generation-Quality Benchmark: Executed Sampler Arms

Table 13 reports the frozen grid's other two datasets, and Fig. 4 draws them. Table 12 and Fig. 3 belong to the MP-20 grid of §5.3 and are carried here rather than there, as is that grid's S.U.N. decomposition below: each carries a qualification, or a trained-baseline caveat, that would displace the MP-20 reading on the page, and none of them is a qualification of the sampler. On the two columns of Table 12 that are the same statistic—terminal validity and stability among samples with a resolved hull—the rows are close (100.0 versus 99.7, and 82.2 versus 79.7%), so a training-free sampler clears the same validity bar as a trained one on this dataset and structural quality is not its deficit; the conditional arm the protocol designates as the proper counterpart, `diffcsp_csp`, has not been scored.

The two rows occupy different stages of the pipeline of §2, and the comparison is bounded by that. The DiffCSP row is a *search*-stage model: it draws coordinates and the cell from noise under a composition and atom-count condition, its competence is bounded by the relaxed-structure database it was trained on, a new chemistry requires retraining, and its score field is under no constraint to be conservative. Ours is the *relaxation* stage, with a frozen potential as the score: no generative training, no database bound, conservative by construction, applicable to any composition, and—because the force field is the same object being evaluated—scored on the same surface the candidates are ranked on. What the two shared columns therefore test is whether a training-free relaxation clears the same generation-quality bar as a trained generator on this dataset, and the answer is that it does. What they cannot test is the complementary question—whether the relaxation stage, given only a composition, reaches the reference structure on its own—because that requires the search stage to be replaced rather than compared, and the arm that does so is the composition-only prior of Appendix C (A5), which has not been run. The stage distinction is also why the efficiency axis of Table 11 is offered only against same-surface baselines: a search-stage model's evaluations are not commensurable with a relaxation stage's, so quoting them side by side would inflate the comparison in our favour.

Two MP-20 readings are recorded here rather than in the main text. First, the S.U.N. reversal decomposes into its two components, which are the same two on both samplers. Protection raises thermodynamic stability on either sampler, by 35.5 pp on ALD (36.6% → 72.1%) and by only 8.3 pp on PF-ODE (74.0% → 82.3%), because the bare PF-ODE sampler is already the more stable of the two; and it lowers within-cell uniqueness on either sampler by almost the same amount—51.9% → 38.1% on ALD and 41.8% → 31.1% on PF-ODE—since the wall confines trajectories to the basin of their starting structure. On ALD the stability gain dominates; on PF-ODE the bare arm's stability leaves little to gain and the uniqueness loss dominates. Second, the high-noise arms are destructive on both sides: at $\sigma_{\max} = 1.5$ Å the bare arm holds only 3.0% stability at a median E_{hull} of 598 meV/atom and the protected arm 4.7% at 613, so the wall restores terminal validity to 100.0% and lifts coverage from 10.0% to 22.0% without recovering the energy distribution. This

Table 12: Our MP-20 row against a third-party trained generative model scored on the same metric suite, with the columns that make the two rows not directly comparable shown rather than hidden. Both rows are MP-20. The comparison is directional: read the note.

Arm	Potential	Valid	S.U.N.	Match	Cover.	E_{hull}	Stable (all)	Stable (known hull)
Ours (C11, PF-ODE)	eSEN	100.0	17.2	80.2	76.2	29	82.2	82.2
DiffCSP (uncond.)	MACE	99.7	5.1	2.2	0.2	17	6.3	79.7

The third-party row's two stability columns deliberately differ. *All* counts every sample, *known hull* counts only those whose hull entry the lookup resolved: for the third-party row that is 7.9% of $n = 1000$ samples, so its all-candidates rate is set by the lookup rather than by the generator, while ours has a resolved hull for every MP-20 candidate and the two columns coincide. Its E_{hull} is likewise a pooled median over those hull-known samples, ours a median of per-composition medians. Three further differences are not correctable in the table: the scored third-party arm is *unconditional*, so it does not satisfy the protocol's identical-test-composition condition; it is evaluated with MACE where our row uses eSEN; and its diversity metrics (AMSD, R-angle KL) are pooled while ours are composition-matched, so both are omitted rather than shown as if comparable.

Table 13: The generation-quality benchmark on Perov-5 and Carbon-24, same protocol and same nine cells as Table 1. Both datasets carry a qualification that is stated at the foot of the table and in the surrounding text; neither is a sampler result, and neither should be read as one.

Cell	$\sigma_{\max}$	Layers	Valid	Stable	S.U.N.	Match	Cover.	AMSD	R-KL	E_{hull}	Cap
Perov-5											
C1	0.5	—	68.5	17.5	7.2	32.6	50.0	0.5	10.45	407	—
C3	1.5	—	65.0	4.3	3.4	3.7	30.0	0.8	12.36	423	—
C5†	0.5	—	68.9	17.7	7.6	32.1	50.0	0.5	10.48	407	—
C8	0.5	L1–L3	100.0	31.0	11.0	37.7	55.0	0.4	9.71	55	—
C9	1.5	L1–L3	100.0	7.9	5.4	3.0	40.0	0.7	12.36	481	—
C11	0.5	L1	100.0	30.1	10.5	37.2	60.0	0.4	9.72	56	—
C5	0.5	—	91.6	26.4	9.3	38.5	60.0	0.3	9.72	99	20.5
C8	0.5	L1–L3	100.0	26.7	8.2	38.3	65.0	0.4	9.51	246	43.0
C11	0.5	L1	100.0	26.7	7.8	38.3	65.0	0.4	9.53	263	43.1
Carbon-24											
C1	0.5	—	56.3	0.3	0.0	12.1	0.6	1.7	0.12	1126	—
C3	1.5	—	50.6	0.0	0.0	3.5	0.7	1.8	0.15	1424	—
C5†	0.5	—	54.9	0.2	0.0	12.0	0.6	1.7	0.12	1134	—
C8	0.5	L1–L3	100.0	1.8	0.0	19.7	0.7	1.5	0.07	1006	—
C9	1.5	L1–L3	100.0	0.1	0.0	2.9	0.5	1.6	0.09	1464	—
C11	0.5	L1	100.0	1.6	0.0	19.8	0.8	1.5	0.07	996	—
C5	0.5	—	79.1	1.9	0.0	22.8	0.7	1.5	0.08	941	88.6
C8	0.5	L1–L3	100.0	2.1	0.0	22.5	0.8	1.5	0.08	953	89.3
C11	0.5	L1	100.0	2.1	0.0	22.5	0.8	1.5	0.08	952	88.8

Columns as in Table 1. *Perov-5*: E_{hull} is a median over 12–16 of the 20 cells, because 44–65% of its candidates have no physical hull value at all (all below the -0.05 eV/atom floor), so that column is a median over a selected subsample as well as a median of per-cell medians. The dropout is not arm-specific: it appears under all nine. *Carbon-24*: **S.U.N. is 0 in every row**; in this run every stable candidate is non-novel and every novel candidate is non-stable—but the two criteria are *not* exclusive in principle (a new allotrope inside the 100 meV/atom window satisfies both), so the empty intersection is an outcome of the run and the collapsed rate is reported as a result rather than dismissed as a degeneracy—and its three PF-ODE rows are budget-truncated rather than converged (Cap 88.6–89.3%, every cell's median at the 1,000-step ceiling), so those three rows measure a partial trajectory. Both qualifications are stated in full in the surrounding text.

is the generation-side counterpart of the noise scan's finding that the wall flattens the failure curve without moving its midpoint, and it is why the recommended operating point keeps $\sigma_{\max} \leq 1$ Å.

Carbon-24's S.U.N. rate is zero in all nine cells, and the two criteria are disjoint in this run rather than by construction. The dataset has a single composition, and its hull table holds a single entry—

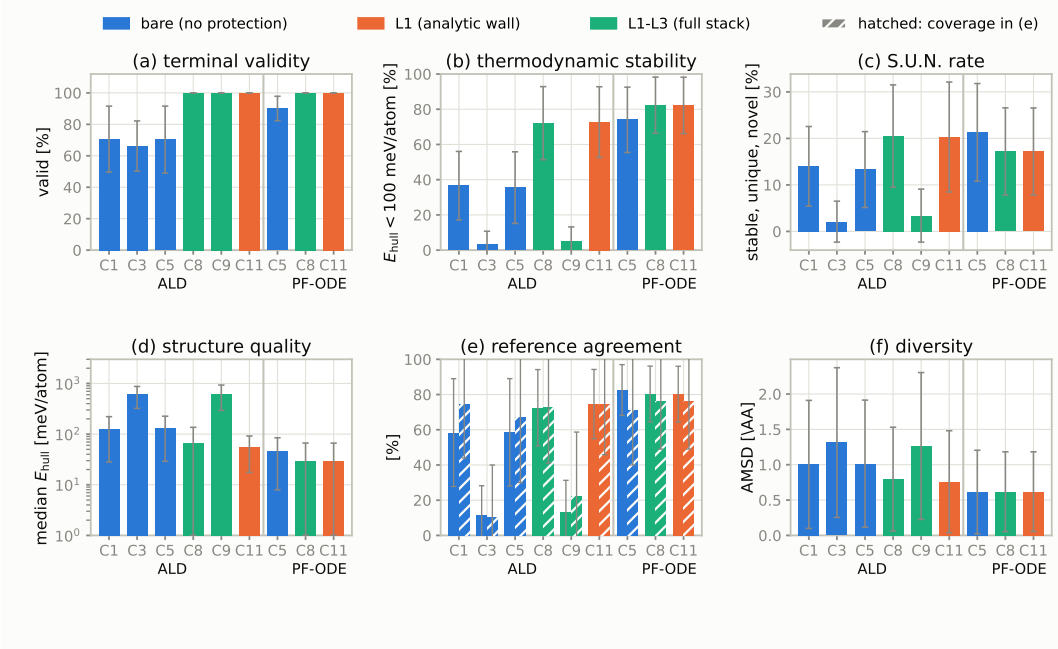


Figure 3: The generation-quality benchmark on MP-20, 1,000 trajectories per cell under eSEN. Bars are the mean over the 20 test compositions and error bars are the spread across those compositions, not across seeds. Hue encodes the defense state—bare, Layer 1 only, and the L1–L3 stack—so the same three colours mean the same three arms as in Fig. 2; where a panel carries two metrics they are separated by hatching rather than by a second hue. The nine cells are grouped by sampler, ALD (C1, C3, C5, C8, C9, C11) left of the rule and PF-ODE (C5, C8, C11) right of it, and are never pooled. (a) Terminal validity, where the five protected arms sit at exactly 100%. (b) Thermodynamic stability, the two-sided band $-50 \leq E_{\text{hull}} < 100$ meV/atom. (c) S.U.N. rate: protection raises it on ALD and lowers it on PF-ODE, and the components that produce the reversal are in Table 1. (d) Median physical E_{hull} on a log axis, which is a median over cells and so a median of medians. (e) Match rate and, hatched, coverage. (f) AMSD, the mean nearest-neighbour RMSD within a composition's ten candidates, where higher is more diverse: the deterministic sampler returns the *less* diverse ensembles (0.6 Å against 0.8–1.3 Å on ALD), which is the ensemble-level counterpart of the basin confinement discussed in the text.

elemental carbon, at $e_{\text{form}} = 0.0$, with the energy zero defined as the minimum over the dataset's own train, validation, and test splits. Stable and Novel are *not* mutually exclusive in principle: a genuinely new allotrope landing inside the 100 meV/atom window would satisfy both, and nothing in the criteria forbids it. What this run produced is disjointness: over the nine cells, all 101 stable candidates are non-novel and all 7,678 novel candidates are non-stable, giving $P(\text{Novel} \mid \text{Stable}) = 0.000$ against 0.88 on MP-20 and 1.00 on Perov-5. Fig. 5(a) shows the four-way breakdown; the collapse is in the intersection, and every marginal is non-degenerate—1.6–2.1% of candidates are stable under the four protected arms at the operating point, and 75–98% are novel—so it is the intersection that is empty, not the stability marginal. Read as a result, this says that on Carbon-24 the samplers reached the stability window only where they reproduced an allotrope the dataset already held, which is what the basin confinement of Appendix G predicts for a sampler that refines its starting structure when the hull's single entry is the dataset's own ground state. Carbon-24's S.U.N. rate should therefore not be compared with the other datasets'—a single-entry hull makes it a different quantity—but it is not excluded as uninformative either, and it is not a cross-dataset S.U.N. trend that this paper reports.

Carbon-24's three PF-ODE arms are budget-truncated, not converged. 88.6–89.3% of their trajectories exhaust the 1,000-evaluation cap, and every one of the sixty cells has a median NFE of exactly 1,000—the whole median trajectory sits at the ceiling, while the mean falls short of it only because a minority of runs terminate early (Fig. 5(b)). Every Carbon-24 PF-ODE number in Table 13 is

therefore a measurement of a partial trajectory and should be quoted with its cap rate beside it. 0.89 also has a physical reading that the grid was not designed to separate: the σ_c model of §5.2 extrapolates Carbon-24's collapse scale to 0.31–0.39 Å, below the $\sigma_{\max} = 0.5$ Å this grid runs at, so the trajectories start already inside the wall. The start-OOD rate is the direct witness: it is 19.2% on Carbon-24 under all seven of its $\sigma_{\max} = 0.5$ arms, against 2.0% on MP-20, and protection does not lower it—the protected Carbon-24 arms sit on the same 19.2% floor that the bare ones do, and can only bring the total OOD rate down onto it, from 93.9–94.0% under the bare ALD arms to 19.2–22.1% under the protected ones. The trajectories are thus handed a starting structure already past the wall, and the sampler then spends its entire budget failing to escape. This is the one place in the study where the saturation regime of Fig. 2(d) is reached at the recommended operating point, and it is a property of a one-composition dataset's length scale rather than of the sampler.

Perov-5's E_{hull} column is a median over a subsample, and its dropout is not arm-specific. 44–65% of its candidates have no physical hull value at all, every one of them below the -0.05 eV/atom floor, so between twelve and sixteen of the twenty cells contribute to the column in Table 13 under every one of the nine arms. Fifteen of the twenty cells are affected, thirteen of them under all nine arms, and the affected set is the same list of compositions throughout—CsIrO2F, NaBO3 and RbBeNOF lose essentially every candidate under every arm, while the milder cases (ScAlNO2, LaCrSO2) lose a minority. This is an eSEN extrapolation limit on those chemistries, not a sampling effect: the dropout follows the composition and not the arm. The column is reported, with its cell count, rather than dropped; the MP-20 and Carbon-24 columns are unaffected and rest on all twenty cells. Fig. 5(c) plots the share by arm, where its flatness across bare and protected arms is the evidence that no defense layer causes it.

Two Perov-5 readings do transfer. Protection again takes terminal validity to exactly 100.0% on all five protected arms, against 65.0–91.6% for the bare ones, and its stability ordering over the layer sets is the same as MP-20's. And its coverage is systematically lower than MP-20's—50–65% against 67–76% at the operating point—for a reason that is also a property of the reference set: Perov-5 holds essentially one structure per composition, so a cell's ten candidates have little to recover.

G.3 Score-Field Comparison, Curl Violation, and Score Mixing

Score-field comparison. For 100 structures sampled along natural diffusion paths, $x_t = \alpha_t x_0 + \sigma_t \epsilon$, we would compute the potential-based score and a learned score at each (x_t, t) and quantify the angular deviation

$$\phi_{\text{ang}}(x,t) = \arccos\left(\frac{s_{\text{NNP}}(x) \cdot s_{\text{learned}}(x,t)}{\|s_{\text{NNP}}(x)\| \, \|s_{\text{learned}}(x,t)\|} \right), \tag{8}$$

the relative magnitude $\|s_{\text{learned}}\|/\|s_{\text{NNP}}\|$, the per-atom cosine coherence, and the squared field difference. Both scores should converge to physics near the data manifold and diverge as the noise level grows; the deviation profile locates where a learned score extrapolates beyond physics.

Curl violation. The curl of a score field is estimated with the Hutchinson trace estimator (Hutchinson, 1990),

$$\|\nabla \times s\|_F^2 \approx \frac{1}{R} \sum_{r=1}^{R} \frac{\|(J - J^\top)\boldsymbol{v}_r\|^2}{\|\boldsymbol{v}_r\|^2}, \tag{9}$$

where $J_{ij} = \partial s_i / \partial x_j$ is approximated by central finite differences at $\delta = 10^{-4}$ Å and $\boldsymbol{v}_r \sim \mathcal{N}(0, \boldsymbol{I})$ for $R = 100$ directions per structure. The protocol is stated in the Cartesian chart, where conservativity is defined (Theorem 2) and where δ is a Cartesian length; evaluating the same measurement on a fractional field would not be a test of conservativity, since the curl operator is chart-dependent and a non-orthonormal L makes an exactly conservative field appear rotational. For a potential-derived score this quantity is numerically zero by construction (Theorem 2); for a learned score it is not, and the comparison quantifies how much non-conservative structure a learned field carries. Non-conservative components drive circulation that no energy function can account for, so this measurement bears directly on whether a potential-derived score should serve as a regularization reference.

Score mixing. Commensurability is a precondition for any mixing, and requires three alignment steps: the learned model diffuses lattice and fractional coordinates jointly, so its fractional com-

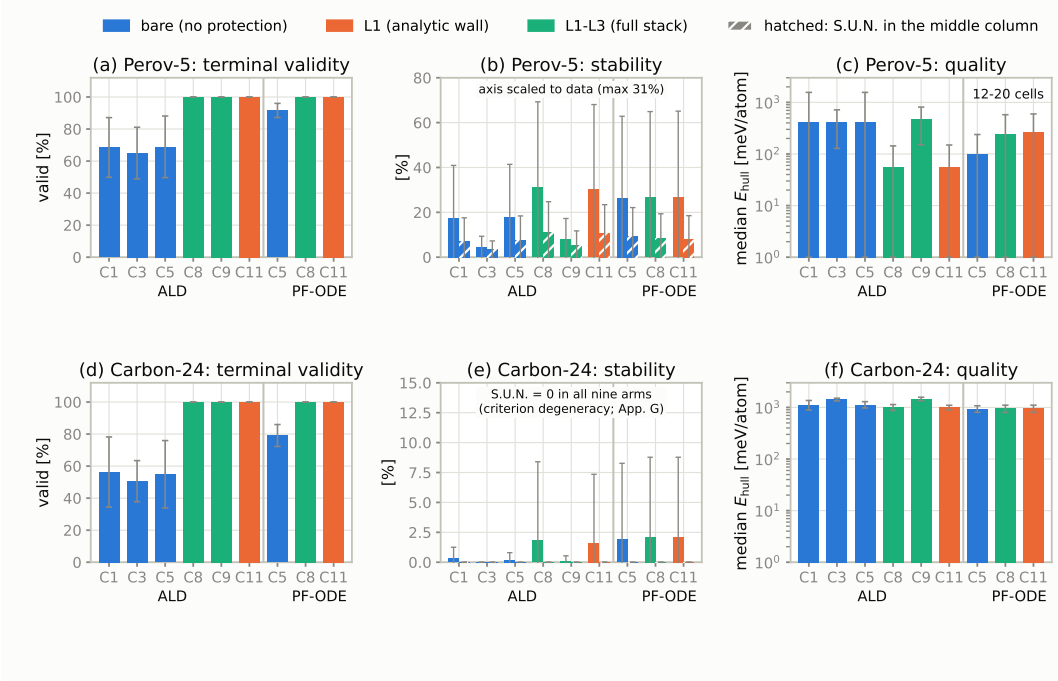


Figure 4: The generation-quality benchmark on the two secondary datasets, same protocol and same cells as Fig. 3. Row 1 Perov-5, row 2 Carbon-24. The middle column's vertical scale is fitted to its data rather than shared with Fig. 3: Perov-5's stability peaks at 31% and Carbon-24's at 2.1%, so a common 0–100 axis would draw both as a flat line at zero. Carbon-24's S.U.N. bars are absent rather than missing—the rate is exactly zero in all nine cells, because in this run every stable candidate is non-novel and every novel candidate is non-stable. The two criteria do not require that in principle, so the empty intersection is an outcome of the run rather than a degeneracy of the metric; the breakdown is in Fig. 5(a) and the full statement in §G.2.

ponent must be extracted and compared against $s^{\text{frac}}_{\text{NNP}}$ (Eq. (1)), with its lattice component used only in the pure-learned reference arm; its discrete timestep must be mapped to an effective noise level through the model's own forward kernel, so that mixing at a given physical noise scale compares scores evaluated at matched σ; and both fields must be rescaled by the median relative magnitude measured on the calibration set, so that the mixing weight interpolates *direction* rather than magnitude. With calibrated fields, the hybrid score is $s_\lambda = (1-\lambda)\hat{s}_{\text{NNP}} + \lambda\hat{s}_{\text{learned}}$ for $\lambda \in \{0, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9, 1.0\}$, sampled with PF-ODE and ALD on 20 MP-20 compositions with Layer 1 active in every NNP-containing arm, plus two overlay points at $\lambda = 0$: the unprotected score and the fully protected stack. The quantity of interest is the shape of the quality-versus-λ curve—whether it shows diminishing returns, a threshold, or an optimum at intermediate weight—and specifically how much of the gap between the unprotected score and a learned score is closed by the analytic wall alone. If the curve is flat over a substantial range of small λ, the short-range wall accounts for a corresponding share of what the learned model acquired from data, which is the empirical form of the claim that a universal component should not be learned.

G.4 Efficiency against Molecular Dynamics on a Common Surface

On the same potential surface we compare the score-based samplers at $\{100, 200, 500, 1000, 2000\}$ evaluations against MD simulated annealing at cooling rates spanning 10^{12}–10^{14} K/s and basin hopping at $\{50, 100, 200, 500\}$ hops, using the protocols of Appendix A (the classical baselines span 10^4–10^6 evaluations). The reported quantities are the fraction of runs reaching $E_{\text{hull}} < 100$ meV/atom and the minimum number of evaluations required to reach it, with AMSD reported alongside so that diversity is not sacrificed silently at low budget.

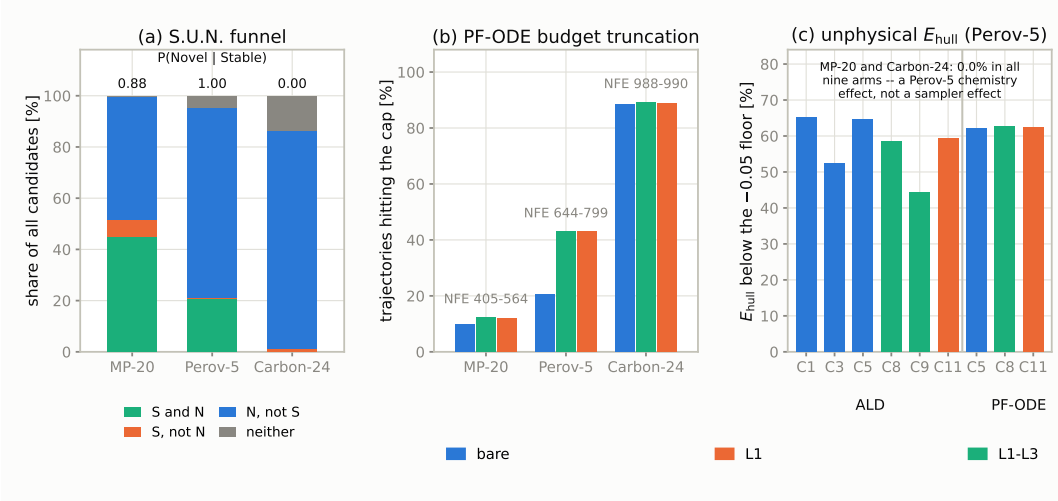


Figure 5: The two qualifications carried by the secondary datasets. (a) The stable/novel breakdown over all candidates, per dataset. On Carbon-24 the stable-and-novel segment is empty in this run—every stable candidate is non-novel—which the two criteria do not require in principle, so the emptiness is reported as an outcome of the run rather than as a degeneracy of the metric; the label over each bar is $P(\text{Novel} \mid \text{Stable})$, which is 0.00 there against 0.88 and 1.00 elsewhere. (b) The share of PF-ODE trajectories exhausting the 1,000-evaluation cap, with the range of mean NFE over the three PF-ODE arms annotated per dataset: Carbon-24's arms sit at NFE 988–990 of a possible $1{,}000$, so its PF-ODE rows in Table 13 measure truncated trajectories. (c) The share of E_{hull} values below the -0.05 eV/atom floor, which is a Perov-5-only effect (0.0% on both other datasets under all nine arms) and is flat across the defense states, so it is a property of the potential on those chemistries rather than of any sampler or protection choice.

The mechanism the comparison isolates is worth stating precisely, because it is not gradient descent—the Langevin drift of the classical baseline is itself the negative gradient. Simulated annealing explores through thermal noise, so basin escape is thermally activated and exponentially slow at low temperature, and most of its evaluations are spent on diffusive wandering. Denoising instead anneals a schedule-controlled noise level and follows the score field, so essentially every evaluation contributes directional progress. The advantage is therefore a property of the exploration mechanism and is independent of the completeness boundary; it is also bounded by the exploration-budget measurement above, which is what makes the comparison informative rather than merely favourable.

G.5 Cross-Potential Validation

The streamlined benchmark subset—C5 and C8 with PF-ODE—runs on MP-20 with both potentials of Table 2, 20 compositions and 100 candidates per potential, reporting per-potential validity, S.U.N. rate, and E_{hull}. For every generated structure the energy is then computed with both potentials, giving the Spearman rank correlation of per-structure energies and the agreement of stability classifications. The scientific question is whether the completeness boundary and the effect of the analytic correction are properties of the class of potentials or of one checkpoint—the decomposition principle predicts the former, since short-range universality is a property of the target physics rather than of any model.

G.6 DFT Validation

Conclusions drawn on a potential's surface are only as good as the transfer to the electronic-structure level, and generated structures may by design visit configurations where the potential's accuracy is not guaranteed. A four-tier protocol addresses this.

1. **Single-point energies.** 200 generated structures stratified by potential-level E_{hull} (50 each from bins < 0, $[0, 50)$, $[50, 100)$, and ≥ 100 meV/atom) across all three datasets, evaluated with

the functional matching the potential's training data—PBE+U with Materials Project U values for the Materials-Project-trained potentials, and r^{2}SCAN for the OMat24-trained one. Metrics: Spearman rank correlation between potential and DFT energies; mean absolute error stratified by E_{hull} bin; and stability-classification agreement at the 100 meV/atom threshold. The hypothesis is strong correlation ($\rho > 0.8$) within the potential's reliable domain, degrading for higher-energy structures.

2. **Relaxation.** The 50 lowest-E_{hull} structures from Tier 1 that pass basic stability checks (no imaginary Γ-point modes, all distances at least 0.75 times the covalent sum) are relaxed with DFT in both positions and cell. Metrics: RMS displacement between potential-relaxed and DFT-relaxed positions, volume ratio, E_{hull} after DFT relaxation, and the fraction remaining stable. The hypothesis is that more than half of the structures predicted stable at the potential level remain stable after DFT relaxation.
3. **Head-to-head against MD.** For five representative compositions, the production configuration and MD simulated annealing are run on the same potential surface; the five lowest-energy structures from each are relaxed with DFT. Metrics: the minimum DFT-level E_{hull} each method achieves and the evaluation cost of achieving it. The hypothesis is that score-based sampling reaches comparable DFT-level minima at far lower cost, which is what would establish that the efficiency measured on the surrogate is algorithmic rather than an artifact of the surrogate.
4. **Dynamical stability.** For the ten lowest-E_{hull} DFT-relaxed structures, Γ-point phonon frequencies via DFPT or frozen phonons, to test whether a subset of thermodynamically stable candidates is also dynamically stable—the stronger indicator of synthesizability.

Settings. Calculations use VASP 6.4 (Kresse & Furthmüller, 1996; Kresse & Joubert, 1999) with projector-augmented-wave pseudopotentials (Kresse & Joubert, 1999) and the functional consistent with each potential's training data (PBE+U with Materials Project U values for MACE-MP-0; r^{2}SCAN for eSEN). k-point grids are generated by `pymatgen`'s reciprocal density of 1000 (Å^{-3} of reciprocal cell volume, not per reciprocal atom), yielding e.g. $15 \times 15 \times 3$ for 8-atom primitive cells and $5 \times 5 \times 3$ for the ≥ 10 Å min-image supercells; plane-wave cutoffs are 520 eV (PBE) and 700 eV (r^{2}SCAN), following the Materials Project and OMat24 protocols. The total cost is approximately 2250 DFT calculations, of which the first two tiers already provide the central transferability evidence.

H Per-Composition Breakdown

The pooled curves of Table 15 are unions over the 20 MP-20 test compositions of per-composition cells, each pooling five seeds and 10 candidates into 50 trajectories per $(\sigma_{\max}, \text{condition})$ pair. The full per-composition breakdown—20 tables reporting the sampler-induced OOD rate for every composition, noise level, and condition—is provided as supplementary material rather than reproduced here, because those cells are the ones most affected by the initial-condition defect of §6.1; like the pooled tables, they are taken from the corrected re-run. The composition-level variation they carry is summarized in Table 16, which also reports the a-priori probe prediction: the potential's repulsive pair force at 0.75 times the nearest-neighbour distance, which costs about three minutes per composition and no sampling, ranks the compositions it resolves correctly and predicts σ_c within a factor of 1.5–2.2.

Failure timing, and the initial-condition test. The failures the scan measures are early-step events on any initial condition, not a steady-state tail. Failures concentrate in the first 3–12 steps irrespective of the start, so the curve reflects large-noise drift steps; the alternative that the sampler concentrates a cluster which then collapses is also downgraded, since clean starts contribute only a quarter to a third of the collisions. A pre-registered marginal test showed that re-sampling initial conditions to enforce $r_{\min} \geq 0.4$–0.6 Å rejects 25–31% of starts but reduces the collision rate at $\sigma = 5$ by only about 13%, far below its $2\times$ gate, so the initial-condition generator was left unchanged.

The σ_c estimator, stated once. Table 16 reports the scale that solves $P_{\text{step}}(\sigma_c) = 0.005$, where P_{step} is the per-accepted-step probability that $r_{\min} < 0.5$ Å, pooled over the accepted steps at step ≥ 1 of clean-start trajectories in one (composition, $\sigma_{\max}$) cell; the solve is a log-linear interpolation of $\log P_{\text{step}}$ against $\log \sigma_{\max}$ over the ten-point grid, applied to the running maximum of P_{step} so that the root found is the first crossing. Both constants in that sentence are chosen rather than derived—

the threshold, calibrated as one collision per 200-step trajectory, and the value substituted for a cell whose rate is exactly zero. We therefore also report a threshold-free statistic, the per-step collision rate at the single condition $\sigma_{\max} = 5.0$ Å, and both raw and pair-normalized versions of the scale. Three properties of the estimator bound how far its output can be read as a material property; all three are computed by `scripts/analyze_sigma_c_robust.py`, which reproduces the tabulated values exactly.

(i) Size. A cell with more atoms has more pairs and hence more chances to be within 0.5 Å at a given step. The MP-20 cells span 2–16 atoms, i.e. 1–120 pairs, wider than the spread the crossing is used to establish, and the crossing correlates with pair count across the 15 compositions it resolves ($\rho = -0.574$ on logs, $p = 0.025$). The five compositions that never reach the threshold are the five smallest cells (2–4 atoms). Normalizing does not remove the trend but reverses it: at $\sigma_{\max} = 5.0$ the raw rate correlates positively with pair count ($\rho = +0.620$, $p = 0.0035$) and the per-pair rate negatively ($\rho = -0.566$, $p = 0.0093$), because in a two-atom cell the single pair saturates—once it is inside the threshold, every subsequent step counts.

(ii) Residence. The published statistic counts every step spent below threshold, so it is the product of two factors: the per-step hazard of *entering* the OOD state and the mean number of steps then spent in it. Over the 20 compositions the decomposition is exact, and the two factors are of comparable importance: at $\sigma_{\max} = 5.0$ the published rate spans $88.9\times$, the entry hazard $30.7\times$ and the residence time $3.3\times$. Both factors correlate with cell size ($\rho = +0.63$ and $+0.57$). A cross-composition rate comparison therefore mixes propensity, cell size, and how long a cell stays stuck once it is stuck, which is why we do not read the spread as a ranking of score quality.

(iii) Definition. The crossing is replaced by a fitted scale: the midpoint of a saturating sigmoid $A/[1 + e^{-(\log\sigma - \log\sigma_c)/\tau}]$ fitted in $\log\sigma$. This estimator is invariant to any per-composition multiplicative chance factor, so it is size-free by construction, and it reports an amplitude A alongside σ_c. It resolves for only 6 of the 20 compositions: the other 14 are still climbing at the top of the scan, by up to $7.8\times$ between $\sigma_{\max} = 3$ and 5, so no midpoint exists inside the scanned range and none is extrapolated. Over the 6 that resolve, σ_c spans $2.7\times$ (1.26–3.42 Å), not $17.5\times$. §5.2 quotes the same family of fit for the *pooled* bare-ALD curve, where it is evaluated in σ on the log-spaced grid rather than in $\log\sigma$ and gives $\sigma_c = 0.334$ Å with $A = 32.1\%$; reading that one curve with the $\log\sigma$ midpoint above instead gives 0.353 Å and $A = 34\%$, a 6% convention dependence of the same kind, and smaller than the between-composition spread, that the per-composition table carries. The values 0.462 Å and 0.618 Å quoted in earlier versions of this work were fitted to the *total* rate, which includes the start-OOD floor and so saturates near 0.74 rather than 0.32, and are not the scale of the induced curve those panels draw.

(iv) Denominator. The induced rate answers a population question—of all trajectories launched, what fraction failed because of the sampler?—and its two conventions differ in the pool they divide by. Conditioning instead on the trajectories that started clean answers the per-trajectory question, and the two separate as the start-OOD floor grows: pooled over the ten noise levels the bare-ALD rate is 23.2% published against 30.5% conditional, but at $\sigma_{\max} = 5$ Å, where 43.1% of starts are already OOD, the same 312 events read 31.2% against 54.8%. The published curve's plateau is therefore in part the shrinking risk pool: held at a fixed pool the conditional curve is still climbing at the top of the scan (50.9, 50.6, 53.0, 53.3, 54.8% at $\sigma = 1, 1.5, 2, 3, 5$ Å). Table 14 pairs the two fits, computed with one estimator on one set of trajectories differing only in the denominator, for every arm and sampler. Every σ_c moves right and every amplitude rises by a factor of 1.6–1.8 between columns, and the ordering within each arm and sampler is unchanged in all six rows: bare before L1 before L1–L3 on both paths, and, within the bare and L1 arms, PF-ODE before ALD (0.462 against 0.334, then 0.947 against 0.634). The one reversal is L1–L3, where ALD's 2.841 Å exceeds PF-ODE's 1.165 Å; the conditional column keeps the reversal (2.869 against 1.449) rather than removing it, so it is the arms and not the denominator convention that produce it. Both L1–L3 curves are flat by then—amplitudes of 1.0 and 1.9%—and a crossing statistic near its own floor ranks curves that it can barely resolve, which is why the last row is reported as a pair of saturated scales rather than as an ordering. All rates here and in Table 15 are reproducible from `scripts/analyze_conditional_ood.py`, which imports the pooled table's own loader so that the two columns cannot drift apart.

Table 14: The denominator convention, paired: same estimator and the same trajectories in each row, dividing the induced count by all trajectories (published) or by the clean-start pool (conditional). Every scale moves right and every amplitude roughly doubles; the arm ordering is unchanged in all six rows. Computed by `scripts/analyze_conditional_ood.py` from the cells of Table 15, whose loader it imports so the two columns cannot drift apart. The L1–L3 arms are the runner's `l1l2l3` cell on ALD and `l1l4` on PF-ODE, as in that table.

		published		conditional	
Arm	Sampler	σ_c (Å)	A (%)	σ_c (Å)	A (%)
Bare	ALD	0.334	32.1	0.438	51.9
Bare	PF-ODE	0.462	6.1	0.547	10.3
L1	ALD	0.634	21.4	0.738	36.6
L1	PF-ODE	0.947	1.8	1.296	3.3
L1–L3	ALD	2.841	1.0	2.869	1.8
L1–L3	PF-ODE	1.165	1.9	1.449	3.5

Table 15: Sampler-induced OOD rate (%) and median E_{hull} of valid final structures (meV/atom) versus $\sigma_{\max}$ for the bare score, Layer 1 only, and the three-layer configuration L1–L3, under ALD and PF-ODE, on MACE-MP-0. The induced rates are plotted in Fig. 2(a) and (d) and the ALD E_{hull} medians in panel (b); this table is their numerical record. Entries are clean-start trajectories with an event at steps ≥ 1, normalized by all trajectories; the start-OOD floor is excluded. E_{hull} medians are over valid final structures only, so at high σ the unprotected rows rest on the minority of trajectories that survived and are not comparable with the protected arms' across the full cell. Each cell is 20 compositions $\times$ 5 seeds $\times$ 10 candidates $=$ 1000 trajectories, from the initial-condition-corrected re-run (§6.1). In the ODE path Layers 2 and 4 have no counterpart (no noise injection, no acceptance step), so its protected arm is L1+L3 by construction; the divergence guard's effect at fixed σ is reported separately (Appendix A.7).

	$\sigma_{\max}$ (Å)									
Condition	**0.1**	**0.2**	**0.3**	**0.5**	**0.7**	**1.0**	**1.5**	**2.0**	**3.0**	**5.0**
Bare, ALD	0.00	0.20	7.40	31.90	34.00	33.00	31.80	31.70	31.10	31.20
Bare, PF-ODE	0.00	0.00	0.10	4.30	5.70	5.70	5.80	6.70	6.60	6.30
L1, ALD	0.00	0.10	0.90	6.70	13.50	18.40	19.50	19.60	22.80	24.50
L1, PF-ODE	0.00	0.00	0.00	1.30	1.40	0.60	1.50	0.80	2.20	1.70
L1–L3, ALD	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.10	0.60	1.00
L1–L3, PF-ODE	0.00	0.00	0.00	1.20	1.40	0.60	1.50	0.70	2.30	1.80
Median E_{hull} of valid final structures (meV/atom)										
Bare, ALD	1.2	3.7	8.6	57.6	163.5	229.2	323.2	418.9	598.9	796.3
Bare, PF-ODE	6.0	4.9	8.1	23.9	61.6	119.3	207.1	302.4	427.4	597.3
L1, ALD	1.3	3.8	8.2	33.1	119.0	206.3	314.4	392.6	555.0	786.3
L1, PF-ODE	6.1	4.9	8.2	26.5	77.3	157.9	291.4	383.2	529.6	749.4
L1–L3, ALD	1.3	3.8	8.3	33.5	121.9	196.9	309.6	391.6	568.8	753.1
L1–L3, PF-ODE	6.1	4.9	8.2	29.2	90.4	230.1	386.1	524.3	727.1	1144.8

I THEORETICAL STATEMENTS

This appendix collects the formal statements behind every result cited in the main text, preceded by a summary of what they say. The results are structural: they follow from the score being the gradient of a scalar potential, and they hold for the augmented score exactly as for the bare one.

Summary. Because s_{NNP} is the gradient of a scalar, its Jacobian is everywhere symmetric by Clairaut's theorem and $\nabla \times s_{\text{NNP}} \equiv 0$; the same holds for s^{aug} (Theorem 2), and no training choice can break this. The fixed-noise annealed chain therefore has the Boltzmann distribution of the augmented energy as its steady state at every noise level in the σ^2-prefactor convention of Eq. (4), and in the implemented convention—where that prefactor is cancelled—a Boltzmann law at the effective temperature $T_{\text{eff}}(\sigma_k) = \sigma_k^2/(k_B\beta)$, so that the noise schedule *is* the temperature schedule (Theo-

Table 16: Per-composition collapse scale σ_c for the bare score under ALD on MACE-MP-0 ($P_{\text{step}}(\sigma_c) = 0.005$, log-linear interpolation on the ten-point noise grid; ∼9,950 accepted steps per cell—step 0 excluded, trajectories that start OOD dropped), from the corrected re-run. *Pred.* is the a-priori pair-potential probe prediction, which requires no sampling. Entries in Å. Entries are ordered; the ordering is only resolvable at the extremes, see the notes.

Composition	σ_c	**Pred.**	**Composition**	σ_c	**Pred.**
Al_2O_3	0.28	—	$LiCoO_2$	1.73	—
$SrTiO_3$	0.50	0.31	TiO_2	2.05	—
CoF_3	0.54	—	$AlFe_3$	2.12	—
AlF_3	0.58	—	ZrO_2	3.11	—
AlCu	0.62	—	ZnS	3.18	1.46
Ca_3N_2	0.68	—	MoS_2	4.01	—
$BaTiO_3$	0.92	—	Ag_2S	4.94	—
$FeNi_3$	1.70	1.10	AlCo / CdS / GaN / LiF / VN	>5	4.30 / 3.90‡

Notes: The 0.28–>5 Å spread (17.5× over the 15 compositions the crossing resolves) refutes the geometric estimate 0.30–0.43 Å of Corollary 1 as the location of σ_c; so does the definition-free midpoint, which resolves on 6 of the 20 compositions and spans 2.7× (1.26–3.42 Å). The crossing statistic is size-confounded and floor-sensitive: it correlates with pair count ($\rho = -0.57$ on logs, $p = 0.025$) and the five compositions it does not resolve are the five smallest cells, and moving the zero-floor from 10^{-12} to 10^{-3} widens the spread to 21.9× and takes the smallest σ_c to 0.23 Å. Trajectory-level cluster bootstrap 95% CIs (400 replicates) are ±0.01–0.03 Å for the three smallest values but reach 1.9 Å at TiO_2 and 2.4 Å at ZnS, so only the extremes of the ordering above are resolved: Al_2O_3 (0.28, CI 0.28–0.28) and the five censored compositions are separated from the rest, while the middle of the table is not ordered at its quoted precision. ‡The probe places GaN at 4.30 and LiF at 3.90 Å, both below the 5 Å edge, and ranks the three compositions it resolves correctly ($SrTiO_3$ 1.6×, $FeNi_3$ 1.5×, ZnS 2.2×; all three *under*-predicted, so the probe is a lower bound on σ_c and a safe basis for choosing $\sigma_{\max}$). AlF_3, non-estimable in the pre-fix tabulation, is resolved at 0.57 Å once the initial-cell defect is removed.

rem 1). In the deterministic limit the flow is monotone energy descent, $dU/dt = \sigma\dot{\sigma}\beta\|\nabla U\|^2 \leq 0$ during generation. Learned score models carry no such guarantee: their Jacobians are unconstrained, so their fields need not be gradients (Plainer et al., 2025). The augmented score reproduces the drift of overdamped Langevin dynamics (Theorem 1), and decomposing its error against the true score gives

$$\|\epsilon(x,\sigma)\| \leq \beta\kappa\|x-\mu\| + \left\|\frac{\mu-x}{\sigma^2} + \beta\nabla U_{\text{NNP}}(\mu)\right\|, \tag{10}$$

with κ the smoothness constant of U_{NNP} on the region visited and μ the posterior mean (Theorem 3). The first term vanishes as $\sigma \to 0$. The second does not, and its zero-noise limit splits into two terms, $\|\nabla \log p_0(x_0) + \beta\nabla U_{\text{DFT}}(x_0)\| + \beta\|\nabla U_{\text{DFT}} - \nabla U_{\text{NNP}}\|(x_0)$: the mismatch between the data distribution and the Boltzmann law the score targets, plus the model's force error at the posterior mean. Only the second of these is a property of the potential's accuracy; the first is cleared by the additional hypothesis $p_0 \propto e^{-\beta U_{\text{DFT}}}$, which a relaxed-structure database does not satisfy at any temperature (Eq. 15, Remark 1). The reliability condition $\sigma \ll R_{\text{basin}}/\sqrt{d}$ (Corollary 1) therefore bounds the smoothness term alone: it is a conservative sufficient bound on that term, not a predictor of where the collapse occurs (§5.2; Appendix I). The Pauli wall and the constraint together enlarge the effective basin (Remark 3), which is what extends the reliable range.

For a score error ϵ, the terminal deviation of a PF-ODE trajectory scales as $\mathcal{O}(\int_0^T \|\epsilon\|dt)$ at leading order. The analogous term for the stochastic sampler is coupling-dependent rather than additive: under a synchronous coupling—both trajectories driven by the same Brownian path—the injected noise cancels exactly in the difference and leaves the same $\mathcal{O}(\int_0^T \|\epsilon\|dt)$ scaling, while under independent noise the trajectories separate by the injected noise even at $\epsilon \equiv 0$ (Proposition 1). Determinism is therefore preferable in the weaker sense that it introduces no noise term of its own, not in the sense of a general error-propagation theorem; the operating contrast is measured, and Appendix C supplies it. For the reflected dynamics on the piecewise-smooth manifold $\mathcal{M}$ the stationary law is likewise the truncated Boltzmann distribution of the augmented energy (Theorem 4); because the Pauli barrier suppresses the density at $\partial\mathcal{M}$, the truncation carries no error where it matters, and any structure with all pairs a margin above the threshold has $U_{\text{Pauli}} = 0$.

Theorem 1 (Equivalence to annealed Langevin dynamics)**.** *Substituting* $s^{aug}(x) = -\beta\nabla_x[U_{NNP}(x) + U_{Pauli}(x)]$ *for* $\nabla_x \log p_t(x)$ *in the reverse SDE (Anderson, 1982) gives*

$$dx = \left[f(x,t) + \beta g(t)^2 \nabla_x [U_{NNP} + U_{Pauli}]\right] dt + g(t)\, d\bar{W}_t. \tag{11}$$

For the variance-exploding process, $f(x,t) = 0$ *and* $g(t) = \sqrt{d\sigma^2/dt}$*, so the drift reduces to* $\beta g(t)^2 \nabla U$*—the gradient-dependent drift of overdamped Langevin dynamics. With its unit noise coefficient this reverse SDE carries a diffusion* $g^2/2$ *and a drift twice the value the Einstein relation permits at that diffusion, so its own fixed-*σ *steady state is* $e^{-2\beta U}$*, not* $e^{-\beta U}$*. Eq. (4) is a second, separate Langevin convention—drift* $\sigma^2 s^{aug}$ *with noise* $\sqrt{2}\sigma$*—and in that convention the conditional stationary distribution at fixed noise level is exactly*

$$p(x \mid \sigma_k) \propto e^{-\beta[U_{NNP}(x) + U_{Pauli}(x)]}, \tag{12}$$

independent of σ_k*. The implementation uses a different but equivalent normalization: the score enters un-normalized and the noise level is carried by the noise coefficient instead,* $x_{t+\Delta t} = x_t + \alpha_k s^{aug}(x_t) + \sqrt{2\alpha_k}\,\sigma_k \boldsymbol{\epsilon}$*, so the drift is* $-\alpha_k \beta \nabla U$ *and the diffusion coefficient is* $D = \alpha_k \sigma_k^2$*; the Einstein relation then fixes the stationary law, which is*

$$p_k(x) \propto e^{-\beta[U_{NNP}(x) + U_{Pauli}(x)]/\sigma_k^2}, \tag{13}$$

a Boltzmann law at the effective temperature $T_{eff}(\sigma_k) = \sigma_k^2/(k_B\beta)$*. With* $T_0 = 1/(k_B\beta) = 300\,K$ *for our* β *and* σ_k *in Å,* T_{eff} *runs from* T_0 *at* $\sigma_k = 1\,Å$ *to* $3 \times 10^{-2}\,K$ *at* $\sigma_{\min} = 0.01\,Å$*. The two conventions are related by the Einstein relation rather than by a rescaling of the score: the* σ^2*-prefactor convention fixes the diffusion coefficient at* σ*-independent value and sits at inverse temperature* β *for every* σ_k*, whereas the implemented chain scales its diffusion coefficient as* σ_k^2 *and therefore targets a temperature that anneals to zero—the appropriate target for a dataset of relaxed structures such as MP-20. It is that temperature schedule, and not a target fixed across noise levels, that carries the chain from exploration to exploitation in our runs; the calibration fixes the step-to-noise ratio of the implemented form.*

Theorem 2 (The score is conservative by construction)**.** *Let* $s_{NNP}(x) = -\beta\nabla_x U_{NNP}(x)$ *with* U_{NNP} *twice differentiable. By Clairaut's theorem on the symmetry of mixed partial derivatives,* $\partial(s_{NNP})_i/\partial x_j = \partial(s_{NNP})_j/\partial x_i$ *for all* i, j*: the Jacobian of* s_{NNP} *is everywhere symmetric and* $\nabla \times s_{NNP} \equiv 0$*. Both statements are made in the Cartesian chart. The curl is chart-dependent—in the fractional chart the same physical field acquires a nonzero curl whenever* L *is not orthonormal—so a nonzero fractional-chart curl would not contradict this result; it reflects the change of representation of Eq. (1), not a non-conservative physical field. The same holds for* $s^{aug} = -\beta\nabla[U_{NNP} + U_{Pauli}]$*. Conservativity is structural—it follows from the definition of the score as an energy gradient, and no training, parameterization, or regularization choice can break it. Three consequences follow.*

- Correct steady state. *For the fixed-noise chain of Eq. (4) (*σ^2*-prefactor convention), the Fokker–Planck steady state satisfies* $p(x) \propto e^{-\beta U(x)}$*, independent of the noise level; the implemented chain carries the noise level in its diffusion coefficient instead of in its score and targets* $e^{-\beta U/\sigma_k^2}$ *(Theorem 1). (The reverse SDE of Theorem 1 is time-inhomogeneous; with the exact time-dependent score* $\nabla \log p_t$ *its marginal is the noised density* p_t *and its terminal marginal is the data distribution* p_0*, whatever that distribution is. Substituting the time-independent field* $-\beta\nabla U_{tot}$ *for* $\nabla \log p_t$ *is exact only where* $p_t \propto e^{-\beta U_{tot}}$*, i.e. only if the data themselves follow the Boltzmann law the score targets and the forward noising matches it; conservativity guarantees the* form *of a steady state, not the marginal preservation that the substitution drops. What the substituted process is guaranteed is the fixed-noise steady state of Theorem 1—and in the implemented convention its* $T_{eff}(\sigma_k) \to 0$ *limit. The gap between those two targets is exactly the data-versus-Boltzmann mismatch of Remark 1, and methods that substitute a potential gradient for the score are accordingly best described as potential-driven annealing or gradient flow, not as marginal-preserving diffusion samplers.)*
- Monotone descent in the deterministic limit. *The probability-flow ODE satisfies* $dU/dt = \sigma\dot{\sigma}\beta\|\nabla U\|^2 \le 0$ *during generation, where* $\dot{\sigma} < 0$*. In the zero-noise limit the discrete annealed step reduces to gradient descent,* $U_{NNP}(x_{t+1}) \le U_{NNP}(x_t) + \mathcal{O}(\alpha^2)$*; with the implemented update the drift step reads* $x_{t+1} = x_t - \alpha_k\beta\nabla U_{NNP}(x_t)$*, i.e.* $x_t - \alpha_0\beta\nabla U_{NNP}(x_t)\,\sigma_k/\sigma_{\max}$ *for the schedule* $\alpha_k = \alpha_0\sigma_k/\sigma_{\max}$*: the drift step itself* shrinks *in proportion to* σ_k *as the chain cools,*

while the noise amplitude shrinks as $\sigma_k^{3/2}$, so the drift-to-noise ratio grows as $\sigma_k^{-1/2}$. It is that ratio, and not the absolute step size, that carries the chain from exploration into exploitation, and the descent guarantee is unchanged.

- Structural contrast with learned scores. *A learned score $s_\theta(x,t)$ has an unconstrained Jacobian and is under no requirement to be the gradient of an energy; parameterizing it as one restores conservative forces and the physical sampling guarantees that follow (Plainer et al., 2025). Non-conservative fields admit no corresponding energy function and therefore no descent guarantee.*

Theorem 3 (Approximation error of the time-independent score)**.** *Let U_{NNP} be κ-smooth on the region visited, $\|\nabla U_{NNP}(x) - \nabla U_{NNP}(y)\| \le \kappa\|x-y\|$ for x, y in that region (a global constant is not meaningful for a learned potential, whose gradient is unbounded in the extrapolation regime), and let $\mu = \mathbb{E}_{x_0|x,\sigma}[x_0]$ be the posterior mean. Then the error $\epsilon(x,\sigma) = s_{true}(x,\sigma) - s_{NNP}(x)$ satisfies*

$$\|\epsilon(x,\sigma)\| \le \beta\kappa\|x-\mu\| + \left\|\frac{\mu - x}{\sigma^2} + \beta\nabla U_{NNP}(\mu)\right\|. \tag{14}$$

The first term is a smoothness term. Tweedie's identity gives $\|x-\mu\| = \sigma^2\|s_{true}(x,\sigma)\|$ exactly, so it scales as σ^2 when the data mode is broad compared with σ and as σ when it is narrow ($\rho \ll \sigma$), as for the near-δ modes of a relaxed-structure dataset; in either case it vanishes as $\sigma \to 0$. The second term is a model-mismatch term and does not *vanish with σ. Writing Tweedie's identity for the limit, $(\mu - x)/\sigma^2 \to \nabla \log p_0(x_0)$ as $\sigma \to 0$ with $x \to x_0$ a data point, it tends to*

$$\left\|\nabla \log p_0(x_0) + \beta\nabla U_{NNP}(x_0)\right\| = \Big\|\underbrace{\nabla \log p_0(x_0) + \beta\nabla U_{DFT}(x_0)}_{\text{data/Boltzmann mismatch}} + \underbrace{\beta[\nabla U_{NNP}(x_0) - \nabla U_{DFT}(x_0)]}_{\text{force error}}\Big\|, \tag{15}$$

and it equals the force error $\beta\|\nabla U_{DFT} - \nabla U_{NNP}\|(x_0)$ only *under the additional hypothesis that the data distribution is the Boltzmann distribution of U_{DFT} at inverse temperature β—$p_0 \propto e^{-\beta U_{DFT}}$—so that $\nabla \log p_0(x_0) = -\beta\nabla U_{DFT}(x_0)$. That hypothesis is not implied by a small force error, by a small first term, or by conservativity, and a database of relaxed structures does not satisfy it at any particular temperature; Remark 1 gives a case where the mismatch term diverges while the force error is identically zero. The term is small when the potential is locally accurate at the posterior mean* and *the data are locally consistent with the Boltzmann target, not merely when the posterior is concentrated within a single basin—a single Gaussian basin with a model curvature off by a factor $1+\epsilon$ already gives a term of order ϵ at vanishing σ. This residual is the model-level error discussed in Remark 2.*

Remark 1 (The data distribution is not the Boltzmann distribution)**.** *The identity $\nabla \log p_\sigma(x) = (\mu - x)/\sigma^2$ is exact, so the accuracy of the substituted field $-\beta\nabla U_{NNP}$ as a* score *is not the same question as the accuracy of U_{NNP} as a* potential*. Three statements that the error decomposition of Eq. (14) keeps apart are often conflated: (a) the cooling marginals contract toward the data as $\sigma \to 0$; (b) the score-field error vanishes in that limit; (c) the residual equals the DFT force error. (a) holds for the exact score by construction and is not a statement about U_{NNP}; (b) and (c) are not implied by a small force error, and (c) requires $p_0 \propto e^{-\beta U_{DFT}}$ (Eq. 15).* The mismatch term can diverge while the force error is zero. *Take $p_0 = \delta_0$ and an approximate potential that is* exactly *right, $U_{NNP} = U_{DFT} = \frac{1}{2}kx^2$. At $x = \sigma z$ with $z \neq 0$ the noised density is Gaussian, $p_\sigma = \mathcal{N}(0,\sigma^2)$, so $s_{true}(x,\sigma) = -x/\sigma^2 = -z/\sigma$ and $s_{NNP}(x) = -\beta k\sigma z$; the posterior mean is $\mu = 0$, and the second term of Eq. (14) is evaluated at that mean, so it equals $\|(\mu - x)/\sigma^2 + \beta\nabla U_{NNP}(\mu)\| = \|-z/\sigma + \beta k \cdot 0\| = |z|/\sigma$, which diverges as $\sigma \to 0$. The potential, its gradient, and the force error are exact throughout. What diverges is the mismatch between the* density *the data actually have and the density the score targets.* What the mismatch measures here, and why it is not the error this paper is about. *For a dataset of DFT-relaxed structures the mismatch is a mismatch of mode* weights *and mode* widths *rather than of forces: a relaxed-structure database concentrates near-δ mass on whichever polymorphs were computed, whereas the Boltzmann law at inverse temperature β weights them by $e^{-\beta E}$ and broadens each minimum to $\sqrt{k_B T/k}$. Both terms of the mismatch vanish at the data modes themselves—$\nabla \log p_0 = 0$ at a density mode and $\nabla U_{DFT} = 0$ at a relaxed structure—so the residual of Eq. (15) is a statement about which minima are visited and with what population, not about the local accuracy of the field a trajectory follows. This work measures the local object: the sampler's target in the implemented convention is $e^{-\beta U_{tot}/\sigma_k^2}$ (Theorem 1) and its $\sigma_{\min} \to 0$ limit is a local minimum of U_{tot} reached from the given initial condition, which is what the relaxation stage of a CSP pipeline is asked to produce and is insensitive to the population weights*

the mismatch describes. The method is accordingly a potential-driven anneal and gradient flow, and no result here should be read as sampling the equilibrium ensemble of U_{NNP}, let alone of U_{DFT}.

Corollary 1 (Reliability condition)**.** *The smoothness term of Theorem 3—and with it the single-basin premise—is controlled when*

$$\sigma \ll \frac{R_{basin}}{\sqrt{d}}, \tag{16}$$

with R_{basin} the characteristic radius of the dominant basin and d the configuration-space dimension. The model-mismatch term of the same theorem is not controlled by this condition: it requires the potential to be locally accurate at the posterior mean, and it is what §5.2 measures. Taking the geometric estimate of the basin radius, $R_{basin} \approx r_{NN} - 0.5\,\text{Å}$ (the clearance a nearest neighbour leaves), with $d = 3N + 9$, gives a nearly flat $\sigma \approx 0.30$–$0.43\,\text{Å}$ across the scan compositions, whereas the measured collapse scale spans $17.5\times$ on the crossing definition of Table 16 (0.28 Å to beyond the 5 Å edge of the scan) and $2.7\times$ on a definition-free midpoint (1.26–3.42 Å)—both far outside the geometric range—because the boundary's location is set by the potential's learned short-range extrapolation rather than by basin geometry. The condition is a conservative sufficient bound on the smoothness term, not predictive of the collapse scale.

Remark 2 (Bounded long-range truncation error)**.** *Theorem 3 bounds the error of using the potential's own gradient in place of the true score; it does not cover the model-level error of U_{NNP} relative to U_{DFT}, which persists as $\sigma \to 0$. A distinct error source is the finite descriptor cutoff (∼5–6 Å for MACE and eSEN): interactions beyond it are truncated to zero, discarding dispersion ($\propto -r^{-6}$) and unscreened electrostatics ($\propto 1/r$ for ionic systems). Unlike the short-range failure of §4.1, where the true force is unbounded and the prediction can carry the wrong sign, this error is bounded: the neglected forces decay algebraically while the model force decays smoothly to zero, so it never diverges or reverses. In a crystal the truncation is structural—every atom has infinitely many periodic images, so the neglected terms are absent from every evaluated structure. It is therefore benign for trajectory stability at fixed cell, but it biases the sampled distribution slightly, concentrated in the early stage where interatomic distances are large. The lattice degree of freedom is the exception: with the binding attraction truncated, a stress-driven lattice update at large volume has no restoring term and can inflate without bound, which the outer wall of Eq. (3) constrains.*

Remark 3 (Effective basin radius under augmentation and reflection)**.** *The Pauli augmentation and the reflecting boundary effectively increase the basin radius. Without protection, R_{basin} is set by the potential's energy landscape alone. With the reflecting boundary at $r_{ij} = f_{typ}(R_i^{cov} + R_j^{cov})$, the effective radius is bounded below by the steric constraint, which prevents the trajectory from entering the extrapolation regime. How far $\sigma_{\max}$ can be extended is an empirical question, and $\hat{R}_{basin}$ is extracted from the fitted collapse scale of §5.2 (Appendix H); the induced-rate scan answers it directly, the protected arms holding at or below 1.00% out to the $5\,\text{Å}$ edge of the scan, where the bare curve has been flat under the published normalization since $\sigma_{\max} = 0.7\,\text{Å}$ and is still climbing, more slowly, under the conditional one (Table 15; Appendix H).*

Theorem 4 (Stationary distribution under reflection)**.** *Let $\mathcal{M}$ be the atomic-position projection of the constraint manifold, Eq. (3), with the outer lattice wall acting on the separate lattice degrees of freedom. Let $U_{tot} = U_{NNP} + U_{Pauli}$ be differentiable on $\mathcal{M}$. The premise is a finite barrier, not a divergent one: the potential of Eq. (2) is an exponential in r_{ij} and tends to the finite amplitude $A_{Z_iZ_j}$ as $r_{ij} \to 0$, so the statement below rests on the barrier-to-temperature ratio evaluated* at *$\partial\mathcal{M}$ rather than on $U_{Pauli} \to \infty$. Assuming sufficient regularity of $\partial\mathcal{M}$ (piecewise C^2 with almost-everywhere defined inward normal), the reflected SDE admits a unique stationary distribution. The activation step of Eq. (2) is centred on the wall, so the Pauli energy is finite on $\partial\mathcal{M}$ but large—$U_{Pauli}(\partial\mathcal{M}) = \frac{1}{2}\sum_{i<j:\, r_{ij}=r_{ij}^{wall}} A_{Z_iZ_j} e^{-B_{Z_iZ_j} r_{ij}^{wall}}$, of order 1–5 eV per violating pair for the parameters of Table 4. The exponential lies in the wall's steepness rather than in its height at the boundary, and what suppresses the boundary is the barrier-to-temperature ratio: $e^{-\beta U_{tot}}$ picks up $e^{-\beta U_{Pauli}(\partial\mathcal{M})}$ in the convention of Eq. (4) and $e^{-\beta U_{Pauli}(\partial\mathcal{M})/\sigma_k^2}$ in the implemented one, overwhelmingly in the low-noise tail and by a factor $\mathcal{O}(10)$ for the weakest element pair at $\sigma_k = 5\,\text{Å}$. The reflection boundary condition is therefore inactive in equilibrium and the stationary distribution is the truncated Boltzmann distribution*

$$p_\infty(x) = \frac{1}{Z_{\mathcal{M}}}\, e^{-\beta U_{tot}(x)} \cdot \mathbf{1}_{\mathcal{M}}(x). \tag{17}$$

The truncation incurs no error in the physically relevant region: for any x with $\min_{i<j} r_{ij}(x) \geq R_i^{cov} + R_j^{cov} + \delta$ ($\delta > 0$) we have $U_{Pauli}(x) = 0$ and $p_\infty(x) \propto e^{-\beta U_{NNP}(x)}$, respectively $e^{-\beta U_{NNP}(x)/\sigma_k^2}$ in the implemented convention. Without the Pauli augmentation the same constraint surface would distort *the stationary distribution near the boundary, since the unmodified score supplies no repulsive gradient to counteract accumulation there; the reflection would then serve no corrective purpose. The two components are a gradient/constraint pair, not substitutes. The statement is about the idealized reflected process of Eq. (4); the discrete step of Algorithm 1 is an approximation to it, and Remark 4 separates the terms of the implemented sampler that fall outside this idealization.*

Remark 4 (What the discretization and the auxiliary layers change)**.** *Six elements of the implemented sampler are absent from Eq. (4), and the stationary law of Theorem 4 is a statement about the idealized process rather than a guarantee that survives all six unchanged.* (i) Projection. *Algorithm 1 moves a violating pair along its separation direction until it lies on $\partial\mathcal{M}$. The transport direction is the inward normal, which is the direction Skorokhod reflection pushes along, so this is the standard discrete analogue of the local-time term rather than a specular reflection; but a projection places the pair* on *the boundary instead of returning it to the interior with a reversed normal velocity, the passes are applied sequentially rather than simultaneously to the whole configuration, and the penetration is corrected entirely within one step. The resulting chain approximates the reflected process to first order in the step size and does not reproduce its stationary law exactly at the step sizes used.* (ii) Drift capping. *The per-atom cap is a numerical stabilizer, not a term of the SDE. It is also almost never active: the cap and the step size both scale as σ_k, so the cap binds exactly when the Cartesian force exceeds* `drift_cap`*$/(\alpha_0 \sigma_{\max} \beta) \approx 39$ eV/Å, a threshold independent of the noise level, while typical forces on a valid structure sit one to two orders of magnitude below it. Where it does bind, the update belongs to neither sampler: motion is truncated rather than redistributed, and no energy is conserved. The* `cap_steps` *diagnostic of Appendix B records the affected fraction rather than leaving it implicit.* (iii) Density-adaptive noise. *Layer 2 scales the noise of atom i by $1/(1+\gamma\rho_i)$, making the diffusion coefficient state-dependent while the drift is left unchanged. For a one-dimensional Itô process $dX = b(X)dt + \sqrt{2D(X)}dW$ the zero-flux stationary density is*

$$p(x) \propto \frac{1}{D(x)} \exp\left(\int^x \frac{b(y)}{D(y)}\, dy \right), \tag{18}$$

not $\exp(\int b/D)$ alone: the $1/D$ prefactor is part of the stationary law, and it is the term that survives when $b = 0$. With a state-dependent D and a drift that is not rescaled in proportion, the L2 arms therefore sample a different target from the L1/L3 arms. In several dimensions the corresponding condition is that b/D be a gradient; a per-atom $D(\rho_i)$ generally does not satisfy it, so Eq. (18) does not carry over to the $3N$-dimensional chain, and no scalar potential is available for the joint law. The layer is retained because what it changes is the noise amplitude in dense regions rather than the location of the boundary, and its effect is measured by the L2-versus-L1L3 comparison of Appendix C rather than assumed. (iv) The lattice gate. *The stress-driven lattice update is switched off outside the near-equilibrium tail, which removes a degree of freedom from the low-noise phase of the schedule: the manifold $\mathcal{M}$ of Eq. (3) is held fixed on that phase rather than sampled jointly with the positions. Theorem 4 is stated for the position process at fixed cell, which is the regime the gate creates; the joint position–cell update, where it is active, is not covered by it, and no joint stationary law is claimed.* (v) Euler discretization. *Eq. (13) and the truncated law of Theorem 4 are stationary distributions of the* continuous *processes, not invariant distributions of the finite-step chain that is actually run. For an Euler–Maruyama discretization of a Langevin diffusion the invariant law differs from the continuous one at $\mathcal{O}(\alpha_k)$, so the correct statement is that the continuous processes have those steady states and that the discrete chain approximates them to first order in the step size. No exact-invariance result (Metropolis adjustment, a higher-order splitting) is used here, and none is claimed; the step size is set by the schedule $\alpha_k = \alpha_0 \sigma_k / \sigma_{\max}$ rather than by an invariance criterion.* (vi) Score normalization. *The normalization mismatch between Eq. (4) and the implemented update is the one non-perturbative difference in this list, and it changes the* target *rather than perturbing it: the implemented chain's steady state is $e^{-\beta U_{tot}/\sigma_k^2}$ at every level, an effective temperature $T_{eff}(\sigma_k) = \sigma_k^2/(k_B\beta)$ that is 300 K at $\sigma_k = 1$ Å and 3×10^{-2} K at $\sigma_{\min}$, whereas the σ^2-prefactor convention holds β fixed (Theorem 1; §4.3). The two are related by the Einstein relation rather than by a rescaling of the score, and the annealing behaviour reported in §5.2 is a statement about the implemented chain and its own target. None of these leaves the failure mode of §4.1 unguarded, and none is claimed to be invisible: the Pauli barrier suppresses the*

density at $\partial\mathcal{M}$ exponentially, so the boundary layer in which (i) and (ii) act carries exponentially little mass, (iii) and (iv) are each isolated by an ablation arm of Appendix C, and (v) is a step-size-ordered approximation rather than a change of target.

Proposition 1 (Score-error propagation under shared and independent noise couplings)**.** *Let s be the exact score, $\hat{s} = s + \epsilon$ the approximate one, with $\|\epsilon\| \le \eta$ on the region visited, and let the drift $\sigma^2 s$ be Λ-Lipschitz there with one-sided (dissipativity) constant $\mu \le \Lambda$, the two being distinct constants: for a confining drift $\mu < 0$ whereas $\Lambda > 0$. Two trajectories of the* same *sampler—one driven by s, one by $\hat{s}$—can be compared under either of two couplings, and the two comparisons disagree.*

- Synchronous coupling. *Let both trajectories share one Brownian path (stochastic sampler) or one initial condition (deterministic sampler). The difference obeys $d\Delta x = \sigma^2[\hat{s}(\hat{x}) - s(x)]\,dt$, in which the injected noise appears on neither side: being common to the two trajectories, it cancels exactly. Grönwall's inequality gives $\|\Delta x(T)\| \le (\eta/\Lambda)(e^{\sigma^2_{\max}\Lambda T} - 1)$: the terminal error carries* no *additive noise term, only the score error amplified by the drift's Lipschitz constant over the integration time.*

- Independent coupling. *Draw the Brownian paths independently, as two separate runs of the same sampler would. The injected noise no longer cancels, and the terminal difference is nonzero* at *$\epsilon \equiv 0$—a property of the noise rather than of the score approximation. The drift still acts on it, so the driftless expression $2\int_0^T g(t)^2 dt$ is not the general answer. Writing $b = \sigma^2 s$ for the drift and letting μ be its one-sided Lipschitz (dissipativity) constant, $\langle b(x) - b(y), x - y\rangle \le \mu\|x - y\|^2$—a two-sided Lipschitz constant cannot be used here, since it does not distinguish a confining drift from an expansive one—Itô's formula on $\|\Delta x\|^2$ and Grönwall's inequality give, in one dimension,*

$$\mathbb{E}\|\Delta x(T)\|^2 \le 2\int_0^T e^{2\mu(T-s)} g(s)^2\, ds, \tag{19}$$

  which the Ornstein–Uhlenbeck process attains: $dX = -\lambda X\,dt + g\,dW$ with a common initial condition gives $\mathbb{E}\|\Delta x(T)\|^2 = (g^2/\lambda)(1 - e^{-2\lambda T})$ exactly, and Eq. (19) with $\mu = -\lambda$ reproduces it. In d dimensions the noise trace carries a factor d; nothing else changes. Two consequences follow. The driftless expression is the $\mu = 0$ case, which is the forward noising process rather than the sampler. And with $\mu \ne 0$ the term acquires a T-dependence of its own: the weight $e^{2\mu(T-s)}$ is largest for the noise injected earliest when the drift is expansive ($\mu > 0$) and smallest when it is confining ($\mu < 0$), so the terminal difference grows with the integration time in the former case even at an exact score.

The noise term therefore enters the stochastic sampler's terminal error only under the independent coupling, where it is present even for an exact score. Neither comparison establishes an advantage for the deterministic sampler in propagating *score error. What is established is weaker and coupling-free: PF-ODE injects no noise, so its error budget contains no noise term of its own, whereas the stochastic sampler's contains the right-hand side of Eq. (19)—of order $\sigma_{\max}$ for the variance-exploding normalization of Theorem 1, where $\int_0^T g^2 dt = \sigma^2_{\max} - \sigma^2_{\min}$ is fixed by the endpoints rather than by T, and T-dependent only through the drift factor $e^{2\mu(T-s)}$. This is not the same as determinism suppressing score error more effectively, because the deterministic flow's own error is amplified exponentially in T, and its long integrations are exactly the ones the evaluation cap truncates. The sampler comparison is therefore reported as measurement, not as a theorem: Appendix C finds unprotected PF-ODE 4.7–7.4× safer than unprotected ALD across $\sigma \in [0.5, 5]$, and not OOD-free. A general error-propagation statement would require a specified coupling, an error metric, and regularity conditions on the drift; the two couplings above show the choice is not innocuous, and this paper does not make it.*

J Outlook: Further Universal Components

The design rule suggested by §6—identify which components of the target physics are universal, supply them analytically, and use learning only for the remainder—extends beyond short-range repulsion along three further axes, and carries implications outside crystals.

Long-range dispersion. Local message-passing architectures with a finite descriptor cutoff capture the universal C_6/r^6 dispersion asymptote poorly, and in a crystal every atom has infinitely many

periodic images, so the neglected terms are absent from every evaluated structure. An analytic dispersion term could be added on exactly the logic of the Pauli wall, with the bounded-truncation analysis of Remark 2 as its error budget.

Truncated electrostatics. For ionic systems the neglected $1/r$ terms are longer-ranged than dispersion and less forgiving of truncation. An analytic Ewald term (Ewald, 1921) using partial charges predicted by the potential—or assigned by charge equilibration—would extend the same decomposition to the electrostatic component.

Nuclear quantum corrections. For light elements at low temperature, harmonic quantum corrections from a potential-derived Hessian could refine stability rankings as a post-processing step, supplied analytically rather than learned.

Building potentials differently. Deliberately including short-range repulsive configurations in training sets could close the gap at the learning stage rather than requiring analytic augmentation. DFT convergence at short interatomic distances makes such data expensive, and it is an open question whether a learned short-range repulsion generalizes as reliably as the analytic form. A modular alternative—separate learned modules for long-range chemistry coupled to analytic modules for universal physics—would improve extrapolation robustness and allow each module to be improved independently.

Outside crystals. The same structure applies to molecular potentials trained on equilibrium geometries, which are equally blind at short range. Transition-state searches deviate furthest from equilibrium and would benefit most from analytic repulsion, and the analytic component is equally usable in free-space rather than periodic settings. The boundary is also a natural acquisition signal for active learning (Vandermause et al., 2020), since configurations near it are the frontier of the potential's reliable domain. Finally, cheap geometric features such as the minimum interatomic distance could serve as physics-informed OOD detectors without ensembles, because the potential's uncertainty is systematically elevated near the constraint surface.