BREAKING OBSERVABLE DEGENERACY IN DIFFUSION MODELS OF AMORPHOUS MATTER

Anonymous authors
Paper under double-blind review

ABSTRACT

Diversity in conditional samples from a generative model can reflect either creativity or hallucination. In the physical sciences, this ambiguity is critical when inverting an experimental measurement, as choosing among the solutions requires an intractable ground truth that the generative model is meant to replace. Using disordered materials as a challenging case study, we show that a score-based denoiser systematically produces samples of high configurational entropy that the conditional input cannot distinguish from the reference. This amounts to a residual excess temperature, raising the concentration of outliers in the system and compromising the physical behavior of the samples. To break this degeneracy, we implement four modifications to sampling routines that suppress this residual noise: a gate interpolating between the reverse SDE and the probability-flow ODE; guidance on the low-frequency bands that the experimental observable leaves undetermined; a structural relaxation projected onto the null space of the observable's Jacobian; and a ladder of partial restarts. Together, these measures halve the excess entropy on benchmark systems and allow us to reach amorphous configurations that quench kinetics have kept out of reach of direct simulation for decades.

1 INTRODUCTION

Inverse problems underlie most of scientific characterization, from resolving protein conformations in cryo-electron microscopy (Scheres, 2012) and assigning molecular structures from spectroscopy (Jonas, 2019) to reconstructing atomic arrangements from diffraction and imaging in materials and engineering (Fienup, 1982; Miao et al., 1999; Billinge & Levin, 2007). In these problems, the structure or quantity of interest is rarely observed directly and must be inferred from indirect, noisy, and often incomplete measurements (Tarantola, 2005). Generative models are increasingly used to close this gap, producing samples compatible with measurements and constraints, or optimized toward target properties (De Breuck et al., 2025). Their evaluation, however, is largely inherited from machine learning, where a model is judged by whether its set of outputs reproduces the statistics of a reference set (Heusel et al., 2017), separates fidelity from diversity (Sajjadi et al., 2018), or differs from memorization (Alaa et al., 2022). These approaches compute properties of distributions rather than of individual samples, and depend strongly on the feature space of interest (Stein et al., 2023). For many inverse problems constrained by physics, this is insufficient. When a generative model inverts an experimental measurement, telling apart creativity from hallucination often requires an intractable oracle that the model was meant to replace. At the same time, an ensemble can match target distribution statistics while its individual members are physically inadmissible.

In materials science, amorphous structures such as liquids, polymers, and glasses further exacerbate this problem. Their structures lack long-range order and are commonly determined by inverting scattering data, most often through the pair distribution function (PDF) $g_2(r)$ (Billinge & Levin, 2007). In practice, however, distinct structures and potentials yield PDFs that are degenerate at any accessible precision (Wang et al., 2020; Gettler et al., 2021). At the same time, obtaining high-quality references is a tremendous challenge. Low-defect structures compatible with low experimental cooling rates are challenging to sample in nanosecond-long simulations, which bias the picture of what amorphous structures should look like. High-quality samples, therefore, require rare, stochastic events across thousands of trajectories (Rosset et al., 2026), some of which may never be sampled in tractable timescales. Thus, generative models for materials trained on imperfect simulations and

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.

conditioned on a degenerate observable carry two complementary sources of error that limit their use in inverse problems: a biased prior and an underdetermined constraint.

Here, we show that a score-based conditional sampler systematically produces amorphous structures of higher configurational entropy than the reference, even when their PDFs are indistinguishable from the target. We quantify this bias per sample using an information-theoretic framework, which separates structures that correlation functions cannot distinguish. We identify an excess entropy, which behaves as a residual temperature due to sampling, raising the concentration of coordination defects and increasing the energy of the samples. To suppress this residual noise without retraining the score network, we introduce four modifications to the sampling routine: (1) a gate interpolating between the reverse stochastic differential equation and the probability-flow ordinary differential equation (Song et al., 2021b; Karras et al., 2022); (2) guidance on the low-frequency band of the structure factor that $g_2(r)$ leaves undetermined; (3) a structural relaxation projected onto the null space of the Jacobian of the observable; (4) a ladder of partial restarts (Xu et al., 2023). On amorphous silicon (a-Si), these modifications halve the excess entropy while keeping the conditioned observables within the variability of the reference ensemble. With a prior trained on a single defect-free structure (Rosset et al., 2026), the modified sampler generates structures that overlap with an independent set of defect-free a-Si models. These are configurations that quench kinetics have kept out of reach of direct simulation for decades. The results apply beyond amorphous structures, and we show that whenever the conditioning input underdetermines the solution, evaluating a conditional generative model requires per-sample, physically grounded criteria in addition to distribution-level agreement.

2 Related Work

Sample quality in generative models. Because a generative model's likelihood is often unavailable to measure sample quality (Theis et al., 2016), quality is assessed by checking whether the generated distribution matches the data distribution. Traditional metrics include distribution distances such as the Fréchet Inception Distance and its variants ($\text{FD}_{\text{DINOv2}}$, CMMD, KID) (Heusel et al., 2017; Stein et al., 2023; Jayasumana et al., 2024; Bińkowski et al., 2018), decompositions into fidelity versus diversity via precision-recall and its density/coverage refinements (Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Naeem et al., 2020) and reference-free diversity scores (Friedman & Dieng, 2023). However, a verdict on an individual sample requires a curated, domain-specific oracle, available only in some fields, e.g., GuacaMol for molecules (Brown et al., 2019) or self-consistency RMSD for protein backbones (Trippe et al., 2023; Yim et al., 2023; Watson et al., 2023). Many materials science problems have no such oracle to compare against, and such oracles often have to be developed on a case-by-case basis.

Generative models for materials. Generating inorganic crystals conditioned on compositions or target properties has been extensively studied from the perspective of structure prediction (Woodley & Catlow, 2008; Oganov & Glass, 2006), representation learning (Damewood et al., 2023), or generative models (De Breuck et al., 2025). Diffusion models such as CDVAE (Xie et al., 2022), DiffCSP (Jiao et al., 2023), and MatterGen (Zeni et al., 2025) tackle the crystal sampling objective via progressive denoising of element identities, atom positions, and unit cells. Often, samples are evaluated according to property thresholds or stability/uniqueness metrics (Xie et al., 2022; Zeni et al., 2025; Szymanski & Bartel, 2025). None of these metrics apply directly to amorphous matter. Its structures have no ground state to score against, are strongly dependent on processing conditions, and are elusive to recover from experimental measurements (Wang et al., 2020; Gettler et al., 2021; Yong et al., 2024).

Generative models and amorphous structures. Few generative models target amorphous matter directly. Variational Autoencoders (Comin & Lewis, 2019; Chen et al., 2026) have been used as proof-of-concept, but often produce structures with excessive disorder or limited diversity. More recent diffusion models (Kwon et al., 2024; Yang & Schwalbe-Koda, 2026; Guo & Schwalbe-Koda, 2026; Finkler et al., 2026; Lin et al., 2026) produce higher quality structures by focusing on the relevant short- and medium-range order metrics, as well as quantifying the relevant distributions constrained on observables. Newer models (Cheng et al., 2026) propose to bootstrap Boltzmann-distributed structures with a given energy function and use them to generate samples across sizes,

temperatures, and compositions. Nevertheless, the risk for any observable-conditioned inversion remains unaddressed. Past research on Reverse Monte Carlo (RMC) shows convergence to the most disordered structure consistent with the data (McGreevy & Pusztai, 1988; Billinge & Levin, 2007). Hybrid schemes (Soper, 1996; Opletal et al., 2008) introduce a potential or hand-tuned constraints as priors, but still struggle to sample low-energy structures (Ehrhardt et al., 2026).

3 Formalizing the observable degeneracy problem

One of the main observables used to characterize disordered structures is the pair distribution function (PDF), $g_2(r)$. The PDF encodes the probability, relative to an ideal gas at the same number density $n = N/V$, of finding two atoms at distance r (Hansen & McDonald, 2013). Its Fourier transform is the structure factor $S(k)$, the normalized scattered X-ray or neutron intensity that an experimental beamline measures, and hence the standard characterization technique. In the exact limit, g_2 carries a strong uniqueness guarantee. For a classical system of N particles with pairwise-additive energy

$$\Phi(\mathbf{r}^N) = \sum_{i<j} v\big(|\mathbf{r}_i - \mathbf{r}_j|\big) \tag{1}$$

in the canonical ensemble at number density n and temperature $T > 0$, Henderson (1974) demonstrated that if two pair potentials v and v' give rise to the same equilibrium pair distribution function, $g_2(r) = g_2'(r)$ for all r, then $v'(r) = v(r) + c$ for some constant c, i.e., the map $v \mapsto g_2$ is injective up to an additive constant. The proof rests on the Gibbs inequality for the two canonical distributions $w_\alpha \propto e^{-\Phi_\alpha/k_B T}$,

$$\int w_0 \ln w_1 \, \mathrm{d}\mathbf{r}^N \;\le\; \int w_0 \ln w_0 \, \mathrm{d}\mathbf{r}^N, \tag{2}$$

with equality if and only if $w_0 = w_1$, combined with the fact that the mean energy depends on the potential only through g_2 (Wang et al., 2020),

$$\frac{\langle \Phi \rangle}{N} = \frac{n}{2} \int_0^\infty v(r)\, g_2(r)\, 4\pi r^2 \, \mathrm{d}r. \tag{3}$$

In practice, generative models such as GLASS (Guo & Schwalbe-Koda, 2026) learn a prior over atomic configurations obtained from atomistic simulations without restraint on $g_2(r)$. At inference, new samples are produced by reversing the diffusion process with a posterior-score guidance term that pushes the generated structure toward an observable such as $g_2(r)$. In principle, Henderson's theorem guarantees the uniqueness of structures given $g_2(r)$, but the equivalence is guaranteed only in the exact limit (Wang et al., 2020). To establish a metric that determines iso-$g_2(r)$ structures, we adopt the distance from Wang et al. (2020) written in the reduced length $r_s = r\, n^{1/3}$ so that structures of different densities can be compared:

$$D_{g_2} \;=\; 4\pi \int \left[g_2(r_s) - g_2^{\mathrm{ref}}(r_s)\right]^2 r_s^2 \, \mathrm{d}r_s. \tag{4}$$

D_{g_2} thus quantifies errors in g_2 affecting thermodynamic averages. To generalize this, we introduce two modifications: first, the integral in Eq. (4) is truncated at the second minimum of g_2, the outer edge of the second coordination shell, the range on which the sampler is often conditioned. Second, the threshold for declaring two functions iso-g_2 is calibrated on the spread of the reference molecular dynamics (MD) ensemble itself rather than on absolute thresholds.

4 Physical origins of structural degeneracy

Counterexamples to Henderson's theorem. Markedly distinct potentials can yield PDFs that are numerically indistinguishable over any range and to any precision ever accessible (Wang et al., 2020). This remains true under abrupt variations of the potential. To demonstrate this point, we trained a GLASS model on the Lennard-Jones (LJ) liquid and sampled structures conditioned on the equilibrium $g_2(r)$ for the system. Figure 1 showcases the counterexamples obtained for this LJ liquid: the potential is varied along a degenerate direction, one in which every member of the family reproduces the same $g_2(r)$ within the criterion of Eq. (4). Two points deserve emphasis. First, the

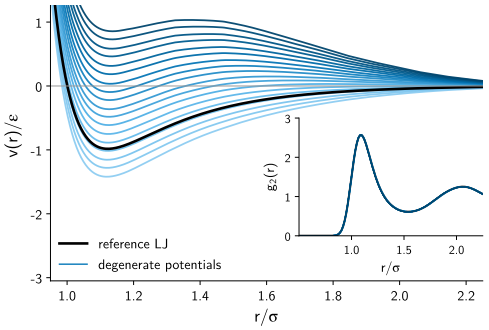


Figure 1: **Distinct pair potentials map to a single, degenerate $g_2(r)$.** Selected set of distinct pair potentials (blue) against the reference Lennard-Jones potential (black), for the LJ liquid at $\rho^* = 0.85$, $T^* = 0.90$, $N = 216$. Each potential was assigned to a structure produced by the GLASS sampler and validated by its own NVT MD run. The resulting $g_2(r)$ (inset) matches the reference with $D_{g_2} \leq 1.5 \times 10^{-3}$, below the simulation error of $\sim 10^{-2}$.

generative model can return families of structures corresponding to distinct potentials for a target $g_2(r)$. Second, because all structures satisfy the target condition but represent distinct physical behavior, a means of distinguishing among such solutions is indispensable, especially if it does not rely on potential energy calculations that are challenging to obtain in real-world applications.

Breaking the degeneracy with information entropy. We propose that structural degeneracies in atomistic samples can be measured beyond many-body correlation functions (Stillinger & Torquato, 2019) via information theory. The information entropy $\mathcal{H}$ of an atomistic system is obtained following the QUESTS method (Schwalbe-Koda et al., 2025), where each atom i is described by a vector X_i of inverse distances to its k nearest neighbors within r_c, and the entropy of the set of N environments is the kernel-density estimate (KDE)

$$\mathcal{H} = -\frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{1}{N}\sum_{j=1}^{N}\exp\left(-\frac{\|X_i - X_j\|^2}{2h^2}\right)\right], \tag{5}$$

in units of nats, bounded by $0 \leq \mathcal{H} \leq \log N$, and default parameters $k = 32$, $r_c = 5\,\text{Å}$ and $h = 0.015$. Given the known tendency of Monte Carlo techniques to sample high-entropy configurations (McGreevy & Pusztai, 1988), we posit that atomistic generative models often sample structures with high information entropy that meet the target. On the other hand, physical constraints may render very high entropy structures unrealistic due to higher associated energies.

To demonstrate the degeneracy-breaking effect of entropy, we used the a-Si case to quantify how diverse the solutions to one $g_2(r)$ are. Figure 2 shows that the information entropy from Eq. (5) effectively separates degenerate solutions from reference configurations (Fig. 2a) despite being iso-PDF (Fig. 2b). To explain these results, Figs. 2c,d show that observable degeneracies overlook other structural quantities such as coordination number distributions and ring statistics. Figures 2a,d illustrate how the number of three-membered rings rises steeply in high-entropy solutions (full analysis

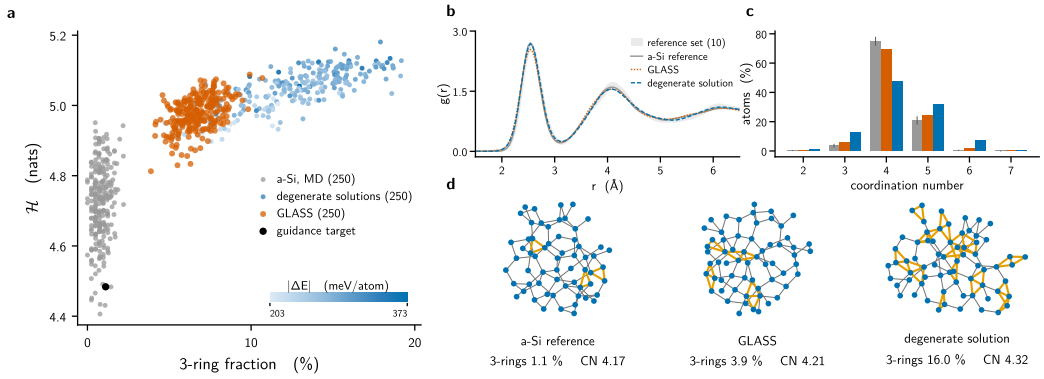


Figure 2: **Breaking the degeneracy of iso-$g_2(r)$ structures.** Distributions of sets of 250 a-Si structures ($\rho = 2.0\,\text{g/cm}^3$, $N = 216$ atoms) for MD references (grey), GLASS solutions (orange), and degenerate solutions (blue). The distributions show the **a**, population, **b**, $g_2(r)$, **c**, coordination number distributions, and **d**, example 60-atom subgraphs per set, with three-membered rings colored in orange. Color in **a** represents the energy above the reference mean obtained with the MACE-MP-0 machine learning potential.

in Appendix A.1). Figure 2a also shows that, though GLASS exhibits state-of-the-art performance in generation quality, its solutions do not yet fully match the reference entropy. Degenerate families go even further, with $\Delta\mathcal{H} = +0.32$ nats above the median reference entropy and energies computed with machine learning interatomic potentials (MLIPs) between 203–373 meV/atom above the reference mean. The fraction of three-membered rings also rises from 0.8% in the reference structures to 12.7% in degenerate systems. Thus, the diversity of generative models comes with a negative bias when inverting characterization data for materials: generated structures are diverse on the high-entropy side of the reference and exhibit room for improvement in sample quality.

Entropy excess as an effective temperature. Figure 3 showcases another perspective on the sensitivity of the degeneracy metrics. For each reference structure, we added an isotropic white Gaussian noise of amplitude σ per atom, mimicking an effective temperature. Such displacements are exactly what Henderson's theorem is insensitive to: at low noise, structures are numerically iso-g_2 despite rising in $\mathcal{H}$. For an example a-Si system at $\rho = 2.0\,\mathrm{g\,cm^{-3}}$, the entropy of the noised references at $\sigma = 0.04\,\text{Å}$ (i.e., 1.7% of the Si–Si bond length) is substantially higher than the reference, while D_{g_2} is still within the numerical error of the reference configurations. Thus, the window within which a noised reference is still iso-g_2 with the reference ensemble covers a large portion of the entropy excess. A generative model whose only structural criterion is $g_2(r)$ therefore cannot tell a reference from its noised copy up to this noise threshold. This reveals the main problem: the sampler introduces a systematic error — a higher effective temperature — that satisfies Henderson's ambiguity rather than generating more physical, lower entropy samples (see full discussion in Appendix A.2).

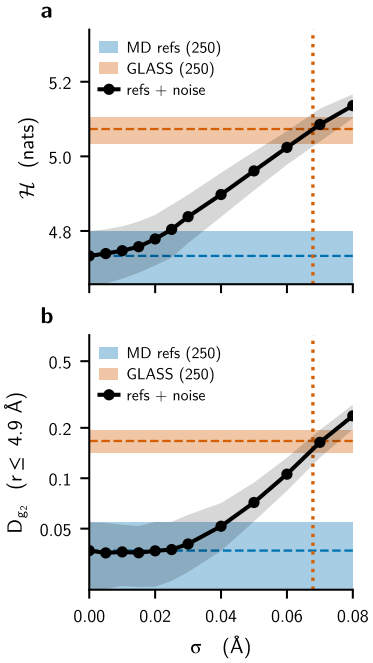


Figure 3: **Henderson's ambiguity and residual temperature. a**, $\mathcal{H}$ and **b**, D_{g_2} ($r \leq 4.9\,\text{Å}$) of 250 MD reference frames for a-Si ($\rho = 2.0\,\mathrm{g\,cm^{-3}}$) after adding a per-atom displacement noise of amplitude σ per component (black: median; grey: interquartile range), against the reference set (blue) and the unmodified GLASS sampler (orange). The dashed line marks $\sigma^* = 0.068\,\text{Å}$, where the entropies meet.

5 Sampling low-entropy solutions

Given the results in the previous section, we assume that samplers with a well-learned prior whose only criterion is $g_2(r)$ return structures of systematically higher configurational entropy than the reference. The excess behaves similarly to a residual, point-by-point noise invisible to PDFs. Thus, to generate higher-fidelity structures, one needs to remove the residual noise from the structures during the generation process to lower $\mathcal{H}$ without moving $g_2(r)$. We propose four modifications to the GLASS sampling strategy, each aimed at reducing one component of the excess and each with a measured partial contribution. None of the modifications requires retraining the score network; they only adapt its inference schedule.

The base sampler. GLASS generates a structure by integrating the reverse-time SDE of the variance-exploding process $x_t = x_0 + \sigma(t)\varepsilon$, $\sigma(t) = kt$, with Euler–Maruyama steps (Song et al., 2021b; Anderson, 1982). We index a pass following Song et al. (2021a): over a decreasing time grid $t_N = T > t_{N-1} > \cdots > t_0$, the state x_i at t_i is carried to x_{i-1} for $i = N, \ldots, 1$, so that x_N is the noisy start and x_0 the sample; with $\Delta t_i = t_{i-1} - t_i < 0$ and $\varepsilon_i \sim \mathcal{N}(0, I)$,

$$x_{i-1} = x_i - \gamma^2(t_i)\big[s_\theta(x_i, t_i) + \ell_{g(r)}(x_i, t_i)\big]\Delta t_i + \gamma(t_i)\sqrt{|\Delta t_i|}\,\varepsilon_i, \qquad \gamma^2(t) = \mathrm{d}\sigma^2/\mathrm{d}t, \quad (6)$$

where $\gamma(t)$ is the diffusion coefficient of the process, $s_\theta \approx \nabla \log p_t$ is the learned score and $\ell_{g(r)}$ is the guidance term that pulls the clean estimate $\hat{x}_i = x_i + \sigma^2(t_i)s_\theta(x_i, t_i)$ (Efron, 2011; Chung et al., 2023) toward the target $g_2(r)$. The first term is the prior, the second the constraint, and the third the injected white noise that makes the sampler stochastic and gives it its diversity.

The modified sampler. The four modifications change Eq. (6) to

$$x_{i-1} = x_i - \tfrac{1+\eta_i^2}{2}\,\gamma^2(t_i)\big[s_\theta + \ell_{g(r)} + \ell_S\big]\Delta t_i + \eta_i\,\gamma(t_i)\sqrt{|\Delta t_i|}\,\varepsilon_i, \tag{7}$$

$$\tilde{x}_0 = \mathcal{Q}(x_0) \quad \text{after each pass,} \qquad x_N = \tilde{x}_0 + \sigma(T)\,\varepsilon \quad \text{to start the next, at its own } T, \tag{8}$$

with $\eta_i = \eta(t_i)$ and the arguments (x_i, t_i) of the three terms omitted. The terms are the following:

(1) Noise gate $\eta(t)$. A factor $\eta \in [0, 1]$ scales the injected noise at inference time. At $\eta = 1$ Eq. (7) is the stochastic sampler of GLASS; at $\eta = 0$ it is the deterministic probability-flow ODE, with no noise at all. The drift coefficient $(1 + \eta^2)/2$ is the value enforced by the Fokker–Planck equation so that every η yields the same marginals for an exact score (Appendix A.3.3; Song et al., 2021b), the continuous form of the DDIM knob (Song et al., 2021a) and of the churn family of Karras et al. (2022). The equivalence breaks with an imperfect score, and noise injected late in the trajectory survives as residual jitter that increases the entropy of the structures. As the noise is not as useful at later stages, η is a step function, $\eta(t) = \mathbf{1}[t > t^*]$. This allows the sampling process to be stochastic while basins are being selected and deterministic during refinement, the continuous analogue of restricting stochasticity in Karras et al. (2022). The step onset was set empirically by a sweep.

(2) Spectral guidance ℓ_S. Structures of higher $\mathcal{H}$ relate to shorter rings, so the generative model emphasizes the short-range network given $g_2(r)$, but cannot constrain the long-wavelength band $k \lesssim 2\pi/r_c$. The second term is a guidance on the structure factor $S(k)$ restricted to the wavevectors commensurate with the periodic cell below a cutoff $k < 0.8\,\text{Å}^{-1}$ to cover the long wavelengths $g_2(r)$ leaves undetermined. The term uses the same posterior-score step and the same residual-normalized step size as $\ell_{g(r)}$ (Chung et al., 2023), so it provides a new "free" observable from the Fourier transform of $g_2(r)$, similarly to GLASS' diffraction guidance (Appendix A.3.5).

(3) Projected relaxation $\mathcal{Q}$. The sampled structure ends away from a local minimum of the physical energy, which may contribute to high-frequency displacements. In our framework, after each generation pass, the structure is relaxed by a potential. We tested both classical interatomic potentials and universal MLIPs (uMLIPs). While uMLIPs are far from perfect for the amorphous space (Fragapane & Deringer, 2026), they still provide meaningful gradients that prevent structural artifacts (e.g., over- or undercoordinated atoms) from forming. Here, we use two representative uMLIPs: PET-OAM-XL 1.0.0 (Bigi et al., 2026) and MACE-MP-0 (Batatia et al., 2025), though other potentials could be employed. Forces are projected onto the null space of $\partial\hat{g}/\partial x$ (Rosen, 1961), simultaneously decreasing the energy while keeping $g_2(r)$ invariant to first order. This contrasts with the hybrid reverse Monte Carlo and FEAR refinements (Opletal et al., 2002; Pandey et al., 2015), where independent penalties are applied. The relaxation is slower than the diffusion model inference step and also truncated to prevent excessive relaxations of intermediate steps.

(4) Ladder of partial passes. The three steps above are composed into a single pass, then repeated from decreasing restart times T: the relaxed sample $\tilde{x}_0$ is re-noised by the exact amount $\sigma(T)\varepsilon$ (Meng et al., 2022), denoised again under Eq. (7) with $\eta \equiv 0$, and relaxed again. This is the Restart iteration of Xu et al. (2023) with a geometrically decreasing amplitude, as in the annealed schedule of Song & Ermon (2019). The entropy often saturates after two full cycles. The complete procedure, with all settings, is stated in Appendix A.3.8.

In brief, our proposed sampling procedure introduces: η, which controls the noise that gives the model its diversity but avoids residual effective temperatures; ℓ_S, which supplies fidelity at long spatial correlations; $\mathcal{Q}$, which adds relaxations to remove high-frequency imperfections and reduce the system energy; and an annealing process that repeats the cycle a finite number of times. Additional details are provided in Appendix A.3 and Table 3.

6 Experiments

Setup. We used a-Si as an archetypical example of an amorphous material that is technologically relevant and challenging to generate at high quality. Structures are generated at density $\rho = 2.223\,\text{g/cm}^3$ with the Stillinger–Weber potential and a melt-quench process using MD simulations implemented in LAMMPS. Full details on the dataset are provided in Appendix A.3.8 and AppendixA.6. The average of 40 structures provided the $g_2(r)$ target during generation, and the 250 held-out structures were used as the reference set. The overlap between the reference distributions

of held-out and guidance structures obtained from the KDE in Eq. (5) is 0.612 against the held-out set's own 0.619, establishing the reference of "novelty" in generated structures (Schwalbe-Koda et al., 2025).

Ablation. Each row of Table 1 progressively adds one modification in Section 5 to the sampling algorithm. D_{g_2} is measured against the mean of the held-out set, which is why the held-out row is not zero. E is evaluated with the PET-OAM-XL 1.0.0 (Bigi et al., 2026) uMLIP, a potential independent of MACE-MP-0, which supplies the forces inside $\mathcal{Q}$. Overlap is the fraction of a structure's environments whose differential entropy against the held-out pool falls below a fixed threshold, set to the median of the entropies measured between two halves of that pool. It is not comparable across densities.

Table 1 shows that each component of the modification progressively lowers $\mathcal{H}$, with no single term dominating the process. The last row is a selected tail, the 25 lowest-$\mathcal{H}$ structures of the complete-recipe arm; $\mathcal{H}$ is hypothesized to be the main metric to optimize, all else being equal in D_{g_2}. Together, the total entropy reduction of 0.38 nats for the top structures corresponds to a 78% reduction in the excess entropy of the system, and stays within the error bar of the reference set. Accordingly, the energy drops to 80 meV/atom above the reference levels and the overlap rises monotonically. The two constrained observables $g_2(r)$ and $S(k)$ end closer to the target than the intrinsic variability within the reference set. Full distributions in Appendix A.5 further demonstrate the effectiveness of the method compared to the original generative method.

Table 1: **Ablation study.** Generation of a-Si structures ($\rho = 2.223\,\mathrm{g/cm^3}$, $N = 216$ atoms) with our proposed method using a prior model for GLASS trained on a-Si at three densities ($\rho = 1.5, 2.5, 3.5\,\mathrm{g/cm^3}$). $n = 250$ structures are generated in each test. Units: $\mathcal{H}$, nats; D_{g_2} and $S(k{<}0.8)$, dimensionless ($\times 10^{-3}$); E, eV/atom.

configuration	$\mathcal{H}$	D_{g_2}	$S(k{<}0.8)$	E	overlap
GLASS, no recipe	4.99 ± 0.06	46	82	-4.96	0.347
$+\ \eta(t)$	4.98 ± 0.05	15	69	-4.89	0.351
$+$ spectral term ℓ_S	4.94 ± 0.05	13	34	-4.91	0.378
$+$ projected relaxation $\mathcal{Q}$	4.86 ± 0.06	15	26	-5.00	0.435
$+$ restart ladder (complete recipe)	4.72 ± 0.06	14	29	-5.06	0.517
best 10% of the recipe (25 of 250)	4.61 ± 0.02	13	25	-5.07	0.559
reference set	4.50 ± 0.11	31	33	-5.15	0.619

Is the ladder just more relaxation? Because $\mathcal{Q}$ runs at the end of every cycle, the complete recipe performs four relaxations in total, so the ladder's ablation row could simply reflect more relaxation. To test this hypothesis, we increased the relaxation length in single-pass outputs, with no re-noising or denoising, and verified it accounts for only part of the gain: at a matched budget these ground-out structures reach $\mathcal{H} = 4.66$ nats against the ladder's 4.59 nats. Hence, of the -0.097 nats credited to the ladder, only -0.035 is extra relaxation and -0.062 is the restart cycle itself. Moreover, prolonged relaxation increases the $g_2(r)$ mismatch to 0.017, as large as the recipe's entire budget, whereas the ladder improves it ($0.020 \to 0.016$), because each cycle repairs the drift its relaxation creates. The cycle therefore lowers the entropy while staying on the measured $g_2(r)$ (Appendix A.3.7).

Structural benchmark. Table 2 scores the original and proposed samplers against an MD reference over a fixed list of structural metrics. The prior is trained from scratch on 22,680 atomic environments across 105 structures (generated as described in Appendix A.6). Both methods run the production loop of GLASS with its default settings, and GLASS+annealing includes the modifiers in Sec. 5. The reference set represents the 35 frames that supplied the $g(r)$ target, and the held-out comparison is the overlap column of Table 1. Our method improves ten of the twelve metrics over the unmodified sampler and is better than the reference variability on four. Two results are worth highlighting: the entropy gap is 2.7 times smaller than the original one, the main target of the strategy; and our method still does not beat the benchmark in ring outliers, which are not captured by any of the proposed sampling improvements. Medium-range order topological parameters remain some of the most challenging features to reproduce in amorphous structures, a topic still under active study (Biswas et al., 2024).

Table 2: **Structural benchmark.** a-Si structures ($\rho = 2.3\,\mathrm{g/cm^3}$, $N = 216$ atoms) scored against molecular dynamics over a fixed list of structural metrics, with a prior trained on a-Si at three densities ($\rho = 1.8, 2.3, 2.8\,\mathrm{g/cm^3}$). $n = 35$ (MD) and $n = 100$ (generated) structures. Cells: deviation from MD; the MD column is the reference's own scatter, and bold marks every metric on which GLASS+annealing improves on GLASS. Overall: mean of the twelve rows normalized by the variability in the MD reference (so that MD reads 1.0).

Metric	Unit ($\times 10^{-3}$)	MD	GLASS	GLASS+annealing
g(r)	χ^2	63.6	43.0	**38.6**
ADF	EMD	9.2	14.4	**8.3**
CN dist	JSD	51.2	48.2	**41.9**
CN outliers	frac	6.0	9.3	**7.4**
mean CN	%Err	7.5	4.9	**3.7**
rings	$\lvert\Delta\rvert$/atom	41.8	49.3	**43.4**
ring dist	JSD	48.2	51.0	**50.1**
ring outliers	frac	0.2	**0.7**	1.0
XRD	EMD	6.9	7.2	**6.9**
ND	EMD	2.8	2.9	**2.7**
QUESTS	1−overlap	49.0	**58.0**	60.0
QUESTS div	$\lvert\Delta\mathcal{H}\rvert$ nats	0.0	163.0	**60.0**
Overall ($\times$MD)		1.0	4.6	2.5

Does the shape of η matter? The gate $\eta(t)$ deserves a separate reading, because the function could be, in principle, arbitrary. Profiles that are one at large t and zero at small t leave the marginals of Eq. (7) unchanged for an exact score, so the choice is empirical; we kept the step because it is the simplest and has a single parameter. We swept the gate position t^* against the constant profiles $\eta \equiv 0$, 0.5 and 1 (§A.3.3). Only profiles that inject noise late in the trajectory are worse in $\mathcal{H}$, so the sweep bounds η from one side rather than selecting a value, and $t^* = 0.6$ is the setting the ablation above was run at. A graded or non-monotonic $\eta(t)$ was not explored and may do better. The gate improves D_{g_2} by a factor of 2.4 to 4.2 across the four densities tested against $\eta \equiv 1$.

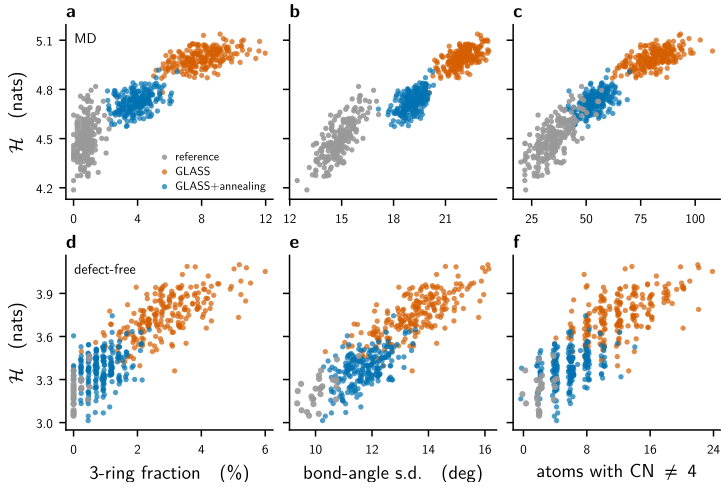


Figure 4: **Two priors, one recipe.** a-Si structures ($\rho = 2.223\,\mathrm{g/cm^3}$, $N = 216$ atoms) against three descriptors $g_2(r)$ does not fix: reference (grey), unmodified GLASS (orange) and GLASS+annealing (blue), $n = 215$ per set. **(a–c)** The prior of Table 1, trained on molecular dynamics at $\rho = 1.5, 2.5, 3.5\,\mathrm{g/cm^3}$. **(d–f)** A prior trained on one defect-free structure (Rosset et al., 2026). Units: $\mathcal{H}$, nats; bond-angle s.d., degrees; coordination, atoms.

Generating high-quality amorphous structures. Despite its use as a prototypical amorphous structure, the goal of generating high-quality a-Si structures that explain their experimental properties has remained unsolved for decades. Given the sensitivity of electronic structure to defects, sampling a-Si structures with low defect concentration is a tremendous challenge, depending on high-quality simulations, long timescales, and even luck. Recently, one such defect-free a-Si struc-

ture was reported by Rosset et al. (2026). We used our annealing sampling scheme to demonstrate that our generative models can sample high-quality a-Si structures comparable with simulations by bootstrapping on the defect-free prior. Figure 4 compares the ablation of Table 1, which was trained on a prior from lower-quality MD, with the same experiment for a prior trained on the single, defect-free Rosset-10 model of Rosset et al. (2026), and against 24 more independent, high-quality structures from the a-Si data set in Rosset et al. (2025) (Appendix A.7).

For generated and reference iso-$g_2(r)$ structures, both rows of Fig. 4 show that the original sampling strategy produces higher-entropy, more defective structures compared to the new method. Our modifications reduce the entropy of the system for the MD datasets (Fig. 4a–c), though they do not fully overlap with the target distribution. However, when using the high-quality prior model, the generated structures fully overlap with the near-defect-free reference (Fig. 4d–f), simultaneously reducing the entropy, the fraction of 3-membered rings, the atoms without tetrahedral bonds, and so on.

Importantly, given our degeneracy-breaking criterion for observables, one can select low-entropy structures, which correlate with overall structural quality in these low-defect regimes (Fig. 4f). Interestingly, generated structures also reach lower levels of entropy compared to the reference, demonstrating the method's ability to sample configurations that remained elusive to conventional sampling methods and posed a fundamental scientific limitation to the field.

7 Discussion and Conclusion

Limitations. The modified sampler improves the quality of the generated structures but does not fully remove the bias. With a prior trained on melt-quench MD simulations using classical force fields, the entropy distribution shifts toward the reference without fully overlapping it. Part of the excess entropy (e.g., ring distributions) is also not simply explained by residual thermal noise, so the sampler may be addressing only one of the sources of variability. Furthermore, the attainable quality is bounded by the training data: the prior inherits the defects of the simulations it learns from. Amorphous datasets remain small because high-quality structures are expensive to generate, and scaling to larger, higher-quality datasets could alleviate this limitation. A related dependence enters through the projected relaxation $\mathcal{Q}$, the only component that invokes an interatomic potential: its quality tracks the accuracy and coverage of the foundation MLIP it uses (here, MACE-MP-0). This is a modular dependence rather than a fundamental restriction, since $\mathcal{Q}$ enters only as a swappable final relaxation whose constrained, truncated form cannot alter the fixed $g_2(r)$. For a system poorly modeled by a uMLIP, $\mathcal{Q}$ can simply be given a system-appropriate or fine-tuned potential, though a systematic evaluation of these effects is outside the scope of this work. Only the numerical accuracy of the final relaxation inherits the potential's reach, so the effect and its correction are general; the diagnosis of an excess configurational entropy applies to any $g_2(r)$-conditioned sampler.

Conclusion. We demonstrated numerous counterexamples for the observable degeneracy problem prevalent in scientific characterization, specifically in chemistry and materials science. Using amorphous structures as a challenging case study, we showed that a score-based sampler conditioned on $g_2(r)$ returns structures that reproduce the target PDF yet carry a systematic excess of configurational entropy. This excess behaves as a residual effective temperature and raises the concentration of defects and the energy of the samples. We introduced four inference-time modifications to suppress this excess without retraining the score network: a noise gate between the reverse SDE and the probability-flow ODE, spectral guidance on the low-k band of $S(k)$, a relaxation projected onto the null space of the Jacobian of $g_2(r)$, and a ladder of partial restarts. On a-Si, the modified sampler reduces the entropy gap to the reference by a factor of 2.7 and almost reaches the correct entropy value, while keeping the conditioned observables within the variability of the reference ensemble. With a prior trained on a single defect-free structure, it generates configurations that overlap with independent defect-free a-Si models, which quench kinetics have long kept out of reach. These results emphasize the nuanced interplay among underdetermined solutions in inverse scientific problems, physical constraints, and the evaluation of generative models needed to advance real-world applications in AI for science.

AI USE STATEMENT

Generative AI tools (Claude) were used in this work to help write code. They were not used to develop the method, to propose or refine the hypotheses, to design the methodology or the experiments, to prepare the data, to formulate or derive any of the mathematical statements, or to draw conclusions from the results; the remaining required-disclosure tasks, namely translation and qualitative or thematic analysis, do not arise in this work. We additionally used these tools to adjust figure scripts, tighten the prose of the manuscript, and assemble and verify the bibliography. All AI-assisted work was reviewed: analysis scripts were executed by the authors and their output checked against the metric values recorded by the scoring pipeline, and every bibliography entry was confirmed against its DOI or arXiv record. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The reference configurations used throughout are classical molecular dynamics trajectories obtained with interatomic potentials taken from the published literature, and each third-party model entering the pipeline is covered by its own public license. No part of the study draws on human subjects, personal data, or laboratory measurements. We identify no dual-use risk in the method or in the structures it produces.

REPRODUCIBILITY STATEMENT

All implementation details and hyperparameters needed to reproduce our results are given in the appendix: the base sampler and the four modifications are specified term by term in Appendix A.3, and the complete procedure, with every numerical setting used in the reported runs, is collected in Appendices A.3.8 and A.7. Section 6 states the system, density, system size and number of structures behind each reported number, together with the reference set and the noise floor of the structural benchmark. The generation and analysis code, and the scripts that regenerate every table and figure in this paper, will be made available anonymously through GitFront.

ACKNOWLEDGMENTS

REFERENCES

Ahmed M. Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 290–306, 2022.

Hans C. Andersen. RATTLE: a "velocity" version of the SHAKE algorithm for molecular dynamics calculations. *Journal of Computational Physics*, 52:24–34, 1983.

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12:313–326, 1982.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, Matthew Avaylon, William J Baldwin, et al. A foundation model for atomistic materials chemistry. *The Journal of chemical physics*, 163(18), 2025.

Filippo Bigi, Paolo Pegolo, Arslan Mazitov, Jonathan Schmidt, and Michele Ceriotti. Pushing the limits of unconstrained machine-learned interatomic potentials. *arXiv:2601.16195*, 2026. PET-OAM-XL 1.0.0 checkpoint: `https://huggingface.co/lab-cosmo/upet`.

Simon J. L. Billinge and Igor Levin. The Problem with Determining Atomic Structure at the Nanoscale. *Science*, 316:561–565, 2007. doi: 10.1126/science.1135080.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018. doi: 10.48550/arXiv.1801.01401.

Parthapratim Biswas, Devilal Dahal, and Stephen R Elliott. Insights into the origin of the first sharp diffraction peak in amorphous silica from an analysis of chemical and radial ordering. *Physical Review B*, 109(10):104207, 2024.

Nathan Brown, Marco Fiscato, Marwin H. S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3): 1096–1108, 2019. doi: 10.1021/acs.jcim.8b00839.

Qiyuan Chen, Ajay Annamareddy, Ying-Fei Li, Dane Morgan, and Bu Wang. Physical-regularized hierarchical generative model for metallic glass structural generation and energy prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 83–91, 2026.

Mouyang Cheng, Denis Blessing, Botao Yu, Gerhard Neumann, Mingda Li, Carles Domingo-Enrich, and Yuanqi Du. ATLAS: A foundation neural sampler for amorphous materials. *arXiv:2607.19198*, 2026.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2206.00941.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.14687.

Massimiliano Comin and Laurent J Lewis. Deep-learning approach to the structure of amorphous silicon. *Physical Review B*, 100(9):094107, 2019.

James Damewood, Jessica Karaguesian, Jaclyn R Lunger, Aik Rui Tan, Mingrou Xie, Jiayu Peng, and Rafael Gómez-Bombarelli. Representations of materials for machine learning. *Annual Review of Materials Research*, 53(1):399–426, 2023.

Pierre-Paul De Breuck, Hai-Chen Wang, Gian-Marco Rignanese, Silvana Botti, and Miguel AL Marques. Generative ai for crystal structures: a review. *npj Computational Materials*, 11(1):370, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2105.05233.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Karen M Ehrhardt, Jack D Sundberg, and Scott C Warren. pyhrmc: Hybrid reverse monte carlo for electron total scattering. *Journal of Computational Chemistry*, 47(3):e70306, 2026.

J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758, August 1982. ISSN 1539-4522. doi: 10.1364/ao.21.002758. URL `http://dx.doi.org/10.1364/ao.21.002758`.

Jonas A Finkler, Yan Lin, Tao Du, Jilin Hu, and Morten M Smedskjaer. Inverse design of amorphous materials with targeted properties. *Advanced Materials*, pp. e22493, 2026.

Natascia L Fragapane and Volker L Deringer. Amorphous materials as a frontier challenge for universal interatomic potentials. *arXiv:2607.11384*, 2026.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research (TMLR)*, 2023. doi: 10.48550/arXiv.2210.02410.

Ryan C. Gettler, Henry D. Koenig, and Matthias J. Young. Iterative Reverse Monte Carlo and Molecular Statics for Improved Atomic Structure Modeling: A Case Study of Zinc Oxide Grown by Atomic Layer Deposition. *Physical Chemistry Chemical Physics*, 23(46):26417–26427, 2021. doi: 10.1039/D1CP03742K.

Jiawei Guo and Daniel Schwalbe-Koda. Generative inversion of spectroscopic data for amorphous structure elucidation. *arXiv:2603.23210*, 2026.

Jean-Pierre Hansen and Ian R. McDonald. *Theory of Simple Liquids: with Applications to Soft Matter*. Academic Press, Oxford, 4 edition, 2013.

R. L. Henderson. A uniqueness theorem for fluid pair correlation functions. *Physics Letters A*, 49 (3):197–198, 1974. doi: 10.1016/0375-9601(74)90847-0.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. doi: 10.48550/arXiv.1706.08500.

William G. Hoover. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Physical Review A*, 31(3):1695–1697, 1985. doi: 10.1103/PhysRevA.31.1695.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, 2024. doi: 10.1109/CVPR52733.2024.00889.

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 (NeurIPS)*, volume 36, pp. 17464–17497, 2023.

Eric Jonas. Deep imitation learning for molecular inverse problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2206.00364.

T. Kumagai, S. Izumi, S. Hara, and S. Sakai. Development of bond-order potentials that can reproduce the elastic constants and melting point of silicon for classical molecular dynamics simulation. *Computational Materials Science*, 39:457–464, 2007. doi: 10.1016/j.commatsci.2006.07.013.

Hyuna Kwon, Tim Hsu, Wenyu Sun, Wonseok Jeong, Fikret Aydin, James Chapman, Xiao Chen, Vincenzo Lordi, Matthew R. Carbone, Deyu Lu, Fei Zhou, and Tuan Anh Pham. Spectroscopy-Guided Discovery of Three-Dimensional Structures of Disordered Materials with Diffusion Models. *Machine Learning: Science and Technology*, 5(4):045037, 2024. doi: 10.1088/2632-2153/ad8c10.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. doi: 10.48550/arXiv.1904.06991.

Yan Lin, Jonas A Finkler, Tao Du, Jilin Hu, and Morten M Smedskjaer. Amshortcut: An inference- and training-efficient inverse design model for amorphous materials. *arXiv:2603.29812*, 2026.

R. L. McGreevy and L. Pusztai. Reverse Monte Carlo simulation: a new technique for the determination of disordered structures. *Molecular Simulation*, 1:359–367, 1988.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2108.01073.

Jianwei Miao, Pambos Charalambous, Janos Kirz, and David Sayre. Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400(6742):342–344, 1999. ISSN 1476-4687. doi: 10.1038/22498. URL `http://dx.doi.org/10.1038/22498`.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning (ICML)*, 2020. doi: 10.48550/arXiv.2002.09797.

Shuichi Nosé. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *The Journal of Chemical Physics*, 81(1):511–519, 1984. doi: 10.1063/1.447334.

Artem R Oganov and Colin W Glass. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of Chemical Physics*, 124(24), 2006.

G. Opletal, T. Petersen, B. O'Malley, I. Snook, D. G. McCulloch, N. A. Marks, and I. Yarovsky. Hybrid approach for generating realistic amorphous carbon structure using metropolis and reverse Monte Carlo. *Molecular Simulation*, 28(10-11):927–938, 2002.

George Opletal, Timothy C Petersen, Brendan O'Malley, IK Snook, DG McCulloch, and Irene Yarovsky. HRMC: Hybrid Reverse Monte Carlo method with silicon and carbon potentials. *Computer Physics Communications*, 178(10):777–787, 2008.

Anup Pandey, Parthapratim Biswas, and D. A. Drabold. Force-enhanced atomic refinement: structural modeling with interatomic forces in a reverse Monte Carlo approach applied to amorphous Si and SiO_2. *Physical Review B*, 92:155205, 2015.

Steve Plimpton. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995. doi: 10.1006/jcph.1995.1039.

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7:836–847, 2025. doi: 10.1038/s42256-025-01055-1. Matbench Discovery, `https://matbench-discovery.materialsproject.org`.

J. B. Rosen. The gradient projection method for nonlinear programming. part II. nonlinear constraints. *Journal of the Society for Industrial and Applied Mathematics*, 9:514–532, 1961.

Louise A. M. Rosset, David A. Drabold, and Volker L. Deringer. Signatures of paracrystallinity in amorphous silicon from machine-learning-driven molecular dynamics. *Nature Communications*, 16:2360, 2025. doi: 10.1038/s41467-025-57406-4. arXiv:2407.16681.

Louise A. M. Rosset, Chinonso Ugwumadu, Stephen R. Elliott, David A. Drabold, and Volker L. Deringer. A defect-free model of amorphous silicon with pristine electronic structure. *arXiv:2606.20500*, 2026.

Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23:327–341, 1977.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. doi: 10.48550/arXiv.1806.00035.

Bernardo P. Schaeffer, Ricardo M. S. Rosa, and Glauco Valle. The effect of stochasticity in score-based diffusion sampling: a KL divergence analysis. *arXiv:2506.11378*, 2025.

Sjors H.W. Scheres. Relion: Implementation of a bayesian approach to cryo-em structure determination. *Journal of Structural Biology*, 180(3):519–530, December 2012. ISSN 1047-8477. doi: 10.1016/j.jsb.2012.09.006. URL `http://dx.doi.org/10.1016/j.jsb.2012.09.006`.

Daniel Schwalbe-Koda, Sebastien Hamel, Babak Sadigh, Fei Zhou, and Vincenzo Lordi. Model-free estimation of completeness, uncertainties, and outliers in atomistic machine learning using information theory. *Nature Communications*, 16:4014, 2025. doi: 10.1038/s41467-025-59232-0.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021a. arXiv:2010.02502.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1907.05600.

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

Alan K. Soper. Empirical Potential Monte Carlo Simulation of Fluid Structure. *Chemical Physics*, 202(2–3):295–306, 1996. doi: 10.1016/0301-0104(95)00357-6.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.48550/arXiv.2306.04675.

Frank H. Stillinger and Salvatore Torquato. Structural degeneracy in pair distance distributions. *Journal of Chemical Physics*, 150:204125, 2019.

Frank H. Stillinger and Thomas A. Weber. Hidden structure in liquids. *Physical Review A*, 25: 978–989, 1982.

Nathan J. Szymanski and Christopher J. Bartel. Establishing Baselines for Generative Discovery of Inorganic Crystals. *Materials Horizons*, 12(19):8000–8011, 2025. doi: 10.1039/D5MH00010F.

Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, January 2005. ISBN 9780898717921. doi: 10.1137/1.9780898717921. URL http://dx.doi.org/10.1137/1.9780898717921.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR)*, 2016. doi: 10.48550/arXiv.1511.01844.

Aidan P. Thompson, H. Metin Aktulga, Richard Berger, Dan S. Bolintineanu, W. Michael Brown, Paul S. Crozier, Pieter J. in 't Veld, Axel Kohlmeyer, Stan G. Moore, Trung Dac Nguyen, Ray Shan, Mark J. Stevens, Julien Tranchida, Christian Trott, and Steven J. Plimpton. LAMMPS – a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, 2022. doi: 10.1016/j.cpc.2021.108171.

Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *International Conference on Learning Representations (ICLR)*, 2023. doi: 10.48550/arXiv.2206.04119.

Haina Wang, Frank H. Stillinger, and Salvatore Torquato. Sensitivity of pair statistics on pair potentials in many-body systems. *The Journal of Chemical Physics*, 153(12):124106, 2020. doi: 10.1063/5.0021475.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Scott M Woodley and Richard Catlow. Crystal structure prediction from first principles. *Nature Materials*, 7(12):937–946, 2008.

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

Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart sampling for improving generative processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.14878.

Kai Yang and Daniel Schwalbe-Koda. A generative diffusion model for amorphous materials. *npj Computational Materials*, 12:29, 2026.

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE(3) diffusion model with application to protein backbone generation. In *International Conference on Machine Learning (ICML)*, 2023. doi: 10.48550/arXiv.2302.02277.

Adrian Xiao Bin Yong, Tianyu Su, and Elif Ertekin. Dismai-Bench: Benchmarking and Designing Generative Models Using Disordered Materials and Interfaces. *Digital Discovery*, 3(9):1889–1909, 2024. doi: 10.1039/D4DD00100A.

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.

A Appendix

A.1 Diversity

This appendix makes explicit, on the population and on individual structures, how diverse the iso-g_2 solutions of Section 4 are. Fig. 5 shows, over the whole hard-gate population, which quantities move at fixed $g_2(r)$; Fig. 6 lays out eight of its structures as an even ladder in $\mathcal{H}$, from the level of the a-Si references to the top of the degenerate pool, each rung the structure whose ring-size and coordination distributions are farthest from everything already selected.

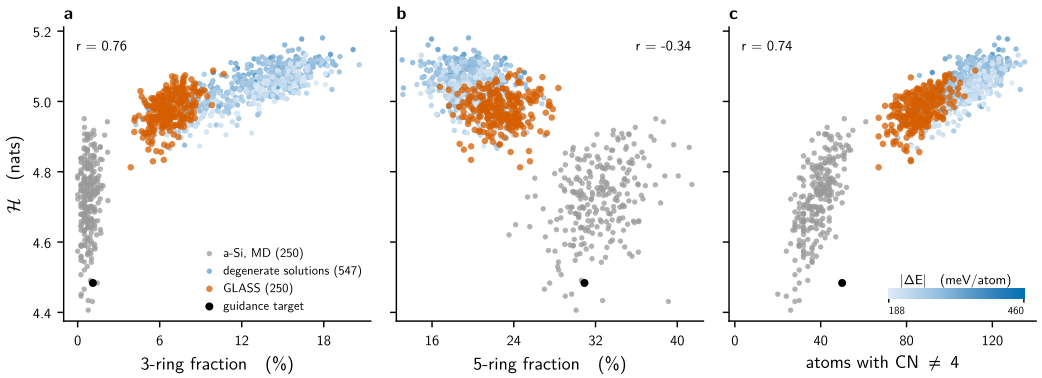


Figure 5: **What deviates at fixed $g_2(r)$.** The 547 hard-gate solutions, each reproducing the guidance target's $g_2(r)$ to a relative mismatch ≤ 0.0545 (the base ensemble's median), against three axes the pair correlation does not fix: **(a)** 3-ring fraction, **(b)** 5-ring fraction, **(c)** atoms with CN $\neq 4$ (bonded neighbors within 2.85 Å). Panels share the vertical axis and quote their Pearson r against $\mathcal{H}$ (0.76, -0.34 and 0.74). Color is $|\Delta E|$ per atom (MACE-MP-0) above the mean of the 250 independent a-Si MD references (grey), which are not themselves iso-$g_2(r)$ to the target. Orange is the 250 GLASS solutions; black is the guidance target, a rescaled structure that sits off the MD cloud. Entropy rises with 3-rings and miscoordination, falls with 5-rings, and every deviation costs energy.

The numbers behind the main-text reading of Fig. 2 are these. Its three drawn structures have mean coordination 4.17, 4.21 and 4.32 and three-ring fractions 1.1, 3.9 and 16.0%; across the population the energy spans 203 to 373 meV/atom above the reference mean, and fewer three-membered rings accompany a lower configurational entropy, with Pearson $r = 0.78$ between the two. Longer ring analyses probe medium- and long-range order instead, and are reported below.

The comparison to make is between the ends of the ladder. Rungs **a** and **b** have almost no three-membered rings and a coordination distribution concentrated at four, as the references have; rungs **g** and **h** are all but taken over by such rings, with 17.7% of the atoms in one against 0.9% in the references, and their coordination spreads to three and five. Between the two ends the pair distribution functions in the second row remain essentially the same, every one within mismatch 0.050 of the target. The fewer the small rings, the closer the coordination sits to four; the more distributed the coordination, the more the network can be rearranged, ring by ring, while $g_2(r)$ stays fixed. Fig. 5 shows the same on the whole population, one axis at a time. Entropy rises with the fraction of three-membered rings and with the fraction of atoms whose coordination is not four (Pearson $r = 0.76$ and 0.74), and falls, more weakly but with a clear trend, with the fraction of five-membered rings ($r = -0.34$); the correlations are quoted as a summary of the tendency, not because the relation is linear. In a defect-free tetrahedral network every atom would have four bonds; the amorphous state admits other coordinations, and every excess of them costs energy, as the color scale shows. On all three axes the molecular-dynamics references are markedly more ordered than either generated set: fewer small rings, fewer miscoordinated atoms and a lower entropy than the diffusion model reaches even before any diversity term is added.

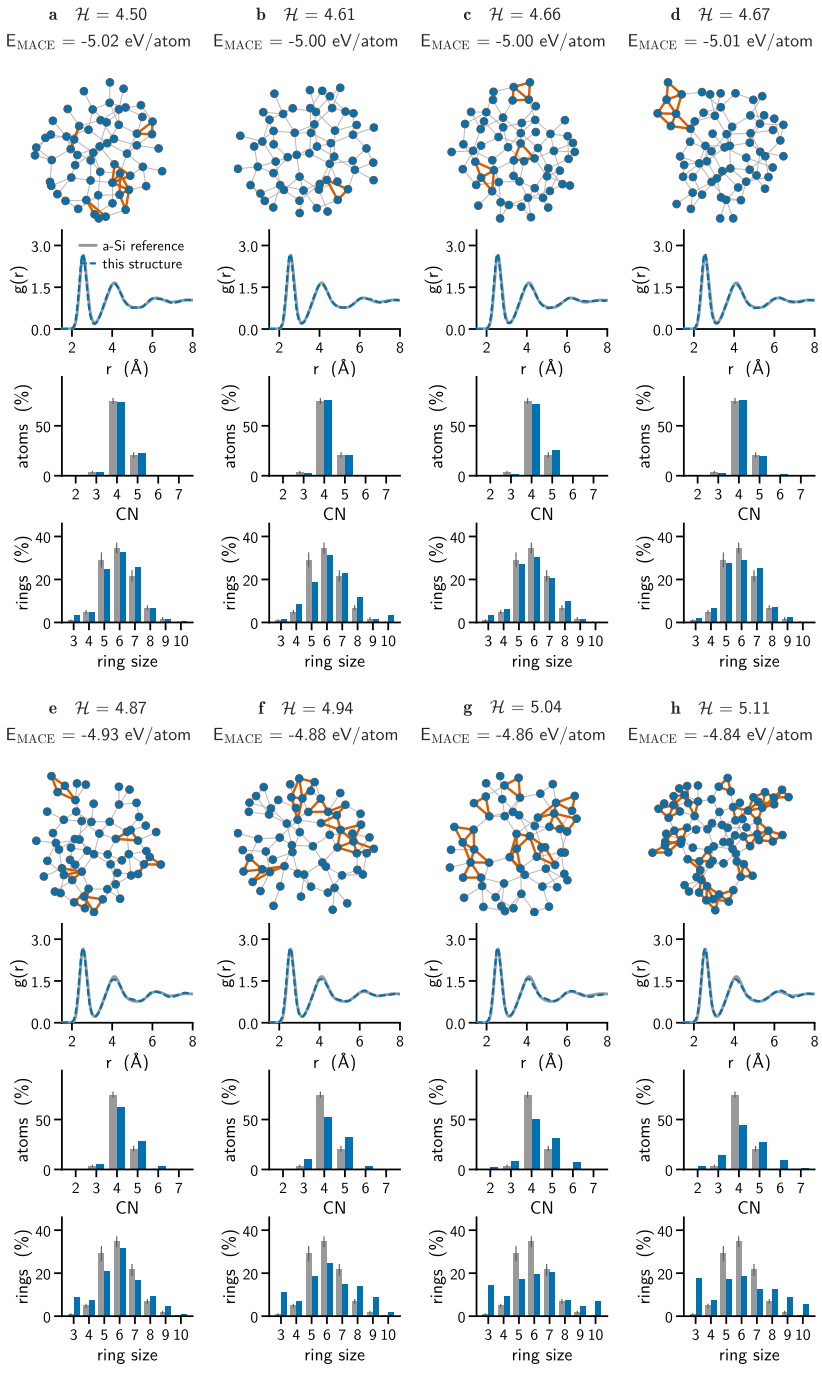


Figure 6: **Eight iso-$g(r)$ structures spanning what the pair correlation leaves free.** An even ladder in $\mathcal{H}$ from 4.50 to 5.11 (the a-Si references sit at 4.55), each rung the structure whose ring-size and coordination distributions are farthest from everything already selected. Rungs **a**–**d** come from the complete-recipe production pool and match the reference in entropy, energy ($E_{\mathrm{MACE}} = -5.02$ to -5.00 eV/atom vs. -5.004) and distributions; **e**–**h** climb to 17.7 % 3-rings (references: 0.9 %). Every structure reproduces the target $g(r)$ to mismatch ≤ 0.050; grey is the reference set (mean $\pm$ 1 s.d.), green edges close a 3-ring.

A.2 TEMPERATURE

The effective-temperature reading of Section 4 was repeated at the second density, $\rho = 2.223\,\mathrm{g/cm^3}$, with the same prior, trained at $\rho = 1.5$, 2.5 and $3.5\,\mathrm{g/cm^3}$, and the same construction: isotropic white displacement noise of amplitude σ added to each of the 250 references, entropy and D_{g_2} read against the reference band and the band of the unmodified sampler. This density lies closer to the training densities of the prior than $\rho = 2.0$ does (0.28 against $0.5\,\mathrm{g/cm^3}$).

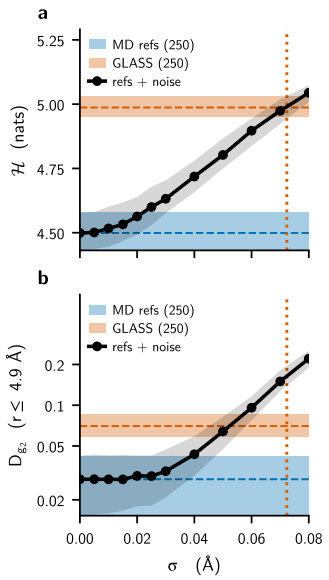


Figure 7: **Effective temperature at $\rho = 2.223\,\mathbf{g\,cm^{-3}}$.** As Fig. 3: **(a)** $\mathcal{H}$ and **(b)** D_{g_2} of the noised references (black median, grey interquartile range) against the reference set (blue) and the unmodified GLASS sampler (orange), median dashed and interquartile range shaded, 250 structures each. The crossing is at $\sigma^* = 0.072\,\text{Å}$.

The correlation survives: the entropy of the noised references rises monotonically with σ and meets the entropy of the sampler at $\sigma^* = 0.072\,\text{Å}$, 3.1% of the bond, for an excess of 0.48 nats. What changes is the order of the two crossings. At $\rho = 2.0$ the noised references reach the sampler's D_{g_2} at the same amplitude at which they reach its entropy; here they reach it earlier, near $\sigma \approx 0.05\,\text{Å}$, where the entropy has covered only about 60% of the excess, and the strict iso-g_2 window, $\sigma \leq 0.013\,\text{Å}$, covers 6% of it against 8% at $\rho = 2.0$. The pair distribution therefore breaks before the entropies meet. This is a limitation of the explanation, stated as such: part of the excess is accounted for as residual high-frequency noise, a temperature modeled as a Gaussian displacement, and part of it is not, since a purely thermal excess would raise the entropy all the way to the sampler's level before $g_2(r)$ gave way, as it does in Fig. 3. Empirically, the farther the sampled density from the prior, the larger the share of the excess that residual noise explains, and the closer, the smaller, though still a substantial part. One would also expect a density nearer the training set to leave the sampler with an entropy nearer the references to begin with; the measured excess does not bear this out, 0.48 against 0.35 nats, so the distance to the prior orders the share explained by noise, not the size of the excess.

Two readings of the excess are compatible with these measurements. Either the model fails to learn some parameter or structural feature at short or long range, or it learns well but learns to model a structure at a higher temperature, of higher entropy, even with $g_2(r)$ as its principal guidance; the two are not separated by any measurement reported here. That temperature is a structural variable a generative model can be conditioned on is shown by ATLAS (Cheng et al., 2026), whose Fig. 2(a) tracks the sharpening of g_2 and the shift of the energy distribution from the liquid to the glass. There, however, temperature is an explicit input and the sampler is matched to a Boltzmann reference at each T, so the comparison places the present excess as a bias of the $g_2(r)$-conditioned sampler rather than as a general property of diffusion models.

A.3 ANNEALING PIPELINE

The system is amorphous silicon with $N = 216$ atoms in a cubic cell, and $\mathcal{H}$ denotes throughout the QUESTS information entropy in nats (bandwidth $h = 0.015$, $k = 32$ neighbors, cutoff $r_c = 5\,\text{Å}$). Numerical effects quoted per term come from the paired ablation reported in the main text.

A.3.1 SETUP: THE OBJECTS OF THE BASE METHOD

(G1) Forward process: $x_t = x_0 + \sigma(t)\varepsilon$, $\sigma(t) = kt$, $k = 0.8\,\text{Å}$; $\gamma^2(t) = \mathrm{d}\sigma^2/\mathrm{d}t = 2k^2t$. **(G2)** Score network $s_\theta \approx \nabla \log p_t$ (weights never modified here). **(G3)** Reverse SDE, Eq. (6) of Song et al. (2021b) for the variance-exploding forward process of its Eqs. (5) and (9) (reverse-time form due to Anderson (1982)), Euler–Maruyama over the grid $t_N = T > t_{N-1} > \cdots > t_0 = 10^{-3}$, $N = 120$, $T = 1$, indexed as in Song et al. (2021a) with x_N the start and x_0 the sample, $\Delta t_i = t_{i-1} - t_i < 0$, $\varepsilon_i \sim \mathcal{N}(0, I)$:

$$x_{i-1} = x_i - \gamma^2(t_i)\big[s_\theta + \ell_{g(r)}\big]\Delta t_i + \gamma(t_i)\sqrt{|\Delta t_i|}\,\varepsilon_i, \qquad i = N, \ldots, 1, \tag{9}$$

Table 3: All modifications, against what the base method already provides. None of the three added terms is a new mechanism: the published sampler Guo & Schwalbe-Koda (2026) exposes a diffraction (XRD/ND) guidance channel through the same posterior-score step, an unconstrained empirical-potential annealing tail (disabled by default), and an iterated re-noise refinement at fixed $t_0 = 0.2$. What changes is the observable, the constraint and the schedule. Within *this* recipe an interatomic potential enters at one point, the forces inside $\mathcal{Q}$; the default configuration of the base method additionally applies potential-based guidance inside the reverse SDE, which this recipe does not use.

term	original	modification	channel closed
score s_θ	$-g^2 s_\theta \Delta t$	coefficient $\times \frac{1+\eta^2}{2}$	$g(r)$ fidelity
noise	$g\sqrt{\lvert\Delta t\rvert}\varepsilon$	$\times\, \eta(t) = \mathbf{1}[t > 0.6]$	
spectral ℓ_S	diffraction channel	retarget to commensurate S_q, k<0.8	low-k bias
relaxation $\mathcal{Q}$	unconstrained annealing tail	projected on $\ker(\partial\hat{g}/\partial x)$	high-k dressing
ladder	iterated refinement, fixed t_0	geometric $t_0 = 0.2, 0.13, 0.085$	(repeats the drain)

where s_θ and $\ell_{g(r)}$ are evaluated at (x_i, t_i). (The implementation evaluates s_θ and g at the destination time t_{i-1} while acting on x_i; the difference is $O(\Delta t)$ and immaterial at $N = 120$.) **(G4)** Constraint term, as published: $\ell_{g(r)} = -\frac{\rho}{\sum\|r\|}\nabla_x \|y^* - \hat{g}(\hat{x})\|^2$, $\rho = 3000$, with $\hat{g}$ the differentiable pair-correlation observable (range $r_c \approx 5$ Å) and $\hat{x}$ defined in (14).

Where the excess entropy sits. The $g(r)$ constraint leaves two spectral bands weakly determined. Below resolution, positional jitter under 0.1 Å barely moves the smeared $\hat{g}$. Beyond range, an observable truncated at r_c cannot constrain $k \lesssim 2\pi/r_c \approx 1.3\,\text{Å}^{-1}$; we act on the four commensurate shells below $0.8\,\text{Å}^{-1}$ (wavelengths 8.6 to 17.1 Å, the longest being the box itself), a conservative subset of that band. The excess of the unmodified sampler was measured in both bands: a high-k vibrational dressing and a $4\times$ excess of long-wavelength fluctuation, with injected noise excluded as its source by direct test. Reading the two bands as a spectral form of the structural degeneracy at fixed pair correlation (Stillinger & Torquato, 2019) is our interpretation, not a result; the same degeneracy is long known in reverse Monte Carlo as the tendency to return the most disordered structure consistent with the data (McGreevy & Pusztai, 1988), which is why hybrid schemes add an energy term. (Henderson's theorem is a different statement, uniqueness of the pair *potential* given $g(r)$, and does not bear on configurational degeneracy.) Each modification below targets one band.

A.3.2 The update, term by term

Original (9), with $\gamma_i = \gamma(t_i)$ and the arguments (x_i, t_i) omitted:

$$x_{i-1} = x_i \underbrace{-\gamma_i^2\, s_\theta\, \Delta t_i}_{\text{learned score}} \underbrace{-\gamma_i^2\, \ell_{g(r)}\, \Delta t_i}_{g(r)\text{ constraint}} + \underbrace{\gamma_i \sqrt{|\Delta t_i|}\, \varepsilon_i}_{\text{white noise}}. \tag{10}$$

Modified, with $\eta_i = \eta(t_i)$, plus two operators outside the loop:

$$x_{i-1} = x_i - \tfrac{1+\eta_i^2}{2}\gamma_i^2 s_\theta \Delta t_i - \tfrac{1+\eta_i^2}{2}\gamma_i^2 \ell_{g(r)} \Delta t_i - \tfrac{1+\eta_i^2}{2}\gamma_i^2\, \ell_S\, \Delta t_i + \eta_i\, \gamma_i \sqrt{|\Delta t_i|}\, \varepsilon_i, \tag{11}$$

$$\tilde{x}_0 = \mathcal{Q}(x_0) \text{ (projected relaxation, §A.3.6)}, \qquad x_N = \tilde{x}_0 + \sigma(T)\varepsilon \text{ (re-noising, §A.3.7)}. \tag{12}$$

A.3.3 Noise-amplitude schedule $\eta(t)$

Take the family of reverse processes with reduced noise, $\mathrm{d}x = \mu\, \mathrm{d}\bar{t} + \eta\gamma\, \mathrm{d}\bar{W}$. Matching its Fokker–Planck equation to the reversed forward one and using $\nabla\cdot(p\nabla \log p) = \Delta p$,

$$-\nabla\cdot(\mu p) + \tfrac{\eta^2\gamma^2}{2}\Delta p = -\tfrac{\gamma^2}{2}\Delta p \implies \boxed{\mu = \tfrac{1+\eta^2}{2}\,\gamma^2\,\nabla \log p_t} \tag{13}$$

For an exact score every $\eta \in [0, 1]$ gives the same marginals, the limits being the reverse SDE of Eq. (6) and the probability-flow ODE of Eq. (13) of Song et al. (2021b), so the coefficient is forced rather than tuned. This is the continuous counterpart of the DDIM dial, Eq. (16) of Song et al.

(2021a), and the family coincides with Eq. (6) of Karras et al. (2022) under $\beta(t) = \eta^2\dot{\sigma}/\sigma$, since $\eta^2\gamma^2 = 2\beta\sigma^2$: their $\beta = \dot{\sigma}/\sigma$ is $\eta = 1$ and $\beta = 0$ is $\eta = 0$.

With an imperfect score $s_\theta = \nabla \log p_t + b$ the equivalence breaks, and it breaks in a direction that is already documented. Section 4 of Karras et al. (2022) attributes the degradation under heavy churn to the practical denoiser's non-conservative field and reports that it does not appear for an analytical denoiser, and Schaeffer et al. (2025) bound the two competing effects in KL divergence: stochasticity corrects the error inherited from the prior but amplifies the score error, so whether it helps depends on the structure of that error and not on its size alone. Here the spurious drift enters with the same coefficient as the score and is halved at $\eta = 0$, and the noise injected below $t = 0.6$, of total variance $\int_0^{0.6} \gamma^2 \mathrm{d}t = k^2(0.6)^2 \approx 0.23\,\text{Å}^2$, is removed at the source. Residual jitter δ costs entropy: white displacement noise added to the references gives $\Delta H = +0.25$ at $\delta = 0.05\,\text{Å}$ and $+0.55$ at $0.1\,\text{Å}$ per component (n=10).

The schedule $\eta(t) = \mathbf{1}[t > 0.6]$ is therefore stochastic while basins are selected and deterministic during refinement, the continuous analogue of restricting stochasticity to $[S_{\mathrm{tmin}}, S_{\mathrm{tmax}}]$ in Algorithm 2 of Karras et al. (2022). Measured, paired on seeds, n=250: $\Delta H = -0.096 \pm 0.004$ at $\rho = 2.0$, against a $g(r)$ mismatch $4\times$ better, $0.059 \to 0.014$. At three other densities the entropy term is within 0.015 of zero while the $g(r)$ term stays better by a factor of 2.4 to 3.1. The entropy contribution is not something to promise on a new system; the $g(r)$ contribution reproduces everywhere, and that is what $\eta(t)$ earns its place on. A six-arm sweep at $\rho = 2.300$ (n=10 each, guidance weight 300) varies the gate position ($t^* = 0.3$, 0.6, 0.9) and the constant amplitude ($\eta \equiv 0$, 0.5, 1). What orders the six is the noise each injects below $t = 0.6$, the part the rest of the trajectory no longer removes. The two arms that inject most there, $t^* = 0.3$ and $\eta \equiv 1$, give $\mathcal{H} = 4.95$ and 4.91; the other four give 4.77, 4.77, 4.81 and 4.81, a range under the protocol's ± 0.05 run-to-run reproducibility, so the sweep resolves one step and not a curve. Miscoordination follows the same ordering, from 46 to 55 atoms of 216. Two things it does not settle. First, $\eta \equiv 0$ is not a noiseless sampler: (8) re-noises at every rung whatever η is, adding $0.04\,\text{Å}^2$ to every arm against the gate's own 0.45, so the deterministic limit of the sampler was never reached. Second, with three gate positions at n=10 the sweep can exclude late injection but cannot locate an optimum.

A.3.4 Evaluation point of all constraints

For the Gaussian kernel (G1), differentiating $p_t = \int p(x_0)\,\mathcal{N}(x_t; x_0, \sigma^2 I)\mathrm{d}x_0$ gives Tweedie's formula, Eq. (1.4) of Efron (2011), with the score network as plug-in estimate as in Eq. (10) of Chung et al. (2023):

$$\hat{x} = \mathbb{E}[x_0 \mid x_t] = x_t + \sigma^2 \nabla \log p_t \approx x_t + \sigma^2 s_\theta, \qquad \nabla_x \log p(y \mid x_t) \approx \nabla_x \log p\big(y \mid \hat{x}\big). \quad (14)$$

Their Eq. (10) is the variance-preserving specialization; the variance-exploding form used here follows from the same exponential-family identity, their Lemma 1. The approximation on the right is Eqs. (11) and (15) of Chung et al. (2023), with the error bounded by their Theorem 1. Every constraint, including the original (G4), is such a posterior-score term evaluated on the clean estimate $\hat{x}$, which is why the ingredients are additive and independent; the decomposition of the conditional score into learned score plus constraint gradients is Eq. (14) of Song et al. (2021b), equivalently Eqs. (12)–(13) of Dhariwal & Nichol (2021).

A.3.5 Spectral term ℓ_S

For the 32 wavevectors ($16 \pm$ pairs, averaged over the half set) $q = \frac{2\pi}{L} n$, $|q| < 0.8\,\text{Å}^{-1}$: $\rho_q = \sum_j e^{iq\cdot x_j} = A_q + iB_q$, $S_q = |\rho_q|^2/N$. The exact relation $S(k) = 1 + \rho_0 \int_0^\infty (g(r) - 1)\frac{\sin kr}{kr} 4\pi r^2 \mathrm{d}r$ requires g at all r; truncation at r_c leaves this band weakly determined, and there the unmodified model places $S_{\mathrm{band}} = 0.120$ against 0.032 in the references. Following (14) with a Gaussian likelihood centered on the reference values (per-wavevector mean over the reference set),

$$\ell_S = -\frac{\rho_S}{\|S - S_{\mathrm{ref}}\|} \nabla_{\hat{x}} \frac{1}{M} \sum_q \big(S_q(\hat{x}) - S_{\mathrm{ref},q}\big)^2, \qquad \frac{\partial S_q}{\partial x_j} = \frac{2}{N}\big[B_q \cos(q\cdot x_j) - A_q \sin(q\cdot x_j)\big] q, \quad (15)$$

with $\rho_S = 3\times 10^4$. This is the posterior-score step of Chung et al. (2023), line 7 of their Algorithm 1, at the residual-normalized step size $\zeta = \rho/\|y - \mathcal{A}(\hat{x})\|$ they recommend in footnote 5

and Appendix D.1, the same rule (G4) already uses. It is a set-point force, damping an excess and refilling a deficit, so no systematic overshoot was observed.

ℓ_S is therefore not a new guidance mechanism, since the base model is built to invert diffraction data (Guo & Schwalbe-Koda, 2026); what changes is the observable. Its released diffraction channel evaluates a Debye sum on unwrapped pair vectors with no minimum image, so below $k \approx 1\,\text{Å}^{-1}$ it reports the box shape rather than the density fluctuation: across our ten references its configurational contrast in this band is 7%, against 97% for the commensurate S_q (relative spread across the ten references, per wavevector). Substituting $\rho_q = \sum_j e^{iq\cdot x_j}$ at the box-commensurate wavevectors restores the correct low-k observable for a periodic cell.

A.3.6 Constrained relaxation $\mathcal{Q}$, the only use of a potential

The sampler's endpoint sits away from the local minima of the physical energy, and that residual displacement reads as entropy. The correction minimizes an interatomic energy E without moving $g(r)$, by Rosen's gradient projection (Rosen, 1961) with Tikhonov damping,

$$\min_x E(x) \;\text{ s.t. }\; \hat{g}(x) = y^* : \qquad x^{(k+1)} = x^{(k)} + \alpha_k P d^{(k)}, \quad d^{(k)} = -\nabla E\big(x^{(k)}\big), \\ P = I - J^{\mathsf{T}}(JJ^{\mathsf{T}} + \epsilon I)^{-1} J, \tag{16}$$

for $k = 0, 1, \ldots$, $J = \partial\hat{g}/\partial x \in \mathbb{R}^{40\times 648}$ (refreshed every 10 steps), $\epsilon = 10^{-3}$, step ≤ 0.01 Å/atom, at most 150 iterations, with an early stop at $\|Pd\|_\infty < 0.05$ eV/Å that in practice never fires: measured on 32 structures, every call spends all 150 iterations, so the relaxation is truncated rather than converged. Two properties hold by construction: $\mathrm{d}\hat{g} = J(\alpha_k P d^{(k)}) = O(\epsilon)$, so the constraint is invariant to first order, and $\Delta E \approx -\alpha_k d^{(k)\mathsf{T}} P d^{(k)} \leq 0$ with $P \succeq 0$, so the step only descends. The step bound keeps the $g(r)$ drift at 0.0071 ± 0.0004, measured on the ten references, against a spread of 0.072 between references on the same observable. With J computed once instead, the drift is 0.045 ± 0.004, six times larger, so the refresh is necessary.

The closest antecedents are in structure refinement rather than in the diffusion literature. FEAR (Pandey et al., 2015) alternates partial reverse Monte Carlo fitting of the structure factor with partial energy relaxation, demonstrated on 216-atom a-Si, and hybrid reverse Monte Carlo (Opletal et al., 2002) combines an empirical potential with a structural constraint as a penalty; $\mathcal{Q}$ enforces the constraint by tangent-space projection instead, so $\hat{g}$ is invariant by construction rather than restored after each step, and the target it descends to is the inherent structure of Stillinger & Weber (1982). The manifold-constrained gradient of Chung et al. (2022), their Eq. (14), likewise confines a gradient to a distinguished subspace, but to a different one and in a different sense: their Theorem 1 places the measurement gradient in the tangent space of the learned data manifold at $\hat{x}$, spanned by the transposed Jacobian of the Tweedie map, under local linearity and an exactly optimal score, whereas P projects onto the null space of $\partial\hat{g}/\partial x$, removing the component that would move a chosen physical observable, and is exact with no assumption on the score. The analogy with the constrained dynamics of SHAKE and RATTLE (Ryckaert et al., 1977; Andersen, 1983) is structural only.

The relaxation state is part of the comparison. Applying $\mathcal{Q}$ to the references themselves lowers their entropy by 0.090 ± 0.005 (n=10). The generated output is relaxed and the references are not: they are 300 K molecular-dynamics snapshots, displaced from their own inherent structures by construction. Any entropy comparison that does not put both sides in the same relaxation state credits the method with part of that difference; the energy comparison of the main text is not exposed to this. E is the one place a potential enters this recipe, and it is a choice: we use MACE-MP-0, and an empirical Tersoff potential run as a dependency-free alternative leaves its own fingerprint in the coordination (CN 4.31 against 4.23 under MACE and 4.18 in the references) while moving $\mathcal{H}$ by less than the run-to-run scatter.

A.3.7 Ladder of partial passes

Because (G1) is Gaussian, $p_T = p_0 * \mathcal{N}(0, \sigma^2(T)I)$, so the re-noising jump $x_N = \tilde{x}_0 + \sigma(T)\varepsilon$, Algorithm 1 of SDEdit (Meng et al., 2022), gives a sample of $\hat{p} * \mathcal{N}(0, \sigma^2(T)I)$, which matches p_T to the extent that the previous pass's output $\hat{p}$ matches p_0. Their Proposition 1 bounds how far one pass moves the state away from its input as a function of $\sigma(T)$. The pass therefore restarts near, not inside,

the trained distribution; what keeps that gap from compounding is that the amplitudes decrease geometrically ($\sigma_{\mathrm{reheat}} = 0.16, 0.10, 0.068$ Å), so re-noising perturbs less than the following relaxation removes. Measured after each rung (n=8, MACE-MP-0 in $\mathcal{Q}$): $4.680 \to 4.600 \to 4.568 \to 4.584$, saturating after two rungs, the last one inside the error. Without the ladder the procedure is one sampler call, $\mathcal{H} = 4.680 \pm 0.024$.

The jump-and-descend cycle is the Restart iteration, Eqs. (3)–(4) and Algorithm 2 of Xu et al. (2023), and placing several such intervals at decreasing noise is their §4.3; their Theorem 2 splits the error into a contracted part carrying a factor $(1-\lambda)^K$ in the number of cycles K and an added sampling error growing with K, so the bound has an optimum at finite K, consistent with our saturation after about two cycles. The base method already provides this loop at a fixed $T = 0.2$, identical to the first rung here; what changes is the geometric decrease of the amplitude, as in the annealed schedule of Song & Ermon (2019) (their §4.1 states the geometric condition on $\{\sigma_i\}$ and line 3 of their Algorithm 1 the matching step size $\alpha_i = \varepsilon\,\sigma_i^2/\sigma_L^2$), and the use of $\mathcal{Q}$ as the per-cycle relaxation.

Is the ladder just more relaxation? $\mathcal{Q}$ runs at the end of every rung, so the complete recipe spends four relaxations against the single-pass arm's one, and the ablation row for the ladder carries both effects. Continuing to relax the $\rho = 2.0$ single-pass outputs, with no re-noising and no denoising, separates them (n=24). First, the relaxation never converges: not one of the 24 reached the 0.05 eV/Å stop even at 1200 iterations, eight times the recipe's setting, so the 150-iteration cap is a truncation and the early stop is decorative. Second, at the recipe's own budget of 600 iterations, grinding the same basin reaches $\mathcal{H} = 4.656$ while the ladder reaches 4.585: of the -0.097 credited to the ladder, -0.035 is the extra relaxation and -0.062 is the cycle. Even at 1200 iterations pure relaxation stops at 4.642. Third, and this is what the cycle is for, long relaxation damages the constraint monotonically: the $g(r)$ drift reaches 0.017 at 1200 iterations, as large as the recipe's entire mismatch. The ladder improves $g(r)$ over the single-pass arm instead, $0.020 \to 0.016$. Each rung's guided denoise repays the drift its relaxation created, and the re-noising jump moves the state somewhere the next relaxation has new work to do.

The rungs run with $\eta \equiv 0$: they start at $T = 0.2, 0.13, 0.085$, all below the gate, so the only stochasticity in the ladder is the re-noising jump itself. That is deliberate. Injecting noise through the rungs instead lowers $\mathcal{H}$ by 0.034 ± 0.017 but nearly triples the $g(r)$ mismatch, $0.0136 \to 0.0383$ (n=48 paired, $\rho = 2.223$): the entropy is bought by leaving the constraint surface, which is the one thing this recipe may not do.

A.3.8 The complete procedure

The four modifications compose into a single loop. A sample starts from $x \sim \mathcal{N}(0, \sigma^2(1)I)$ on the periodic cell and is carried through four passes of the reverse SDE at decreasing initial times, $T = 1,\ 0.2,\ 0.13,\ 0.085$, discretized with 120, 24, 16 and 10 Euler–Maruyama steps down to $t = 10^{-3}$. Every pass after the first is preceded by the exact re-noising of §A.3.7, whose amplitudes $\sigma(T) = 0.16,\ 0.10,\ 0.068$ Å decrease geometrically, and every pass is followed by the projected relaxation $\mathcal{Q}$ of §A.3.6. Within a pass, each step evaluates the score once, forms the clean estimate $\hat{x}$ of (14), and adds the two guidance gradients on that estimate: $\ell_{g(r)}$ of (G4) with $\rho = 3000$ against a 40-bin pair correlation of range $r_c = 5$ Å, and ℓ_S of (15) with $\rho_S = 3 \times 10^4$ on the 16 box-commensurate wavevectors below $0.8\,\text{Å}^{-1}$. The sum enters the drift with the coefficient $c_\eta = (1+\eta^2)/2$ of (13) and the injected noise with amplitude $\eta(t) = \mathbf{1}[t > 0.6]$, so the first pass is stochastic above the gate and deterministic below it, while all three ladder rungs, which begin below the gate, run at $\eta \equiv 0$. The relaxation refreshes its Jacobian every 10 iterations, damps the normal equations with $\epsilon = 10^{-3}$, caps the displacement at 0.01 Å per atom, removes the center-of-mass component of each step, and stops after 150 iterations. The energy E is MACE-MP-0 (small); it is the only place an interatomic potential enters the procedure, and the score network is used as trained and never modified.

A.4 Ablation at $\rho = 2.0$

Table 4 repeats the ablation of Table 1 at $\rho = 2.0$ g/cm^3, the density of the ten reference structures shipped with GLASS, in a cubic cell of $L = 17.1416$ Å, with the same prior, the same free pa-

Table 4: Ablation at $\rho = 2.0\,\mathrm{g/cm^3}$: $N = 216$ silicon atoms in a cubic cell of $L = 17.1416\,\text{Å}$, $n = 250$ structures per generated row; the last generated row is the 25 lowest-H structures of the complete-recipe arm. H_{QUESTS} is in nat (k=32, r_c=5 Å, h=0.015). D_{g_2} **and** $S(k{<}0.8)$ **are reported in units of** 10^{-3}: both are dimensionless, so the column entries are the values $\times 10^3$; here D_{g_2} is taken to the first reference structure, the target the generation was conditioned on, which is why the reference row is not zero. E_{PET} is in eV/atom (PET-OAM-XL 1.0.0 (Bigi et al., 2026; Riebesell et al., 2025)), rounded to two decimals. Overlap is a fraction, scored against the 10 reference structures shipped with GLASS, $\rho = 2.5$ glasses rescaled to 2.0; the reference row is its leave-one-out value. The score prior is the silicon model shipped with GLASS, trained at $\rho = 1.5$, 2.5 and 3.5 g/cm^3, the same prior as in Table 1; $\rho = 2.0$ sits 0.5 g/cm^3 from the nearest training density, against 0.28 for $\rho = 2.223$. Free parameters are those of Table 1, unchanged.

configuration	$\mathcal{H}$	D_{g_2}	$S(k{<}0.8)$	E	overlap
GLASS, no recipe	4.98 ± 0.05	59	120	-4.84	0.337
+ $\eta(t)$	4.88 ± 0.05	14	81	-4.80	0.443
+ spectral term ℓ_S	4.82 ± 0.05	12	37	-4.81	0.510
+ projected relaxation $\mathcal{Q}$	4.69 ± 0.05	20	25	-4.89	0.618
+ restart ladder (complete recipe)	4.59 ± 0.05	15	34	-4.94	0.667
best 10% of the recipe (25 of 250)	4.50 ± 0.01	14	33	-4.95	0.696
reference set	4.55 ± 0.06	65	32	-4.94	0.730

rameters and the same columns. Two things differ from the main-text density. The reference pool is the ten shipped structures, $\rho = 2.5$ glasses rescaled to 2.0, so D_{g_2} is taken to the first of them, the actual guidance target, which is why the reference row is not zero, and the overlap threshold is calibrated on that pool. And the gate $\eta(t)$ carries an entropy effect here, -0.10 nats paired on seeds, that it does not carry at $\rho = 2.223$, where it stays within the run-to-run scatter, nor at the two other densities tested, where it is within 0.015 nats of zero. The difference has a plausible origin in the prior: the score model was trained at $\rho = 1.5$, 2.5 and 3.5 g/cm^3, so $\rho = 2.0$ lies 0.5 g/cm^3 from the nearest training density while $\rho = 2.223$ lies 0.28 from it, and the farther the sampled density, the more residual noise the sampler leaves for a term that acts directly on it. The one column the gate moves the wrong way is the energy, up by 0.04 to 0.07 eV/atom in the η row of both tables, which the relaxation then removes. The last generated row is the 25 lowest-$\mathcal{H}$ structures of the complete-recipe arm, a selected tail whose dispersion is not that of a sample. Every other term contributes as at $\rho = 2.223$: the spectral term brings $S(k{<}0.8)$ to the reference level, the relaxation and the ladder drain the entropy, and the complete recipe ends within one reference standard deviation of the reference mean.

A.5 Entropy distributions of the ablation arms

Tables 1 and 4 report each arm as a mean and a standard deviation; Fig. 8 shows the 250 structures behind every row as a histogram against the 250 molecular-dynamics references at the same density, so the overlap with the reference pool is read directly. The arms are unimodal and of similar width at both densities (0.045–0.062 nats), narrower than the references (0.107 and 0.115 nats); each ingredient shifts the whole distribution as a block, none broadens or splits it. The densities differ in where the block ends. At $\rho = 2.223$ the stock sampler and the gated arm coincide, the spectral term and the relaxation move the block by 0.05 and 0.08 nats and the ladder by a further 0.14; the complete recipe covers the upper half of the held-out pool, 94% of its structures inside the pool's range but only 5% within one standard deviation of its mean. At $\rho = 2.0$ the gate alone moves the block by 0.10 nats, and the complete recipe ends at 4.59, on the ten rescaled references it was guided to (4.55 ± 0.06, Table 4; 63% within one standard deviation) and 0.13 nats below the grey independent MD set, which it covers entirely. The unmodified sampler overlaps neither pool.

A.6 Molecular dynamics generation of the benchmark training set

The a-Si reference structures were produced by classical melt–quench molecular dynamics in `LAMMPS` (Plimpton, 1995; Thompson et al., 2022) (`units metal`, periodic boundaries), with silicon described by the Kumagai modified Tersoff potential (`pair_style tersoff/mod`) (Kumagai et al., 2007). Each 216-atom cubic cell was energy-minimized and then evolved entirely

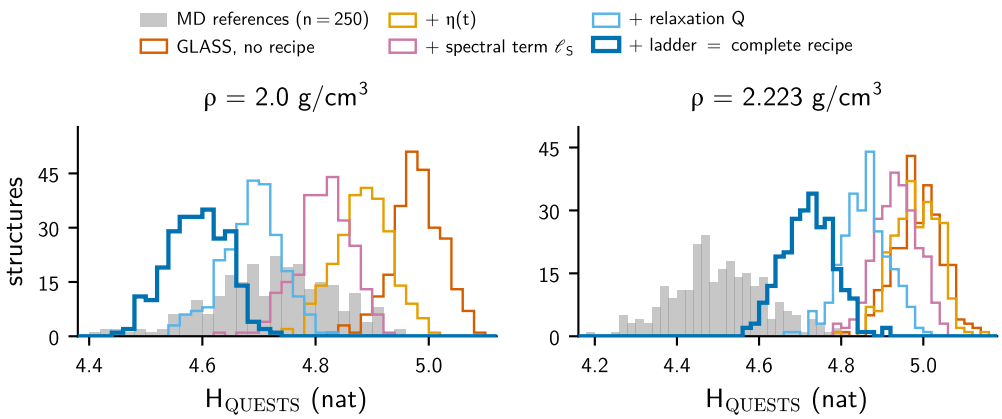


Figure 8: $\mathcal{H}$ **distributions of the ablation arms.** Histograms of the 250 structures in each arm of Tables 1 and 4, at $\rho = 2.0$ (left) and $\rho = 2.223$ g/cm^3 (right), in bins of 0.02 nats; the same five arms in both panels, from the unmodified sampler (orange) to the complete recipe (blue, thick). Grey is the 250 Stillinger–Weber melt-quench references at each density: the held-out pool of Table 1 on the right, and on the left the independent MD set that was never used as a target. QUESTS with k=32, r_c=5 Å, h=0.015; $\mathcal{H}$ is not comparable between the two panels.

in the canonical (NVT) ensemble with a Nosé–Hoover thermostat (Nosé, 1984; Hoover, 1985) ($T_{\mathrm{damp}} = 100\,\Delta t$) at a 0.5 fs timestep: melted at 3000 K for 50 ps, cooled linearly to 300 K at 5 K/ps, and relaxed for a further 50 ps at 300 K. Because the dynamics run at fixed volume, the density is set by the cell; 35 independent structures were produced at each of three densities ($\rho = 1.8$, 2.3, and 2.8 g/cm^3; 105 in total). The 1 and 4 are the Stillinger–Weber analogous.

A.7 The defect-free prior

The second row of Fig. 4 replaces the melt–quench prior by one trained on a single structure. A few settings differ from the canonical ones of §A.3.8; they are collected here.

Prior. The training set is the single Rosset-10 structure of Rosset et al. (2026), 216 atoms at $\rho = 2.2230$ g/cm^3. Architecture and optimizer are those of the melt–quench prior: 5 convolutions, width 200, graph cutoff 5 Å, $k = 0.8$, EMA decay 0.9999, learning rate 10^{-3}. Only the schedule is rescaled to the single structure: duplication factor 400, 3600 epochs, batch size 4, 0.9 train–validation split.

Reference set. The 25 references of Fig. 4d–f are the Rosset-10 model together with 24 independent structures of the a-Si data set of Rosset et al. (2025) carrying at most two miscoordinated atoms, rescaled to this density and relaxed at fixed cell under MACE-MP-0; neighbors are counted within 2.85 Å. The guidance target is a different object: the $g_2(r)$ and the $S(k{<}0.8)$ of the Rosset-10 structure alone, so the other 24 enter the comparison and never the generation.

Generation. The unmodified arm ($n = 215$) is unconditional, with no guidance term: $t \in [0.005, 0.95]$ over 1024 steps, three restarts, and the Langevin corrector of GLASS ($n_{\mathrm{corr}} = 2$, step 0.5, active below half of $t_{\max}$). Starting configurations come from Poisson-disk packing at a minimum separation of 2 Å.

The recipe arm ($n = 215$) runs the same integrator with the four modifications of Section 5 and four settings changed from §A.3.8. The pair observable is coarsened from 40 bins to 12, with the Gaussian smearing widened from 0.15 to 0.45 Å; the guidance strength is $\rho = 100$ rather than 3000; the first pass uses 1024 steps rather than 120 and each rung of the ladder is four times longer, 96, 64 and 40 steps at $T = 0.2$, 0.13 and 0.085; and the projected relaxation is extended from 150 iterations to 600 (150 of the 215 structures) or 900 (the remaining 65), refreshing its Jacobian every 5 iterations instead of 10 and capping the displacement at 0.02 Å per atom instead of 0.01. The spectral strength $\rho_S = 3 \times 10^4$ and the gate $\eta(t) = \mathbf{1}[t > 0.6]$ are those of §A.3.8.

These departures are not cosmetic. Holding the pair observable at its canonical 40 bins leaves the recipe at $\mathcal{H} = 3.63 \pm 0.06$ ($n = 6$), barely below the 3.72 ± 0.02 of the unconditional sampler

($n = 100$), whereas coarsening to 12 bins reaches 3.40 ± 0.05; and $\rho = 3000$ is calibrated for the multi-density prior and does not transfer to one trained on a single structure.