WHAT DEFINES A HIGH-QUALITY GENERATED MATERIAL? STRESS-TESTING GENERATIVE MODELS ON AMORPHOUS MATTER

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models for materials excel at producing crystals and molecules with targeted properties. However, most technologically relevant materials are metastable, defected, or disordered systems that challenge the definitions of uniqueness, stability, and novelty on which typical metrics and benchmarks rest. Here, we introduce AmorphBench, a protocol that evaluates how well generative models sample inorganic amorphous structures according to physical principles. We define 6 metric families and 21 quality metrics covering short- and medium-range order, topology, spectroscopic observables, mechanics, energetics, and phase-space coverage. By scoring properties and experimental observables rather than structural agreement alone, we expose multiple limitations of the learned distributions, where even single outlying atoms lead to outright unphysical behavior. Applying the protocol to 8 generative and structure-fitting methods across 5 systems and 14 density and composition conditions, we show that generative model quality for materials is far from saturated. Scaling tests indicate that performance improves with training data, a promising signal for amorphous systems, where reference datasets are expensive to obtain. We release the benchmark and its reference datasets to guide the development of generative models that meet the bars for data efficiency and sample quality required by real-world applications.

1 INTRODUCTION

Generative models have become a key tool for the inverse design of materials and molecules, proposing structures conditioned on composition, engineering constraints, and target properties (Zunger, 2018; Gómez-Bombarelli et al., 2018; Sanchez-Lengeling & Aspuru-Guzik, 2018; Zeni et al., 2025). The rapid progress of generative models has been possible due to well-defined benchmarks assessing validity, uniqueness, and novelty of candidate samples (Xie et al., 2022; Brown et al., 2019). For instance, in inorganic crystals, validity is defined via thermodynamic quantities, while compact periodic representations allow straightforward evaluation of sample uniqueness and novelty relative to known databases (Szymanski & Bartel, 2025). In many cases, machine learning interatomic potentials (MLIPs) are also employed to rapidly estimate properties of samples, providing fast and quantitative metrics that guide the development of improved models (De Breuck et al., 2025).

Despite this progress, many technologically relevant materials—including liquids, battery materials, metallic glasses, and polymers—are not periodic crystals, but amorphous structures that are overlooked by most atomistic generative models (Yong et al., 2024; Liu et al., 2025). Interpreting and sampling configurations that lack long-range order has largely relied on human experts (Elliott, 1990; Billinge & Levin, 2007). Even proposing reasonable configurations with conventional simulations is a substantial scientific contribution that requires expert interpretation of the results (McGreevy, 2001; Cliffe et al., 2010; Deringer et al., 2021). Automating this process with generative models would bring them closer to real-world materials applications while addressing fundamental physics challenges. However, evaluating amorphous samples is harder than evaluating crystals or molecules for several reasons. First, an amorphous material has no unique ground-truth structure: each configuration is one sample from an ensemble that is best modeled at scales of $\mathcal{O}(10^3$–$10^6)$ atoms and long simulations (Morrow et al., 2024; Finkler et al., 2026). Second, distinct structures can reproduce the same experimental measurements, so matching them does not guarantee physical

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.

validity; a structure can satisfy multiple conditions while still exhibiting unphysical energies (Wang et al., 2020; Gettler et al., 2021). Third, there is no straightforward way to determine whether a generated amorphous configuration is truly "novel" or merely memorized from the training set (Yang & Schwalbe-Koda, 2026). Fourth, MLIPs still struggle to make accurate predictions of amorphous materials due to the diversity of states (Bertani & Pedone, 2026; Benassi et al., 2026). Finally, no single quality indicator captures structure, topology, energetics, and ensemble diversity at once. Together, these challenges have stymied progress in combining generative modeling with amorphous matter, aside from a few recent success cases (Kwon et al., 2024; Yang & Schwalbe-Koda, 2026; Guo & Schwalbe-Koda, 2026; Finkler et al., 2026; Lin et al., 2026).

Here, we introduce AmorphBench, a benchmark for generative models of amorphous materials released with a Python package, curated datasets, and an interactive leaderboard (Fig. 1). AmorphBench defines a set of 21 figures-of-merit rooted in the physics of disordered structures and grouped into six families: short-range order, topology, diffraction, mechanics, energetics, and phase-space coverage. Using these quality metrics, we establish the underlying variability of structures sampled from classical simulations across five systems and multiple processing conditions, and use this baseline to test multiple generative models, sampling strategies, and traditional algorithms. AmorphBench reveals multiple failure modes that are not captured by existing evaluation metrics. First, structural metrics alone suggest that the problem is nearly solved, but energy, mechanical properties, and topology reveal that models still struggle to sample the exact distributions of amorphous configurations. Second, sample quality is inconsistent, as models with competitive median scores often produce overlapping atoms or other unphysical features. Finally, model performance does not transfer across bonding types: a model that generates valid metallic glasses can fail on covalent networks and vice versa. AmorphBench provides quantitative assessments of physical admissibility for atomistic generative models aimed at capturing distributions, and, at the example of other benchmarks for materials science (Chiang et al., 2026; Malik et al., 2026; Riebesell et al., 2025), can inform the development of transferable generative models for materials science.

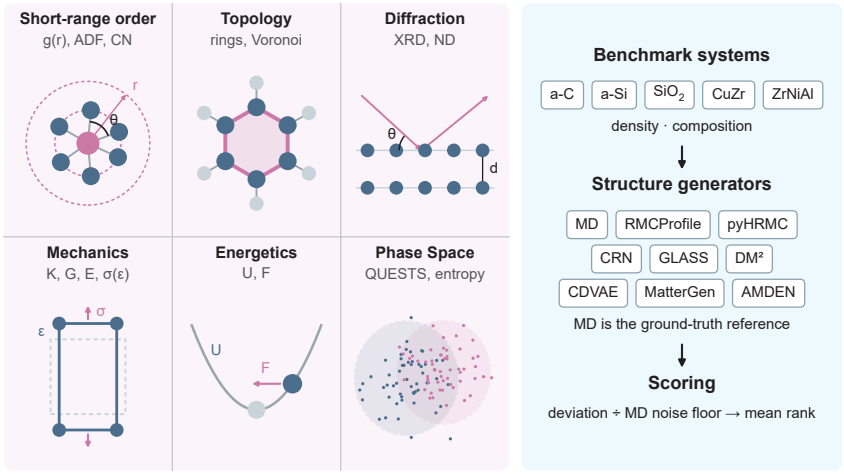


Figure 1: **Overview of the amorphous benchmark framework.** Eight methods are evaluated across six metric families spanning 14 density/composition conditions over five amorphous systems.

2 AmorphBench: Datasets, Metrics, Baselines, and Models

2.1 Systems and Datasets

Given that an amorphous material has no unique ground-truth structure, a generative model can only be evaluated against an ensemble of reference configurations whose variability is itself known. We therefore built AmorphBench on five systems whose disordered structure is well characterized by classical molecular dynamics (MD) (Frenkel & Smit, 2002), so that the spread between independent MD samples defines a trusted baseline for each generated structure: amorphous carbon (a-C), amorphous silicon (a-Si), amorphous silica (a-SiO_2), and the metallic glasses CuZr and Zr-Ni-Al.

Simulation protocols, potentials, and per-system constants for each system are given in Appendix B. The systems were also chosen to span the bonding types across which model performance fails to transfer. a-C mixes sp^2 and sp^3 bonding in a ratio set by density, so a model must capture a density-driven change in hybridization and ring statistics (McCulloch et al., 2000; Dernov et al., 2025). a-Si is the prototypical continuous random network (CRN) of tetrahedrally coordinated atoms, where deviations from fourfold coordination provide a sensitive measure of structural quality (Bernstein et al., 2019; Morrow et al., 2024), and a-SiO_2 is a chemically ordered, two-species network former (Liu et al., 2022). CuZr and Zr–Ni–Al are densely packed metallic glasses characterized by composition-dependent short- and medium-range order, including icosahedral and related coordination motifs (Mattern et al., 2009; Ding et al., 2014). The ternary Zr–Ni–Al system additionally exhibits pronounced chemical short-range order arising from preferential interactions between its constituent species (Georgarakis et al., 2010). Together, these systems cover multiple applications ranging from protective coatings and thin-film electronics to optical glasses and high-strength structural alloys. To test whether models follow changes in processing rather than memorizing a single structure, we swept one physical axis for each system except a-SiO_2: density for a-C and a-Si and composition for CuZr and Zr–Ni–Al, which gave 14 conditions in total. Each condition contains 50 MD structures, split 35/5/10 into training, validation, and test sets. Though this seemingly limits the benchmark to the low-data regime, each structure has 216 to 500 atoms that are learned independently, leading to up to 25,000 distinct atomic environments. Later, we test the scaling of generative models against a system with up to 3M atomic environments in the full dataset. Nevertheless, the ultra-low-data regime is also closer to the reality of simulations or structure determination in amorphous matter.

2.2 METRICS

As an amorphous configuration is drawn from an ensemble, a generated structure has no unique, analytical ground truth against which it can be compared. One of our main contributions is to make this assessment systematic and grounded in physically relevant principles. First, to quantify the capability of models to generate valid structures, we verify if they are statistically indistinguishable from reference MD samples. No single observable determines this indistinguishability: structures with correct local order can still be energetically inaccessible. AmorphBench therefore defines 21 figures-of-merit in six families (Fig. 1), each probing a physically distinct aspect of disorder and chosen to be complementary. Then, all metrics are computed per structure against the MD reference and summarized by the median over a method's structures according to the families below.

Short-range order. Local order is captured by three terms, all of which are compared individually to the mean result from MD simulations. First, the error between the PDFs $g(r)$ is obtained according to the reduced χ^2 function ($\sigma = 0.1$). Second, bond-angle distribution functions (ADF) capturing three-body interactions are compared according to the Earth Mover's Distance (EMD) (Rubner et al., 2000). Third, coordination number (CN) distributions provide an intuitive perspective on a per-atom basis, and are compared in three different ways: the difference between CN distributions by Jensen–Shannon distance (JSD); the relative error of mean CN; and the fraction of atoms whose CN lies outside the range of the ground-truth MD histogram. The outlier fraction often reveals defects too rare to shift the mean or the distribution distance.

Topology. Medium-range order, which pair and angle correlations leave unconstrained, is characterized by Franzblau rings (Franzblau, 1991) and by Voronoi indices (Sheng et al., 2006). Rings are compared three ways, mirroring coordination: the absolute difference in rings per atom, the JSD between ring-size distributions, and the fraction of rings whose size lies outside the range of the MD histogram. Voronoi indices are compared with a multidimensional EMD under Euclidean ground distance. Rings describe network connectivity in covalent glasses, whereas Voronoi polyhedra describe packing in metallic glasses, so the family spans both bonding types.

Diffraction. Alongside $g(r)$, diffraction patterns connect generated structures to the observables used to validate amorphous models experimentally. X-ray diffraction (XRD) and neutron diffraction (ND) patterns are computed with the Debye scattering equation (Debye, 1915; Johansen et al., 2024) over the atoms in the cell without periodic images, peak-normalized, and compared by EMD; they are finite-cluster proxies rather than the bulk structure factor $S(q)$.

Energetics. Each structure is evaluated with a single-point calculation using the ground-truth force fields for each system. Energy calculations are performed without minimization, and compared according to errors in per-atom energy and EMD for the force-magnitude distribution.

Mechanics. The elastic and plastic responses probe network stiffness, connectivity, and dynamic deformation beyond static structure, and reflect relevant criteria in glassy materials. Bulk, shear, and Young's moduli are obtained as Hill averages from the elastic-constant protocol on relaxed structures (Yang & Schwalbe-Koda, 2026). Cold yield stress (with a 0.2% strain offset) and ultimate tensile strength (UTS) come from $T = 0$ uniaxial deformation averaged over five lattice translations, since a single translation can shift the yield stress by more than a factor of two. The relaxation distance, the RMS non-affine displacement needed to reach the nearest minimum, measures how far a sample lies from a mechanically stable configuration. All six are reported as relative errors.

Phase-space coverage. Per-structure metrics cannot detect mode collapse nor "novelty" of generated structures. We compare atomic-environment distributions with QUESTS (Schwalbe-Koda et al., 2025) on per-density sets. Densities are rescaled to avoid offsets due to density. Ideally, the overlap of environments between generated and simulated structures should be indistinguishable from the mean overlap between two simulated structures, as two different MD simulations do not lead to the same structure. We report the fraction of a method's environments outside the MD manifold (1 - overlap), and the error between the information entropies $\mathcal{H}$ of each structure, $|\mathcal{H}_{\text{src}} - \mathcal{H}_{\text{MD}}|$. Both quantities detect whether structures mode-collapsed into reference data (overly high overlap) or into unreasonable configurations (low overlap, high entropy). "Novel" structures are determined as those with overlaps and entropy error as close to MD as possible.

Scoring and validity. MD reference metrics are obtained from the 10 test structures of each condition; the 35 training structures train the generative models and set the RMC and HRMC targets. A leave-one-out error against mean distributions is computed following the metrics described above, then the average error is normalized to 1.0, accounting for the natural variability of the system. Each method is scored on its first ten structures per condition in filename order, including potentially unphysical ones, to avoid biasing the evaluation toward top candidates. On each metric, structures are pooled and ranked by deviation from MD, and methods are sorted by the mean rank of their errors. In addition, three binary validity criteria are reported: presence of out-of-composition elements, presence of interatomic distances below 1.0 Å, and an identically zero ADF.

2.3 MODELS

We evaluate two classes of methods: baseline data-fitting and network-construction algorithms, which have been the standard route to atomistic models of amorphous materials for decades, and machine-learned generative models, which aim to sample such structures directly from data at a fraction of the cost, but have not yet been comprehensively stress-tested on amorphous matter.

Baselines. Reverse Monte Carlo (RMC) (McGreevy & Pusztai, 1988) displaces one atom at a time and accepts each move by Metropolis (Metropolis et al., 1953) on the error between the structure's pair distribution function (PDF) and the mean PDF of the training structures, with a hard-sphere wall as its only other constraint; it therefore sees a single diffraction-like observable. Hybrid RMC (HRMC) (Opletal et al., 2002) adds to the same cost a Boltzmann-weighted energy term, evaluated with the interatomic potential that the benchmark uses to score energies and forces and quenched to 50 K; this shared potential is a deliberate advantage. RMC was started from liquid MD seeds to improve convergence and allow us to discuss the "best-case scenario" for RMC-like regimes. HRMC was started from random configurations; a liquid-start HRMC test is reported in the appendix for comparison. On the generation algorithm, the continuous random network (CRN) method starts from a four-coordinated network and applies Wooten–Winer–Weaire bond transpositions (Wooten et al., 1985) accepted by Metropolis on the Keating valence force field energy (Keating, 1966) on elemental systems. For a-SiO_2, it later places a bridging oxygen on every Si–Si bond and relaxes the network with the benchmark potential. Because it requires a tetrahedral network, CRN was run only for a-Si and a-SiO_2 and is ranked on fewer rows. Method details are given in the appendix.

Generative models. Generative models are trained from scratch on 35 complete MD structures at each condition, from scratch. MatterGen is also tested as a fine-tuned model. Three models were designed for amorphous materials, all diffusion models over atomic positions in a periodic cell (GLASS and DM^2 at fixed composition): GLASS uses a message-passing network trained per

system on its pooled conditions (Guo & Schwalbe-Koda, 2026); DM^2 uses an equivariant NequIP network trained per condition (Yang & Schwalbe-Koda, 2026); and AMDEN uses an equivariant network trained per system and conditions natively on density (Finkler et al., 2026). Two models that are strong performers in the field of inorganic crystals and exhibit flexibility with the composition and lattice were also tested. CDVAE (Xie et al., 2022), adapted to fixed-composition amorphous cells following Dismai-Bench (Yong et al., 2024), is the only model that produced elements absent from its training composition. MatterGen (Zeni et al., 2025), fine-tuned per system and conditioned on the chemical system, cannot condition on density, so samples were accepted only within 0.25 g/cm^3 of the target. Despite the modifications, crystal models can be at disadvantage by degrees of freedom that amorphous cells at fixed density and composition do not need. Each model is evaluated unconditionally (suffix u) and, except CDVAE, also with conditional sampling (suffix c), which adds a gradient toward the training-set mean PDF at each denoising step (Guo & Schwalbe-Koda, 2026).

3 Results

3.1 Generating a network amorphous silicon

Table 1: **Full metric profile for amorphous silicon (a-Si at 2.3 g/cm^3).** Median method deviations from MD reference normalized by MD's leave-one-out noise floor (1.0 indicates scatter within normal MD variation; $\pm$ is the bootstrap standard error of each method's median). Bold indicates lowest value per row; asterisks mark cells below one minus MD's bootstrap band; ‡ indicates $\geq 50\%$ draw failures.

Metric	RMC (n=10)	HRMC (n=10)	CRN (n=10)	GLASS-c (n=10)	DM2-c (n=10)	CDVAE (n=10)	MatterGen-c (n=5)	AMDEN-c (n=10)
g(r)	24 ±0.2	33 ±0.2	13 ±1.0	**0.5 ±0.1***	3.7 ±0.4	1.1 ±0.2	10 ±2.5	12 ±1.1
ADF	21 ±0.3	26 ±0.2	4.5 ±0.1	0.8 ±0.2*	4.0 ±0.2	**0.8 ±0.1***	7.7 ±0.6	18 ±0.4
CN dist	7.2 ±0.1	8.6 ±0.2	3.7 ±0.3	**0.8 ±0.1***	1.0 ±0.1	1.0 ±0.1	4.2 ±0.4	4.8 ±0.3
CN outliers	46 ±2.6	65 ±3.5	**0.0 ±0.0***	1.3 ±0.4	1.3 ±0.3	0.7 ±0.3	8.0 ±2.3	17 ±1.8
mean CN	2.8 ±0.1	10 ±0.4	3.4 ±0.1	0.4 ±0.1*	**0.3 ±0.3***	0.8 ±0.1	7.9 ±0.9	5.9 ±0.6
rings	5.2 ±0.5	1.8 ±0.3	2.8 ±0.3	0.7 ±0.2*	0.9 ±0.3	**0.5 ±0.1***	7.5 ±1.2	0.8 ±0.6*
ring dist	6.8 ±0.1	8.0 ±0.2	1.7 ±0.1	**0.8 ±0.1***	1.6 ±0.1	1.0 ±0.1	2.3 ±0.1	5.4 ±0.2
ring outliers	23 ±4.2	18 ±4.0	**0.0 ±0.4**	0.0 ±0.1	0.0 ±0.1	0.0 ±0.0	0.0 ±0.4	14 ±3.2
Voronoi	1.9 ±0.0	2.2 ±0.0	1.1 ±0.0	1.0 ±0.0	1.0 ±0.0	**0.9 ±0.1***	1.0 ±0.0	1.6 ±0.0
XRD	1.9 ±0.3	2.5 ±0.1	3.5 ±0.4	1.6 ±0.2	1.8 ±0.4	**1.3 ±0.7**	9.7 ±2.2	2.2 ±0.3
ND	1.9 ±0.2	2.2 ±0.1	2.8 ±0.3	1.3 ±0.2	1.4 ±0.3	**1.1 ±0.6**	8.7 ±2.0	1.8 ±0.1
bulk K	39 ±2.4	40 ±206	13 ±1.2	**3.5 ±0.6**	13 ±1.7	5.7 ±1.0	5.7 ±4.7	29 ±2.2
shear G	15 ±1.5	18 ±1.3	5.7 ±0.6	3.6 ±0.9	9.3 ±1.4	5.7 ±0.8	**2.1 ±2.4**	14 ±0.9
Young E	17 ±1.6	19 ±1.3	6.1 ±0.6	3.6 ±1.0	9.9 ±1.4	5.8 ±0.8	**2.9 ±2.1**	15 ±1.0
relax dist	26 ±0.7	31 ±1.1	0.7 ±0.3*	**0.6 ±0.2***	6.4 ±0.8	0.9 ±0.3	1.3 ±0.4	23 ±0.5
relaxed n [converged/total]	10/10	10/10	10/10	10/10	10/10	8/10	5/5	10/10
yield σ	1.4 ±0.1	1.4 ±0.1	1.3 ±0.5	0.7 ±0.2*	1.0 ±0.1	**0.5 ±0.6***	1.0 ±0.3	1.5 ±0.3
UTS	6.9 ±0.4	6.8 ±0.4	1.6 ±0.4	1.5 ±0.6	2.8 ±0.2	**0.8 ±0.4***	4.0 ±1.1	6.4 ±0.3
Energy/atom	136 ±1.7	172 ±2.0	6.1 ±1.3	3.7 ±0.4	16 ±0.8	**2.1 ±0.3**	6.2 ±0.4	80 ±2.1
Forces	64 ±0.6	68 ±0.9	19 ±1.3	**1.2 ±0.3**	5.5 ±0.4	4.4 ±0.4	8.6 ±0.5	30 ±1.7
QUESTS	7.3	7.9	1.7	1.1	**1.0**	1.1	3.1	4.8
QUESTS div	14	15	1.0	1.1	**0.1**	0.9	5.7	12
foreign n [count/total]	0/10	0/10	0/10	0/10	0/10	2/10	0/5	0/10
Overall (×MD)	14	15	2.8	1.1	1.6	**1.0**	5.7	12
Mean rank	6.7	7.6	4.1	1.8	3.1	2.4	4.5	5.7

Network systems such as a-Si are the archetypical glassy material, some exhibiting over 50 years of structural models (Lewis, 2022). Despite comprehensive studies, many sampling problems in the structure of a-Si remain elusive due to the kinetics of the glass transition (Rosset et al., 2025; 2026). Our dataset on a-Si at the experimental density exposes a distinct failure mode for each class of method (Table 1). RMC, the traditional route from diffraction data to structural models, is difficult to converge to a physical structure even from liquid seeds: its PDF residual reached reasonably low values against the target within the computational budget (Appendix C.2.2), yet all other unconstrained metrics remain far from convergence. Adding the ground-truth potential does not close this gap, and surprisingly HRMC deteriorates the result in terms of convergence, likely because of the initialization and number of quench steps (Appendix C.3.1). CRN, which is a purely unconditional

generation algorithm, performs reasonably: its ring statistics and structural rows are much closer to MD than those of RMC and vary little between seeds, but its network is more ordered than the reference, with 99.5% fourfold atoms against 85.9% for this MD reference (Appendix C.4.1), so distribution-matching rows such as $g(r)$ and CN penalize idealization rather than unphysical structure. The generative models span a wide range. GLASS-c, DM^2-c, and CDVAE reproduce most of the structural distribution, with GLASS-c reaching a close match on $g(r)$. For this model, the error of generated structures is smaller than the variability between independent MD replicas. On the other hand, energy, forces, relaxation distance, and elastic constants show that even the well-performing GLASS-c is unable to fully reproduce the ground-truth energy, identifying these rows as informative discriminating tests. QUESTS overlap and entropy show that memorization of training environments is not present in any structure, but high complexity structures are often generated with (H)RMC, MatterGen, and AMDEN. CDVAE, although designed for crystals, produces plausible amorphous configurations but deviates most in the mechanical properties. However, 24% of CDVAE's covalent structures fail a validity test, mostly because elements absent from the training composition are introduced and otherwise because atoms overlap (15 and 2 of its 17 failures) (Table 4, Section 3.4). Discrepancies across metric groups showcase that no single figure-of-merit dominates in measuring the quality of generated structures (Sec. E.1). Nevertheless, trends in variability and bias can be extracted: yield strength shows lower variability, whereas distribution outliers exhibit failure modes both from lack and excess of outliers (Fig. 13).

3.2 Generating a metallic glass

Table 2: **Full metric profile for metallic glass ($Cu_{50}Zr_{50}$ at 7.4 g/cm^3).** Method deviations relative to MD reference noise floor. Setup, formatting, and markers are identical to Table 1.

Metric	RMC (n=10)	GLASS-c (n=10)	DM2-c (n=10)	CDVAE (n=10)	MatterGen-c (n=10)	AMDEN-c (n=10)
g(r)	22 ±0.4	**1.9 ±0.1**	4.1 ±0.4	2.4 ±0.4	4.4 ±6.4	16 ±1.9
ADF	5.5 ±0.2	1.3 ±0.2	**1.3 ±0.2**	1.6 ±0.2	1.7 ±0.3	5.7 ±0.5
CN dist	3.8 ±0.1	1.2 ±0.1	**1.0 ±0.2**	1.2 ±0.1	1.5 ±0.3	1.9 ±0.2
CN outliers	**0.4 ±0.2***	1.6 ±0.2	1.2 ±0.3	1.2 ±0.1	1.2 ±0.2	1.0 ±0.4
mean CN	**1.0 ±0.3**	3.8 ±2.1	2.1 ±1.4	3.8 ±1.2	7.0 ±1.6	6.1 ±1.3
rings	2.3 ±0.2	1.7 ±0.8	**0.9 ±0.7**	1.9 ±0.7	2.6 ±0.6	2.2 ±1.5
ring dist	3.9 ±0.4	1.3 ±0.3	**1.0 ±0.1**	1.5 ±0.5	3.1 ±0.6	5.2 ±0.3
ring outliers	1.7 ±1.0	**0.0 ±0.0**	0.0 ±0.0	0.0 ±0.0	0.2 ±0.2	1.0 ±0.5
Voronoi	1.7 ±0.1	0.9 ±0.1*	**0.8 ±0.0***	0.9 ±0.0*	0.9 ±0.1*	2.4 ±0.2
XRD	13 ±0.5	1.6 ±0.2	**1.6 ±0.2**	2.7 ±0.2‡	5.9 ±2.2	6.5 ±1.0
ND	3.5 ±0.4	1.1 ±0.2	**1.0 ±0.1**	2.1 ±0.6‡	3.6 ±1.4	1.9 ±0.1
bulk K	7.6 ±0.9	2.6 ±0.9	2.2 ±0.3	3.2 ±1.0‡	7.7 ±3.5	**1.8 ±0.5**
shear G	2.3 ±0.5	2.2 ±0.5	2.3 ±0.7	3.4 ±0.8‡	1.7 ±0.7	**1.6 ±0.9**
Young E	2.5 ±0.6	2.2 ±0.5	2.3 ±0.7	3.4 ±0.8‡	1.8 ±0.8	**1.6 ±0.9**
relax dist	58 ±1.0	7.7 ±1.6	**5.1 ±1.2**	12 ±2.0‡	23 ±3.2	39 ±1.3
relaxed n [converged/total]	10/10	10/10	10/10	4/10	10/10	10/10
yield σ	2.0 ±0.5	**1.2 ±0.6**	1.5 ±0.2	2.0 ±0.5‡	2.3 ±0.6	1.7 ±0.5
UTS	3.4 ±0.8	**1.1 ±0.6**	1.8 ±0.9	4.3 ±1.0‡	4.2 ±0.4	5.1 ±0.4
Energy/atom	286 ±9.6	10 ±1.0	**4.0 ±0.6**	14 ±4.6‡	66 ±28	79 ±8.1
Forces	92 ±0.6	**1.5 ±0.2**	22 ±0.6	7.9 ±0.6‡	21 ±0.9	1.5 ±0.3
QUESTS	1.6	1.0	**0.8**	0.8	0.9	1.3
QUESTS div	4.8	1.5	5.1	4.6	**1.2**	6.0
foreign n [count/total]	0/10	0/10	0/10	6/10	0/10	0/10
Overall (×MD)	3.5	**1.5**	1.6	2.4	2.6	2.2
Mean rank	4.9	2.3	1.8	3.9	3.7	4.5

Unlike the covalent network of a-Si, where errors concentrate in bond angles, coordination defects, and rings, $Cu_{50}Zr_{50}$ is a densely packed, non-directionally bonded glass (Table 2). Its typical structural descriptor is therefore Voronoi packing rather than ring topology, without a CRN analogue. The mechanical response is often a primary design criterion, and thus needs to be well-reproduced. The benchmark shows that generating structures with reasonable packing is straightforward. Even RMC can reproduce coordination reasonably within MD error, and GLASS-c, DM^2-c, CDVAE,

and MatterGen-c all fall below the MD error on the Voronoi metric (0.8–0.9×). The same RMC structures, however, exhibit high errors in their potential energy surface and low convergence upon relaxation. The structural and energetic errors of a metallic glass therefore differ by two orders of magnitude, and seemingly correct packing does not place a structure near a minimum of the potential energy. Energetics are also harder for the generative models than on a-Si. GLASS-c, among the best on the structural rows, rises from 3.7× error on a-Si to 10× on energy per atom. No model generates structures whose energies are within MD error. AMDEN-c illustrates the decoupling between observables. It matches the elastic moduli best of all methods and matches forces as well as GLASS-c, yet it exhibits high errors on energy and ultimate tensile strength. Metallic glasses are therefore not easier to generate than covalent networks because of the lack of directionality. Instead, models fail for complementary reasons related to composition, sampling, and energetics. Each bonding type introduces its own class of imperfections, supporting the use of a comprehensive evaluation scheme to assess sample quality.

3.3 Rankings across the full benchmark

Table 3: **Cross-system performance standings.** Mean ranks over the 21 scored metric rows, with places assigned from pooled per-structure ranks, averaged over each system's conditions and across all five systems (± indicates seeded-bootstrap standard error; lower is better; 1 = closest to MD). Per-system standings are shown in parentheses.

System	RMC	HRMC	CRN	GLASS-c	DM2-c	CDVAE	MatterGen-c	AMDEN-c
a-C	5.4	6.2	—	**1.5**	2.8	2.6	3.1	5.0
a-Si	6.4	7.1	5.5	**1.8**	3.0	2.6	4.1	5.4
SiO_2	6.6	7.1	4.1	**2.1**	3.4	2.6	4.6	5.5
CuZr	4.9	—	—	**1.9**	2.4	3.4	3.8	4.6
ZrNiAl	5.4	5.9	—	**2.6**	3.0	3.2	4.1	3.9
mean rank	5.7 ±0.04	6.6 ±0.04	4.8 ±0.06	**2.0 ±0.07**	2.9 ±0.06	2.9 ±0.11	4.0 ±0.08	4.9 ±0.06
mean ×MD	8.2	13	5.7	**1.3**	2.2	1.7	3.7	8.2

Averaged over all five systems and their density and composition variations, GLASS-c ranks first with a mean rank of 2.00 ± 0.07 and places first in every system (Table 3). Notably, first in every system is not first on every metric: each cell averages a method's places over the 21 metric rows, so a mean rank of 1.0 would require first place on every row in all 14 conditions. GLASS-c is first on 43% of its 294 metric–condition pairs and second on a further 29%; the rows where it most often loses are QUESTS overlap, the ring outlier fraction, and the bulk modulus, and it loses them mostly to CDVAE and DM^2-c. Those two are statistically tied behind it, at 2.88 ± 0.11 and 2.91 ± 0.06. For a-Si and SiO_2, CRN and AMDEN-c are tied, despite AMDEN's stronger performance for metallic glasses. Nevertheless, the first place still does not exhibit full parity with MD, and has ample room for improvement. On the network formers (a-C, a-Si, and silica), methods underperform on coordination number, bond angles, and validity, since defects are easier to generate in open, low-density networks. On the metallic glasses, packing is matched by nearly every method, and the discriminating factors become energy, relaxation distance, and plastic response. The performances of vanilla RMC and HRMC, however, show that the generative models outperform traditional algorithms. Monte Carlo methods can still be more accessible to experts looking to impose hand-crafted constraints on their systems or attempt structure prediction with zero priors. Overall, however, neither a converged PDF fit nor an annealed energy term necessarily represent a physical structure, as demonstrated in the previous tables. CDVAE is the unexpected result: a crystal model with no amorphous-specific design ties for second across both bonding types. Its performance, however, carries a major caveat, with 24–30% failure rate dominated by introduction of foreign elements in the desired systems (Section 3.4).

3.4 Validity and yield

The previous results show that median quality across metrics address both structure and property of a given sample. However, a practitioner also needs to know how many draws are required to obtain a usable sample on a per-model basis, which is determined by validity and yield. We define validity through three low-cost, necessary conditions that no physical structure violates: composition within the training set, no interatomic distance below 1.0 Å, and a non-zero ADF. Yield is the fraction

Table 4: **Validity gate pass rates and draw accounting.** Structure generation counts, validity pass rates across three gating criteria (composition, interatomic distance, bond angle distribution), and worst-performing conditions per method across all 14 conditions. HRMC (all arms) pools the random-start, liquid-start, and 0.80-wall arms, which pass every gate on every structure.

Method	Conditions	Drawn	Valid	Rate	Worst condition
MD	14	140	140	100%	—
RMC	14	140	140	100%	—
HRMC (all arms)	14	220	220	100%	—
CRN	4	40	40	100%	—
GLASS-c	14	140	140	100%	—
GLASS-u	14	140	140	100%	—
DM2-c	14	170	146	86%	16/40 (C_1.9)
DM2-u	14	250	140	56%	10/120 (C_1.9)
CDVAE	14	280	205	73%	9/20 (Zr1Ni3Al1_6.7)
MatterGen-c	13	104	104	100%	—
MatterGen-u	11	101	101	100%	—
AMDEN-c	14	136	103	76%	0/10 (C_1.9)
AMDEN-u	14	138	99	72%	0/10 (C_1.9)

of draws that are both delivered at the requested condition and valid (Table 4; per-gate counts in Appendix E.4). The criteria are necessary but not sufficient: RMC passes all of them on every structure while still exhibiting low quality samples. On the other hand, CDVAE produces high-quality structures, but because it predicts atom types freely, it also corrupts composition (73% valid overall; every metallic failure is a foreign element). DM^2 and AMDEN occasionally place overlapping atoms, with minimum distances down to 0.04 Å, with failures exacerbated in low-density covalent networks. DM^2-u needed 120 draws for 10 valid structures at a-C 1.9 g/cm^3, and AMDEN yielded none at that condition. Similarly, MatterGen avoids generating low-density structures, requiring density conditioning to reward samples within certain density windows. RMC, HRMC, CRN, and both GLASS arms enforce validity by construction, as they fix composition, density, and enforce a minimum distance. Because the validity criteria are not costly to check, invalid structures can be discarded and redrawn. However, scoring only the first ten valid structures mostly rewards CDVAE. Yield and conditional quality are therefore kept as separate claims. Within our benchmark, only GLASS-c achieves both across all 14 conditions.

3.5 Data Scaling

All benchmark tests are performed in the low-data regime, which matches the current reality of computational studies in amorphous materials. Each reference structure requires an expensive simulation over large length and time scales, especially if using machine learning potentials or, worse, density functional theory calculations. This regime may itself explain part of the failures above: CDVAE and AMDEN were originally developed on much larger datasets and may perform better at those scales, but no amorphous dataset of comparable size or transferability currently exists. To test how sample quality depends on data, we built a separate dataset, sodium–calcium borosilicate (NCBS, Na_2O–CaO–B_2O_3–SiO_2), spanning a compositional space of 117 different compositions and allowing a more diverse set of structures and compositions to be generated compared to the existing benchmark. The scaling study was restricted to GLASS, the best-performing model across the benchmark, because of its computational cost. Trained at a constant 384 epochs on 30, 100, 300, and 1,000 structures (about 3,000 atoms each) with four seeds each, GLASS reduced its held-out validation loss from 1.36 to 0.81 (Fig. 2), following a power law with exponent -0.144 ($R^2 = 0.89$), or about a 28% reduction per decade of data. Seed-to-seed variability fell from 0.078 to 0.002. At a fixed epoch count, larger sets also receive more optimizer steps. The trend nevertheless indicates that the models remain data-limited rather than capacity-limited. Larger amorphous datasets spanning many compositions could then yield transferable samplers that bypass costly melt-quench simulations, particularly for large cells and slow cooling rates, where MD is most expensive. The results, however, already show that existing generative models are capable of handling high-dimensional composition and structural spaces, including the challenging space of amorphous matter. Neverthe-

less, developments in dataset design and model architectures may tackle the simultaneous problem of low-data and high-sensitivity of these materials, as highlighted by the metrics above and the complete assessment on the Appendix.

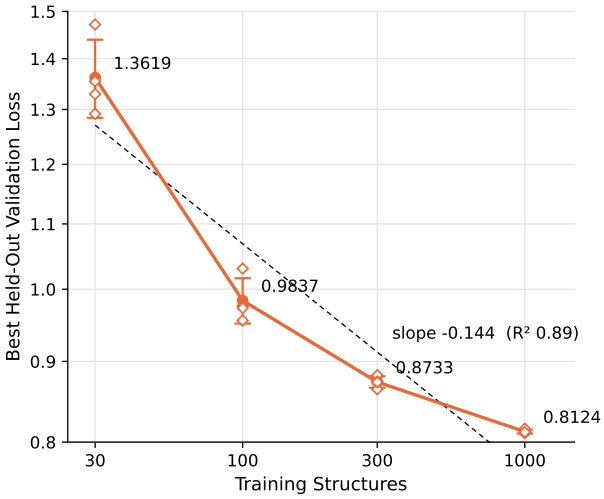


Figure 2: **Held-out validation loss versus training-set size at a constant epoch budget.** Evaluated at 384 epochs for all N. Markers show the mean over available seed replicates (± 1 SD, with individual seeds as hollow diamonds); the dashed line represents the log-log least-squares fit (slope -0.144, $R^2 = 0.89$).

4 Discussion and Conclusion

Limitations. Systematic testing of the performance of models revealed a gap between ground-truth and generated structures. However, the reference is a classical potential rather than experimental or full density functional theory data, so the metrics measure agreement with the physics that produced the MD ensemble rather than an absolute metric. Furthermore, while variability and normalization was employed as much as possible, each method is scored on 10 structures per condition; bootstrap errors and pairwise ordering probabilities carry this uncertainty, and close standings are reported as ties rather than placements. Finally, the analysis is realistic in terms of real-world task, but still lies at the data-scarce regime. Future efforts can generate amorphous configurations at scale using machine learning potentials and robust strategies to sample configurations.

Conclusions. AmorphBench shows that structural agreement alone would declare generative modeling of amorphous matter nearly solved. The leading models reproduce coordination, packing, rings, and diffraction within the MD variability, yet they are still far from excellent energy, forces, relaxation distance, and phase-space coverage. Scientific insights with generative models are possible with expert supervision (Guo & Schwalbe-Koda, 2026), but failure modes still largely depend on system type. Covalent networks often fail through coordination defects, bond-angle errors, and overlapping atoms in open structures, whereas metallic glasses match packing readily but fail on energetics and plastic response, so no single structural criterion captures sample quality across systems. GLASS ranks first in every system with full validity. CDVAE, a crystal model, ties for second despite validity issues at sampling, illustrating a case where sample quality and validity are decoupled in amorphous materials generation. The continued improvement of GLASS with larger training datasets indicates that current models are capable of learning increasingly diverse structures, incentivizing the creation of larger datasets spanning compositions and processing conditions. Our benchmark is accompanied by Python scripts, an interactive leaderboard, and datasets that establish references that measure quality in generative models for amorphous materials beyond structural features and toward a comprehensive suite that characterizes the complexity of these materials adequately.

AI USE STATEMENT

In this work, we used generative AI tools (Claude) to assist in implementing parts of the analysis code. We have not used generative AI tools to develop the method, formulate hypotheses, design experiments, generate data, or interpret results, and the remaining required-disclosure tasks (e.g., translation, qualitative data analysis) are not applicable to this work. Additionally, we used generative AI tools for editing plotting scripts, improving the readability of the manuscript, and formatting and checking references. We reviewed all AI-assisted work: AI-assisted code was tested by the authors against reference outputs, and every reference was checked 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

This work does not involve human subjects, personal data, or experimental measurements. All reference structures were generated with classical molecular dynamics using published interatomic potentials, and third-party models were used under their public licenses. To keep the comparison fair, all methods were trained on the same data splits and scored with the same pipeline on all generated structures, including invalid ones. We do not foresee harmful applications of this benchmark.

REPRODUCIBILITY STATEMENT

We provide the anonymized AmorphBench package as supplementary material, the scoring code, and models. Datasets and metrics are described in Sections 2.1–2.2, and training, fine-tuning, and PDF-guidance settings for all methods, including hyperparameter sweeps, are given in Appendices A–D. Uncertainties are computed with a seeded bootstrap. The generated structures, package, datasets, trained checkpoints, and leaderboard will be released publicly upon acceptance.

REFERENCES

G. T. Barkema and Normand Mousseau. High-Quality Continuous Random Networks. *Physical Review B*, 62(8):4985–4990, 2000. doi: 10.1103/PhysRevB.62.4985.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nature Communications*, 13 (1):2453, 2022. doi: 10.1038/s41467-022-29939-5.

Matilde Benassi, Annalisa Pallini, Marco Bertani, Sofia C. Sarnataro, Marta Corno, and Alfonso Pedone. Assessing Universal Machine Learning Interatomic Potentials for Aluminosilicate Glasses. *Journal of the American Ceramic Society*, 109(7):e70962, 2026. doi: 10.1111/jace.70962.

Noam Bernstein, Bishal Bhattarai, Gábor Csányi, David A. Drabold, Stephen R. Elliott, and Volker L. Deringer. Quantifying Chemical Structure and Machine-Learned Atomic Energies in Amorphous and Liquid Silicon. *Angewandte Chemie International Edition*, 58(21):7057–7061, 2019. doi: 10.1002/anie.201902625.

Marco Bertani and Alfonso Pedone. Machine Learning Interatomic Potentials for NaPSO Glasses: The Critical Role of Training Data. *Solid State Sciences*, 174:108204, 2026. doi: 10.1016/j.solidstatesciences.2025.108204.

Simon J. L. Billinge and Igor Levin. The Problem with Determining Atomic Structure at the Nanoscale. *Science*, 316(5824):561–565, 2007. doi: 10.1126/science.1135080.

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, March 2019.

Yuan Chiang, Tobias Kreiman, Christine Zhang, Matthew Kuner, Elizabeth Weaver, Ishan Amin, Hyunsoo Park, Yunsung Lim, Jihan Kim, Daryl Chrzan, et al. Mlip arena: advancing fairness

and transparency in machine learning interatomic potentials via an open, accessible benchmark platform. *Advances in Neural Information Processing Systems*, 38, 2026.

Matthew J. Cliffe, Martin T. Dove, David A. Drabold, and Andrew L. Goodwin. Structure Determination of Disordered Materials from Diffraction Data. *Physical Review Letters*, 104(12):125501, 2010. doi: 10.1103/PhysRevLett.104.125501.

Pierre-Paul De Breuck, Hai-Chen Wang, Gian-Marco Rignanese, Silvana Botti, and Miguel A. L. Marques. Generative AI for Crystal Structures: A Review. *npj Computational Materials*, 11(1): 370, 2025. doi: 10.1038/s41524-025-01881-2.

P. Debye. Zerstreuung von Röntgenstrahlen. *Annalen der Physik*, 351(6):809–823, 1915. doi: 10.1002/andp.19153510606.

Volker L. Deringer, Noam Bernstein, Gábor Csányi, Chiheb Ben Mahmoud, Michele Ceriotti, Mark Wilson, David A. Drabold, and Stephen R. Elliott. Origins of Structural and Electronic Transitions in Disordered Silicon. *Nature*, 589(7840):59–64, 2021. doi: 10.1038/s41586-020-03072-z.

Andrei Dernov, M. Kowalik, A. C. T. van Duin, and T. Dumitrică. Mapping the Structural–Mechanical Landscape of Amorphous Carbon with ReaxFF Molecular Dynamics. *Journal of Applied Physics*, 137(6):065107, 2025. doi: 10.1063/5.0246126.

Jun Ding, Yong-Qiang Cheng, and Evan Ma. Full Icosahedra Dominate Local Order in $Cu_{64}Zr_{34}$ Metallic Glass and Supercooled Liquid. *Acta Materialia*, 69:343–354, 2014. doi: 10.1016/j.actamat.2014.02.005.

Bradley Efron. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

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. doi: 10.1002/jcc.70306.

Stephen R. Elliott. *Physics of Amorphous Materials*. Longman Scientific & Technical, Harlow, 2nd edition, 1990.

Jonas A. Finkler, Yan Lin, Tao Du, Jilin Hu, and Morten M. Smedskjaer. Inverse Design of Amorphous Materials with Targeted Properties. *Advanced Materials*, 38(39):e22493, 2026. doi: 10.1002/adma.202522493.

D. S. Franzblau. Computation of Ring Statistics for Network Models of Solids. *Physical Review B*, 44(10):4925–4930, 1991. doi: 10.1103/PhysRevB.44.4925.

Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, San Diego, 2nd edition, 2002.

Johannes Gasteiger, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. In *Machine Learning for Molecules Workshop, NeurIPS*, 2020. arXiv:2011.14115.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal Directional Graph Neural Networks for Molecules. In *Advances in Neural Information Processing Systems*, 2021. arXiv:2106.08903.

K. Georgarakis, A. R. Yavari, M. Aljerf, D. V. Louzguine-Luzgin, M. Stoica, G. Vaughan, and A. Inoue. On the Atomic Structure of Zr–Ni and Zr–Ni–Al Metallic Glasses. *Journal of Applied Physics*, 108(2):023514, 2010. doi: 10.1063/1.3446131.

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.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268, 2018.

Jiawei Guo and Daniel Schwalbe-Koda. Generative Inversion of Spectroscopic Data for Amorphous Structure Elucidation. *arXiv preprint arXiv:2603.23210*, 2026.

Tim Hsu, Babak Sadigh, Nicolas Bertin, Cheol Woo Park, James Chapman, Vasily Bulatov, and Fei Zhou. Score-based denoising for atomic structure identification. *npj Computational Materials*, 10(1):155, 2024.

Frederik L. Johansen, Andy S. Anker, Ulrik Friis-Jensen, Erik B. Dam, Kirsten M. Ø. Jensen, and Raghavendra Selvan. A GPU-Accelerated Open-Source Python Package for Calculating Powder Diffraction, Small-Angle-, and Total Scattering with the Debye Scattering Equation. *Journal of Open Source Software*, 9(94):6024, 2024. doi: 10.21105/joss.06024.

P. N. Keating. Effect of Invariance Requirements on the Elastic Strain Energy of Crystals with Application to the Diamond Structure. *Physical Review*, 145(2):637–645, 1966. doi: 10.1103/PhysRev.145.637.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

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(2):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.

Laurent J Lewis. Fifty years of amorphous silicon models: the end of the story? *Journal of Non-Crystalline Solids*, 580:121383, 2022.

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.

Han Liu, Zhangji Zhao, Qi Zhou, Ruoxia Chen, Kai Yang, Zhe Wang, Longwen Tang, and Mathieu Bauchy. Challenges and opportunities in atomistic simulations of glasses: a review. *Comptes Rendus. Géoscience*, 354(S1):35–77, 2022.

Yuanbin Liu, Ata Madanchi, Andy S Anker, Lena Simine, and Volker L Deringer. The amorphous state as a frontier in computational materials design. *Nature Reviews Materials*, 10(3):228–241, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Shreshth A Malik, Tiarnan Doherty, Panagiotis Tigas, Muhammed Razzak, S Roberts, Aron Walsh, and Yarin Gal. MADE: Benchmark environments for closed-loop materials discovery. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=nrXxVDYMMF`.

N. Mattern, P. Jóvári, I. Kaban, S. Gruner, A. Elsner, V. Kokotin, H. Franz, B. Beuneu, and J. Eckert. Short-Range Order of Cu–Zr Metallic Glasses. *Journal of Alloys and Compounds*, 485(1–2):163–169, 2009. doi: 10.1016/j.jallcom.2009.05.111.

D. G. McCulloch, D. R. McKenzie, and C. M. Goringe. *Ab Initio* Simulations of the Structure of Amorphous Carbon. *Physical Review B*, 61(3):2349–2355, 2000. doi: 10.1103/PhysRevB.61.2349.

R. L. McGreevy. Reverse Monte Carlo Modelling. *Journal of Physics: Condensed Matter*, 13(46): R877–R913, 2001. doi: 10.1088/0953-8984/13/46/201.

R. L. McGreevy and L. Pusztai. Reverse Monte Carlo Simulation: A New Technique for the Determination of Disordered Structures. *Molecular Simulation*, 1(6):359–367, 1988. doi: 10.1080/08927028808080958.

M. I. Mendelev, D. J. Sordelet, and M. J. Kramer. Using Atomistic Computer Simulations to Analyze X-Ray Diffraction Data from Metallic Glasses. *Journal of Applied Physics*, 102(4):043501, 2007. doi: 10.1063/1.2769157.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114.

Joe D. Morrow, Chinonso Ugwumadu, David A. Drabold, Stephen R. Elliott, Andrew L. Goodwin, and Volker L. Deringer. Understanding Defects in Amorphous Silicon with Million-Atom Simulations and Machine Learning. *Angewandte Chemie International Edition*, 63(22):e202403842, 2024. doi: 10.1002/anie.202403842.

Shinji Munetoh, Teruaki Motooka, Koji Moriguchi, and Akira Shintani. Interatomic Potential for Si–O Systems Using Tersoff Parameterization. *Computational Materials Science*, 39(2):334–339, 2007. doi: 10.1016/j.commatsci.2006.06.010.

G. Opletal, T. C. 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. doi: 10.1080/089270204000002584.

Janosh Riebesell, Rhys EA 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(6):836–847, 2025.

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.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. doi: 10.1023/A:1026543900054.

Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

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(1):4014, 2025. doi: 10.1038/s41467-025-59232-0.

H. W. Sheng, W. K. Luo, F. M. Alamgir, J. M. Bai, and E. Ma. Atomic Packing and Short-to-Medium-Range Order in Metallic Glasses. *Nature*, 439(7075):419–425, 2006. doi: 10.1038/nature04421.

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*, 2021.

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.

J. Tersoff. Modeling Solid-State Chemistry: Interatomic Potentials for Multicomponent Systems. *Physical Review B*, 39(8):5566–5568, 1989. doi: 10.1103/PhysRevB.39.5566.

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.

Matthew G. Tucker, David A. Keen, Martin T. Dove, Andrew L. Goodwin, and Qun Hui. RMCProfile: Reverse Monte Carlo for Polycrystalline Materials. *Journal of Physics: Condensed Matter*, 19(33):335218, 2007. doi: 10.1088/0953-8984/19/33/335218.

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.

F. Wooten, K. Winer, and D. Weaire. Computer Generation of Structural Models of Amorphous Si and Ge. *Physical Review Letters*, 54(13):1392–1395, 1985. doi: 10.1103/PhysRevLett.54.1392.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal Diffusion Variational Autoencoder for Periodic Material Generation. In *International Conference on Learning Representations*, 2022.

Kai Yang and Daniel Schwalbe-Koda. A Generative Diffusion Model for Amorphous Materials. *npj Computational Materials*, 12(1):29, 2026. doi: 10.1038/s41524-025-01901-1.

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.

X. W. Zhou, R. A. Johnson, and H. N. G. Wadley. Misfit-Energy-Increasing Dislocations in Vapor-Deposited CoFe/NiFe Multilayers. *Physical Review B*, 69(14):144113, 2004. doi: 10.1103/PhysRevB.69.144113.

Alex Zunger. Inverse design in search of materials with target functionalities. *Nature Reviews Chemistry*, 2(4):0121, 2018.

APPENDIX

A MODELS

The methods differ less in sophistication than in what they are actually allowed to see. RMC is given a pair distribution function; HRMC is given a pair distribution function and an interatomic potential; CRN is given a topology rule and an energy function; and the generative models are given 35 complete MD structures per condition.

A.1 RMC

RMC (McGreevy & Pusztai, 1988), or reverse Monte Carlo, displaces one atom at a time and accepts or rejects each move by Metropolis on the χ^2 between the computed PDF and the mean PDF of the 35 training structures, with a hard-sphere wall as its only other constraint. To give RMC its best chance, we started it from liquid MD seeds (Appendix C.2).

A.2 HRMC

HRMC (Opletal et al., 2002), or hybrid reverse Monte Carlo, is RMC with an energy term: each proposed move is scored on the PDF residual and on the change in potential energy, the latter through a Boltzmann factor at a temperature quenched over the run to 50 K. We used the interatomic potential that the benchmark uses to score energy and forces, a deliberate advantage for HRMC. The main tables use a random start; a liquid start and, for CuZr, a random start with a relaxed wall appear in the full tables, and CuZr's HRMC only through the relaxed-wall arm (Appendix C.3).

A.3 CRN

CRN, or continuous random network, builds a structure from a topology rule rather than from data. It applies Wooten–Winer–Weaire bond transpositions (Wooten et al., 1985) to a four-coordinated silicon network, accepted by Metropolis on the Keating energy (Keating, 1966), so every atom stays exactly four-fold, as in the networks of Barkema & Mousseau (2000); for silica, a bridging oxygen is placed on every Si–Si bond and the network relaxed with the benchmark potential. CRN needs a tetrahedral network, so it runs only on amorphous silicon and silica. It never sees the MD structures, only the condition's density and cell, a minimum-distance screen, and, for silica, the benchmark's potential (Appendix C.4).

A.4 GLASS

GLASS, or generative learning of amorphous structures from spectra, is a score-based diffusion model over atomic positions in a periodic cell. During training, Gaussian noise is added to the coordinates of a training structure and a message-passing graph neural network learns to predict the direction back toward the clean structure at every noise level. To generate, we start from noise and integrate the reverse diffusion process, using the network's prediction at each step. That alone is the unconditional model, GLASS-u. For the conditional model, GLASS-c, at each step we form an estimate of the clean structure, compute its pair distribution function with a differentiable forward model, compare it to the training-set mean, and add the gradient of that mismatch to the score. Where we have a differentiable implementation of the generating potential we also add the gradient of its energy: the Tersoff potentials of amorphous carbon and amorphous silica, but not the modified Tersoff potential of amorphous silicon or the EAM potentials of the metallics.

A.5 DM^2

DM^2 is a diffusion model built specifically for amorphous materials and is the closest prior work to GLASS. It uses an equivariant NequIP network trained to predict the displacement that was added to a structure as noise. It generates by repeatedly rattling and denoising the positions down a fixed ladder of noise levels from 1.0 to 0.001. We trained one model per condition on the 35 training structures and did not use the authors' released checkpoints, which were trained on much larger datasets. The unconditional entry, DM^2-u, is the sampler run as released. For the conditional entry,

DM^2-c, we added PDF guidance to the noisy stage of the sampler using the same differentiable PDF that GLASS uses, with the training-set mean PDF as the target (Appendix D.3). DM^2 has no energy guidance. Notably, the authors' own conditional model conditions on a cooling rate, an axis none of our conditions vary—our silica condition is a single density, not a cooling-rate sweep—so we did not use this knob.

A.6 CDVAE

CDVAE, or crystal diffusion variational autoencoder, is a generative model built for crystals that we adapted to fixed-composition amorphous cells following Dismai-Bench. A DimeNet++ encoder compresses a structure into a 256-dimensional latent vector. From that latent vector the model predicts the number of atoms, the lattice, and the composition, and a GemNet decoder then places the atoms by denoising random positions over 50 noise levels. We trained one model per condition on the 35 training structures and sampled it unconditionally.

A.7 MATTERGEN

MatterGen is a diffusion model for crystals that jointly denoises the atom types, the fractional coordinates, and the lattice, and can be steered toward a target property with classifier-free guidance. We fine-tuned it per system for 1000 epochs on the training structures, conditioned only on the chemical system, and sampled with its native 1000 steps. MatterGen has no way to condition on density, so we accepted a draw into a condition only if its density fell within 0.25 g/cm^3 of the target, up to 200 draws per system/condition. Notably, it failed to fill 5 of the 14 conditions this way, including 0 of 10 at three of them—amorphous carbon at 2.4 and 2.9 g/cm^3 and amorphous silicon at 2.3 g/cm^3. We report those as coverage failures rather than leaving the cells blank. The conditional entry, MatterGen-c, adds PDF guidance to the same fine-tuned checkpoints under the same acceptance protocol, with the PDF of a single validation structure as the target. Notably, its guidance settings for amorphous carbon, amorphous silicon, and silica come from a pilot ranked against test structures rather than from the validation sweep used for CuZr and ZrNiAl (Appendix D.1). It filled 9 of the 14 conditions.

A.8 AMDEN

AMDEN (Finkler et al., 2026) is a diffusion model designed for amorphous materials that operates over periodic atomic positions with an equivariant graph network predicting position noise at each step. AMDEN conditions on the target density via *ghost atoms*: placeholder atoms carrying a dedicated element class that pad every structure up to a fixed capacity. Because atomic position and element identity are diffused jointly, the reverse process can relabel any atom between the ghost class and a real species, which is the only way the model can change a structure's density. As a result, atom count and composition vary from draw to draw rather than staying fixed at the MD template's values: 203 to 254 atoms against MD's 216 on the elemental systems and, unguided, up to 45.6 atomic percent off the ZrNiAl target stoichiometry. Atoms still labeled ghost after denoising are discarded.

B PROTOCOL

We targeted prominent amorphous systems such as amorphous carbon, amorphous silicon, silica (a-SiO_2), copper-zirconium, and zirconium–nickel–aluminum (Zr–Ni–Al, four compositions); for all but amorphous silica (one condition at 2.2 g/cm^3) we swept one physical axis per system: density for a-C ($1.9/2.4/2.9$) and a-Si ($1.8/2.3/2.8$ g/cm^3), composition for CuZr (Cu:Zr $=$ 30:70, 50:50, 70:30 at $7.0, 7.4$, and 7.9 g/cm^3, one density per composition) and ZrNiAl (Zr:Ni:Al $=$ 1:1:1, 1:1:3, 1:3:1, 3:1:1 at $5.8, 4.6, 6.7,$ and 6.1 g/cm^3); $216/216/300/500/500$ atoms for a-C/a-Si/silica/CuZr/ZrNiAl. Appendix C.1 describes how the structures were generated and split.

The MD reference data was generated with the 1989 Tersoff potential for a-C (Tersoff, 1989), the Kumagai modified Tersoff potential (`LAMMPS pair_style tersoff/mod`) for a-Si (Kumagai et al., 2007), the Munetoh Tersoff potential for silica (Munetoh et al., 2007), Mendelev's 2007 EAM for CuZr (Mendelev et al., 2007), and finally Zhou's 2004 EAM for ZrNiAl (Zhou et al., 2004). The same potential is used for the energy and force metrics and for HRMC, so those rows measure agreement with the physics that produced the reference rather than with a second model of it. For ZrNiAl the mechanics use a tapered version of the same EAM, because the published form is not smooth at its cutoff; the energy and force rows use the published form. Every per-system constant is pinned in the scoring code: rings are counted up to 10 members (40 for silica); QUESTS uses a bandwidth of 0.025 for a-C and silica and 0.015 otherwise, rescales environments to 2.4, 2.3, and 2.2 g/cm^3 for a-C, a-Si, and silica and to each composition's own density for the metallic glasses, estimates entropies from 216, 216, 300, 500, and 500 environments, and resamples 20 times; the elastic constants use a strain of 10^{-4} and all scoring runs against a frozen snapshot of the code.

C TRAINING DETAILS

C.1 DATA SPLIT DETAILS

All reference structures come from classical MD melt-quench runs. Each of the 50 structures per condition is the end point of an independent run rather than a frame of one trajectory, so the structures within a condition are independent draws from one protocol. Every run used LAMMPS (Thompson et al., 2022) in metal units with a 0.5 fs timestep under NVT: a random packing with a 2.0 Å minimum distance was minimized, melted for 50 ps at 5000 K (a-C and silica), 3000 K (a-Si), or 2000 K (CuZr and ZrNiAl), quenched to 300 K at 5 K/ps, and held for 50 ps at 300 K; the silica and metallic runs add a 50 ps mixing stage at 300 K before the melt. We then split each condition at random (seed 0) into 35 training, 5 validation, and 10 test structures. The training structures are the training data of every generative model, the PDF-guidance target of GLASS and DM^2, and the targets and walls of RMC and HRMC. The validation structures select the GLASS and CDVAE checkpoints and drive CDVAE's learning-rate schedule; neither CDVAE nor DM^2 stopped early, and both ran for a fixed budget. They also select the AMDEN guidance settings and the MatterGen guidance settings for CuZr and ZrNiAl (Appendix D.1). Every method, including those not trained on structures (RMC, HRMC, CRN), is scored against the ten test structures, which no method was trained on; the one setting chosen against them is MatterGen-c's guidance on a-C, a-Si, and silica (Appendix A.7).

C.2 RMC TRAINING DETAILS

We used RMCProfile (Tucker et al., 2007) with ten replicas per condition, each started from a liquid seed taken from MD's melt and run for three 24 h chunks (72 h). The run length is fixed with no convergence stop. The target is written on RMCProfile's own grid ($r = 0.04, 0.08, \ldots$ Å, the points it pairs with the model's PDF by index) up to 8 Å or half the box, whichever is shorter; the maximum move is 0.10 Å; and a hard-sphere wall sits at 0.90 (covalent) or 0.80 (metallic) of the shortest pair distance in the training structures. All scattering lengths are equal, so the fitted total PDF is the plain all-atom $g(r)$.

C.2.1 DATA WEIGHT

RMCProfile's weight card is the standard deviation σ assigned to every data point, so it sets the acceptance temperature of the walk. At $\sigma = 1$ the PDF barely constrains the moves: χ^2/n_q rises instead of falling, the first $g(r)$ peak collapses from 1.73 to 0.11, and the atoms settle onto the hard-sphere wall (Table 5). We compared $\sigma \in \{0.05, 0.02, 0.01\}$ in 10-minute runs on a-C at 1.9 g/cm^3; all three descend, keep the first peak, and stay off the wall. We chose $\sigma = 0.02$ because 0.05 under-constrains the fit and 0.01 fits the ensemble noise: a single training structure against the 35-structure mean already sits at $\chi^2/n_q \approx 9$ at $\sigma = 0.01$, against ≈ 2 at $\sigma = 0.02$, where a converged fit sits at order one.

Table 5: **RMC data-weight comparison, a-C at 1.9 g/cm^3.** The $\sigma = 1$ row is one replica run at that weight; the other rows are 10-minute runs from the same seed. $G(r)$ MSE against the training-mean target; first peak of $g(r) - 1$ (target 1.73); minimum pair distance.

σ	χ^2/n_q (start → end)	G(r) MSE	Minimum distance (Å)	First peak (g − 1)
1 (initial)	0.029 → 0.186 (rises)	0.165	0.99 (wall)	0.11
0.05	11.69 → 1.33	0.027	1.087	0.909
0.02	73.1 → 2.5	0.024	1.151	0.944
0.01	292 → 17.4	0.019	1.195	1.034

C.2.2 CONVERGENCE GRAPHS

Figure 3 shows χ^2/n_q against wall time for every condition. It falls from 69–377 at the seed to a median of 0.97–1.26 per condition at 72 h; 12 of the 14 conditions are within $1.5\times$ of their final value within 30 minutes, a-Si at 2.3 within 2 h, and a-Si at 1.8 g/cm^3 within 8 h. The 24 h restarts leave no trace: paired per replica over all 140 fits, χ^2/n_q changes by $+0.5 \pm 15$ percent between 24 and 72 h, so the fixed budget is well past convergence.

C.3 HRMC TRAINING DETAILS

We used pyHRMC (Ehrhardt et al., 2026) with ten replicas per condition, the atom-hop move (0.2 Å maximum), and a cost that sums the PDF residual, weighted by $\sigma = 0.5$, and E/RT with E in kcal/mol. The temperature starts at 2000 K and, after every valid candidate move, is multiplied by $q^{n/10^6}$, where n is the cumulative step count, down to a hard floor of 50 K; the PDF weight is quenched with it. We ran three arms (Table 6): a random start with the wall at 0.95 of the training-set $g(r)$ onset; a liquid start from RMC's seeds with the wall at 0.95 for covalent and 0.80 for metallic systems; and, for CuZr, a random start with the wall at 0.80.

Table 6: **HRMC campaigns.** Wall as a fraction of the training-set $g(r)$ onset; replicas summed over the system's conditions; budget is the cumulative step target.

System	Start	Wall	Replicas	q	Budget (steps)	Status
a-C	random	0.95	30	0.8724	300,000	fixed budget
a-Si	random	0.95	30	0.9412	300,000	fixed budget
SiO_2	random	0.95	10	0.9618	1,040,000	at 50 K floor
ZrNiAl	random	0.95	40	0.9916	560,000	at 50 K floor
a-C	liquid	0.95	30	0.9461	570,000	at 50 K floor
SiO_2	liquid	0.95	10	0.9679	730,000	at 50 K floor
ZrNiAl	liquid	0.80	40	0.9939	750,000	at 50 K floor
CuZr	random	0.80	30	0.9745	300,000	at 50 K floor

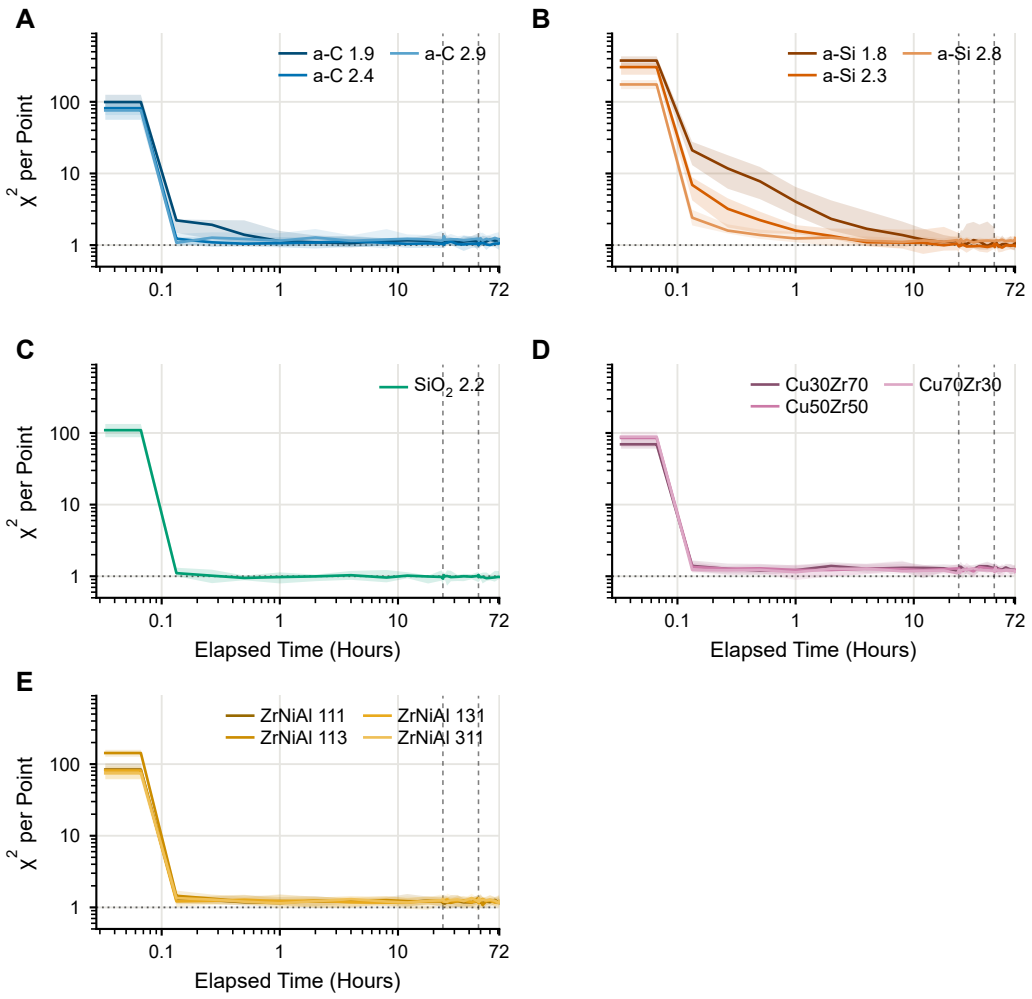


Figure 3: **RMC convergence.** χ^2 per fitted point against elapsed time, one panel per system: (A) a-C, (B) a-Si, (C) silica, (D) CuZr, (E) ZrNiAl. Median over ten replicas per condition, band min–max; the dotted line marks $\chi^2/n_q = 1$ and the dashed verticals the 24 h restarts.

C.3.1 Quench Calibration

Because the temperature updates only on valid candidate moves, q does not transfer between systems, and we calibrated it per campaign from its own temperature trace, assuming the accumulated anneal grows quadratically in the step count. Fitting the long traces instead gives an exponent of 0.92 to 1.48 early and 0.64 to 1.14 late: the anneal grows roughly linearly, so every fixed budget lands hot. The 15 ZrNiAl random-start replicas that completed 300,000 steps, for example, ended at 89 to 190 K (median 119 K) against the 50 K target. We therefore kept each q and extended the budget instead: a replica counts as finished only when it has reached its budget and ended at or below 55 K, and a campaign whose replicas all finish hot has its budget raised by 25 percent. Budgets therefore differ by campaign by design, and every arm except the a-C and a-Si random starts (Table 6) is compared at the same final temperature rather than the same step count.

C.3.2 Starting State and Wall

At 0.95 of the $g(r)$ onset, the random-start wall on CuZr sits above the closest pair in the MD reference itself (2.18 against 2.122 Å at $Cu_{30}Zr_{70}$), so that arm cannot reproduce the structure it fits; the 0.80 arm removes the constraint. In the full tables (Appendix E.3), the liquid start helps most

where the anneal is the bottleneck, on a-C and silica, while the 0.80 wall leaves HRMC last in every CuZr condition.

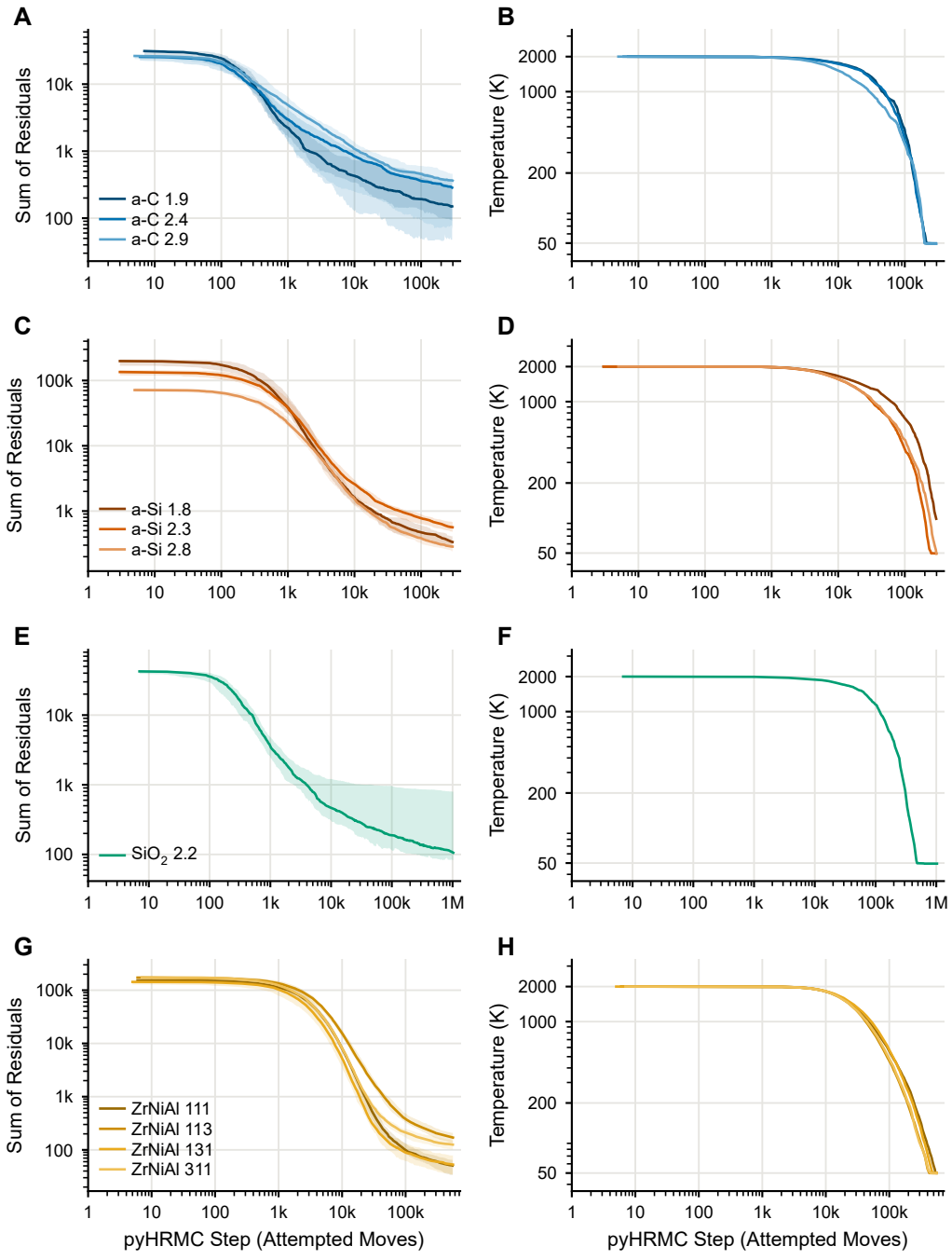


Figure 4: **HRMC convergence, random start.** Sum of residuals (left) and the temperature schedule that drives it (right): (A, B) a-C, (C, D) a-Si, (E, F) silica, (G, H) ZrNiAl. Curves are medians over ten replicas, bands min–max; each curve stops at its slowest replica. The step counter counts attempted moves; a point is logged only when a move is accepted.

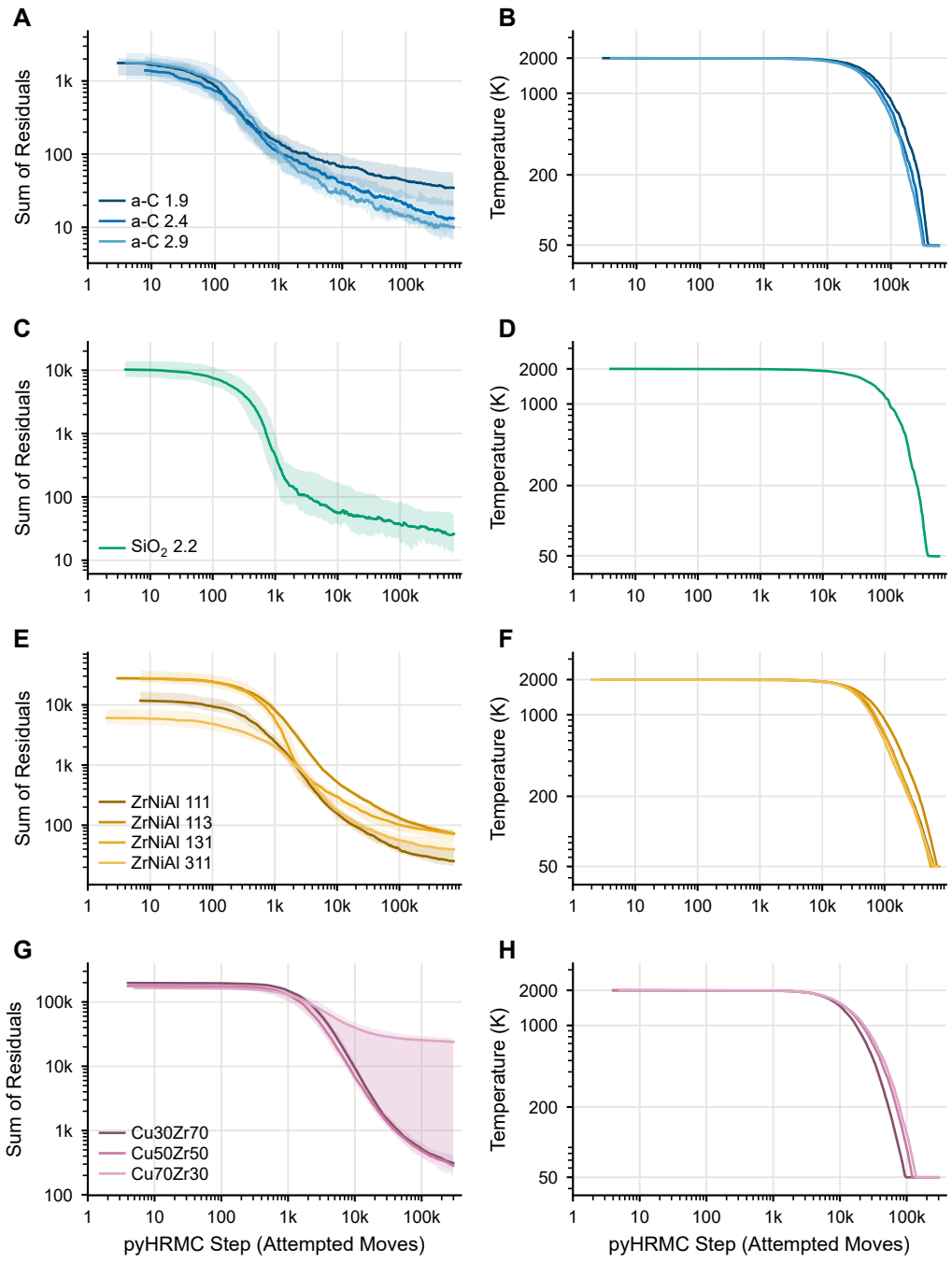


Figure 5: **HRMC liquid-start and relaxed-wall arms.** Sum of residuals (left) and temperature (right): (A, B) a-C, (C, D) silica, and (E, F) ZrNiAl from liquid seeds; (G, H) CuZr from a random start at the 0.80 wall. Conventions as in Figure 4.

C.4 CRN Training Details

We start from a four-coordinated silicon network in the condition's cell, 216 atoms for a-Si and a 100-atom former for silica, and apply Wooten–Winer–Weaire bond transpositions (Wooten et al., 1985) in the cycle of Barkema & Mousseau (2000): each cycle is an anneal of 100 trial transpositions per atom, accepted by Metropolis at $k_B T = 0.25$ eV on the Keating energy (Keating, 1966) ($\alpha = 2.965$ eV/Å^2, $\beta = 0.285\alpha$, and $d = 2.35$ Å at the diamond density, rescaled as $\rho^{-1/3}$ to the

condition's density), followed by a zero-temperature quench of 18 attempts per atom. The run stops when the per-cycle drift between two consecutive 10-cycle means of the energy falls below 5×10^{-4} of the total energy drop so far, capped at 150 cycles; this took 26 to 57 cycles for a-Si and 23 to 63 for the silica formers. Four-membered rings are removed at the end. A structure with any pair closer than 0.95 of the training-set $g(r)$ onset is redrawn with a new seed, which happened three times for two structures, both at a-Si 2.8 g/cm^3. For silica, we place a bridging oxygen on every Si–Si bond of the former and minimize with the Munetoh Tersoff potential. We ran ten seeds per condition. The median Keating energy falls from 0.62–0.79 to 0.28–0.40 eV/atom on a-Si and from 1.28 to 0.67 eV/atom on the silica formers, and the share of accepted transpositions from about 2 percent to 0.06–0.22 percent (Figure 6). The production runs did not log per-cycle energies, so these curves come from replays of the same seeds; every replay reproduces its shipped structure and final energy exactly.

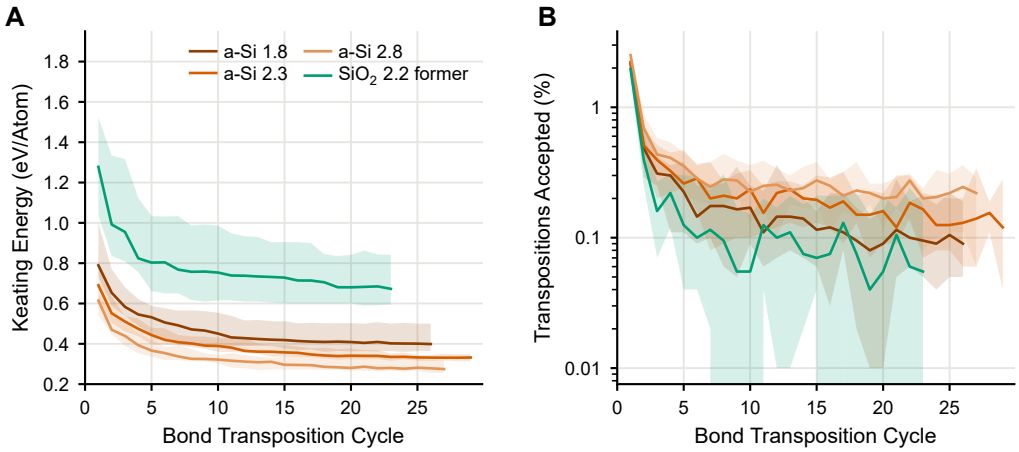


Figure 6: **CRN convergence. (A)** Keating energy per atom and **(B)** share of accepted bond transpositions per anneal–quench cycle. Curves are the median over the 10 replicas of each condition, bands the min–max, drawn to the last cycle every replica reaches. The silica curve is the 100-atom silicon former network before oxygen decoration.

C.4.1 Coordination Under the Benchmark Cutoff

Every atom of the CRN bond network is exactly four-fold. Measured with the benchmark's per-structure $g(r)$ cutoff, 99.5 percent of a-Si atoms at 2.3 g/cm^3 are four-fold, against 85.9 percent for MD, and any fixed cutoff between 2.67 and 2.74 Å recovers the four-regular network exactly (Figure 14, bottom). CRN's structures are therefore more idealized than the reference, not less, and its coordination rows measure that sharpness.

C.5 GLASS Training Details

GLASS (Guo & Schwalbe-Koda, 2026) is a score-based diffusion model (Song et al., 2021) over atomic positions in a periodic cell with a variance-exploding noise schedule, $\sigma(t) = 0.8\,t$. We trained one model per system on all of its conditions pooled and kept the checkpoint at the minimum loss on the pooled validation set (15 structures; 5 for silica, 20 for ZrNiAl; Table 7). The training settings are the defaults from Guo & Schwalbe-Koda (2026), except silica's learning rate (Appendix C.5.1), and were not swept. Every structure starts from a Poisson-disk packing at the condition's density and composition, with a minimum distance of 0.90 (covalent) or 0.80 (metallic) of the shortest pair in the training structures. GLASS-u uses the same checkpoints and sampler without PDF or Tersoff guidance.

C.5.1 Trivial-Solution Collapse

The score network can collapse to the trivial solution, a zero score everywhere, at which the denoising loss sits at $\mathbb{E}\|\epsilon\|^2 = 3$ exactly. The five benchmark models reached validation losses of 0.85 to

Table 7: **GLASS settings.** One value applies to all five systems unless a system is named. Training (top) and sampling (bottom); the selected epoch is each model's validation-loss minimum.

Setting	Value
Message-passing layers	5
Hidden width	200
Cutoff (Å)	5
Noise scale k	0.8
Learning rate	0.001 (SiO_2 0.0003)
Batch size	12
Dataset duplication	4
EMA decay	0.9999
Epochs	12,000
Train / validation	105 / 15 (SiO_2 35 / 5; ZrNiAl 140 / 20)
Selected epoch	5,633–11,961 (per system)
Sampler steps	256
Start time t	0.95
End time t	0.005
Corrector steps × size	2 × 0.5
Restarts	3
PDF guidance ρ	500
Tersoff λ_0	0.30 (a-C, SiO_2); none otherwise

1.60 (Figure 7). Notably, silica's validation loss bottoms at epoch 5,633 of 12,000 and rises afterwards (0.85 to 1.12), so its checkpoint comes from mid-training; the other four minima lie between epochs 9,052 and 11,961.

C.6 DM^2 Training Details

DM^2 (Yang & Schwalbe-Koda, 2026) is an equivariant NequIP (Batzner et al., 2022) network (three convolutions, hidden irreps $64 \times 0e + 32 \times 1e$, cutoff 5 Å) trained to predict the displacement added to a structure as noise (Hsu et al., 2024). We trained one model per condition from scratch on its 35 training structures, duplicated 128 times, with AdamW (Loshchilov & Hutter, 2019) at 2×10^{-4}, batch 16, and 30,000 nominal updates (107 epochs), and used the last epoch. Generation starts from a uniform random gas and runs 2,900 annealed denoising steps with σ from 1.0 to 0.001, followed by 100 clean-up steps.

C.6.1 Training Budget and Validation Split

Two properties of the released training code matter for reading DM^2's results. First, the update count is computed on the full duplicated set while training uses a 90 percent split, so the effective budget is about 27,000 updates rather than 30,000. Second, duplication happens before the 90/10 split, so copies of one structure fall on both sides and the validation loss is not held out; no model selection depends on it, since the last epoch ships. At 107 epochs DM^2 is the only one of the three models in Figure 7 whose loss is still visibly descending at the end of its budget, so its entries are, if anything, under-trained rather than over-fit.

C.7 CDVAE Training Details

CDVAE (Xie et al., 2022) encodes a structure with DimeNet++ (Gasteiger et al., 2020) (width 128, four blocks) into a 256-dimensional latent vector, predicts the atom count, lattice, and composition from it, and decodes positions and atom types with a GemNet-T (Gasteiger et al., 2021) decoder (width 128) by denoising over 50 levels (σ from 10 to 0.01 for positions and 5 to 0.01 for types); the KL weight is 0.01 and the radius graph 7 Å with 20 neighbors. We trained one model per

condition for 1,000 epochs with Adam (Kingma & Ba, 2015) at 10^{-3}, batch 8, gradient clipping at 0.5, and a learning rate reduced by 0.6 after 30 epochs without improvement (floor 10^{-4}), keeping the checkpoint with the lowest validation loss.

C.7.1 CHECKPOINT SELECTION

CDVAE's validation loss oscillates by a factor of several between consecutive epochs, and the minimum-loss checkpoint fires on whichever single epoch dips lowest. Against the median of the 5 percent of epochs around it, every selected CDVAE checkpoint sits 2.43 to 3.34 times lower (2.80 on average), and every curve ends 1.8 to 4.7 times above its checkpoint; for GLASS and DM^2 the same ratio is 1.01 and 1.02 (Figure 7). The CDVAE models were therefore selected on a noise dip rather than at a converged value. We scored the structures those checkpoints generated and did not reselect, but this bears on how CDVAE's standing should be read.

C.7.2 OUT-OF-COMPOSITION ELEMENTS

CDVAE predicts atom types, and its decoder can emit any element up to the configured maximum atomic number: 100 for our covalent models and 41 for the metallic ones. Of the 140 scored structures, 36 carry an element outside the training composition, 15 covalent and 21 metallic (Table 15). No other model produced a foreign element; MatterGen also generates atom types but is conditioned on the chemical system. Such a structure is a failed draw on every row that needs the potential or species identity and is scored on the eleven species-agnostic rows.

C.8 MATTERGEN FINE-TUNING DETAILS

MatterGen was originally tailored for designing crystalline materials (Zeni et al., 2025). Because amorphous space is complex and MatterGen is a considerably larger model (46.8M, compared to 1.2M for AMDEN and $\sim$4M for GLASS), we fine-tuned MatterGen on each amorphous system using the original Alex-MP-20 (Zeni et al., 2025) pre-trained checkpoint. We do not modify MatterGen's original training scheme, which conditioned only on the chemical system. Chemical system is represented as a multi-hot vector over all possible elements. Therefore, density conditioning is not present during training and enters only at inference through PDF guidance (Appendix D.1) and the acceptance bucket filter described above. Unconditional MatterGen generation is not reported in the primary text because it was not designed to deliberately aim at a density bucket.

C.8.1 SCRATCH TRAINING VERSUS FINE-TUNING

As a separate, exploratory question from the primary fine-tuned entry, we trained MatterGen on amorphous silicon from scratch (no Alex-MP-20 (Zeni et al., 2025) pre-training), using the released repository's own scratch-training configuration and learning rate (10^{-4}) rather than the fine-tuning learning rate, for 1000 epochs with no early stopping. Training converged in the sense that the loss curves flattened, but raw generated output from this checkpoint has severe atom overlaps (minimum interatomic distances of 0.50–0.53 Å, comparable to the 0.08 Å MatterGen's own paper reports for unrelaxed samples) and degenerate, unreduced cells (extreme skew angles and cell dimensions), consistent with the absence of a post-generation relaxation step for this model in our pipeline. We treat this as a known limitation of using raw scratch-trained MatterGen output rather than evidence that scratch training itself is worse than pre-trained fine-tuning; the two initializations have not yet been compared under matched relaxation or scoring, and we do not report scratch MatterGen in the main benchmark for this reason.

C.8.2 POSITIONAL NOISE-SCALE SENSITIVITY

MatterGen's fractional-coordinate diffusion adjusts its noise magnitude for cell size via a positional $\sigma_{\max}$ parameter (default 5.0 in the released configuration). This parameter is fixed by the pre-trained checkpoint during fine-tuning. Therefore, overriding it at fine-tuning time has no effect, so it is only a meaningful hyperparameter for scratch training. We swept $\sigma_{\max} \in \{1.5, 2.5, 5.0, 10.0\}$ for amorphous silicon scratch training, all four checkpoints trained to the full 1000 epochs.

Table 8 reports an assessment of 10 unguided samples (chemical-system conditioning only) on the four scratch checkpoints plus the fine-tuned checkpoint used as the primary benchmark entry.

We checked for minimum interatomic distance and cell degeneracy. The atom-overlap and cell-degeneracy pathology is monotonic in $\sigma_{\max}$. 1.5 has physically reasonable minimum distances and only 3 of 10 samples with a cell-length degeneracy, whereas 2.5 already shows both overlap and cell-degeneracy problems in the majority of samples. 5.0, the default of Zeni et al. (2025), generates 9 of 10 degenerate samples. 3 of the 9 collapse to a fully singular, zero-volume cell, leaving only 7 of 10 with a measurable minimum distance. 10.0 does not produce usable output at all, producing fully degenerate, zero-volume cells. This shows that MatterGen's paper-described noise-magnitude mechanism has a real, sweepable effect on the known scratch-training pathology for silicon. While reducing $\sigma_{\max}$ increases the quantity of non-degenerate samples, it remains unclear whether MatterGen amorphous scratch training can surpass crystalline pre-trained, amorphous fine-tuned in the small data, amorphous regime.

Table 8: **MatterGen scratch-training $\sigma_{\max}$ sweep, amorphous silicon.** 10 unguided samples per checkpoint, chemical-system conditioning only. The fine-tuned row is the pre-trained-initialization primary-benchmark checkpoint (Section 3), included for comparison. The mean minimum interatomic distance of $\sigma_{\max}$ 5.0 is skewed because 3 of 10 samples had a fully singular, zero-volume cell. 10.0 only generated samples with fully singular, zero-volume cells.

$\sigma_{\max}$	Min. interatomic distance	Degenerate cells (of 10)
1.5	0.52–2.24 Å (mean 1.75)	3
2.5	0.23–0.85 Å (mean 0.44)	7
5.0 (model default)	0.18–1.14 Å (mean 0.40)	9
10.0	–	10
Fine-tuned	2.20–2.26 Å (mean 2.23)	0

C.8.3 Fine-tuning Hyperparameter Sweep

Because MatterGen was originally designed for inorganic crystals, we wanted to confirm that it was not unfairly penalized by an unfavorable fine-tuning configuration in this amorphous setting. We therefore performed a fine-tuning hyperparameter sweep for MatterGen across every system, varying learning rate (1×10^{-5} to 6×10^{-5}) and model training seed (0, 1, 2), evaluated by validation loss at the final trained epoch. For most (system, seed) combinations, learning rate had little effect on final-epoch validation loss (under 3% best-to-worst spread). The one exception was CuZr's seed 0, which was bimodal rather than smoothly learning-rate-dependent, converging well at three of the six learning rates tested and stalling at the other three. The training seed's effect was larger and more consistent than the learning rate across every system (42.0%–58.7% spread). The primary benchmark uses, for each system, the best learning rate and seed combination.

Figure 8 shows this directly for each system. within each system, the three learning-rate points move far less than the three seeds for every seed except CuZr's seed 0, which jumps between two distinct outcomes depending on learning rate rather than tracking a trend. Which seed ranks best is system-dependent. Seed 0 gives the lowest final-epoch validation loss for Si, CuZr, and ZrNiAl, while SiO_2 instead favors seed 1.

C.9 AMDEN Training Details

AMDEN's primary entry is trained from scratch per system, uses a fixed per-system ghost-atom capacity ceiling ρ_{target} (0.159938 atoms/Å^3 for C, 0.066042 for Si, 0.072767 for SiO_2, 0.072835 for CuZr, 0.075408 for ZrNiAl) and a density property that conditions how many of those ghost slots materialize toward a requested target (see Appendix A.8). The ceiling is fixed once per system at $1.10\times$ that system's maximum training-set number density, so every condition is padded generously enough that the pad-only mechanism never needs to remove atoms from any real training or generation structure. The ceiling itself does not vary with the requested target and so does not by itself steer how many ghost slots convert to real atoms (Appendix E.5.2). The other training hyperparameters used were fixed interaction cutoff $r_{cut} = 7.5$Å, $\sigma_{\max,\text{pos}} = 1.5$, and learning rate 10^{-4}. We trained it per system for a fixed 1000 epochs on the training structures, sampled with 200 steps, and did not use PDF guidance, property conditioning, or the authors' released checkpoints. Draws were

accepted into a condition under the same 0.25 g/cm^3 window and 200-draw budget as MatterGen; it filled 10 of 10 structures in 13 of the 14 conditions, with the ZrNiAl 1:1:3 composition yielding eight. The conditional entry, AMDEN-c, adds PDF guidance to the same per-system checkpoints, with the PDF of a single validation structure as the target, guidance settings selected per density on validation-PDF RMSE, and the acceptance protocol unchanged; at the ZrNiAl 1:1:3 composition it accepted 6 of its 200 draws.

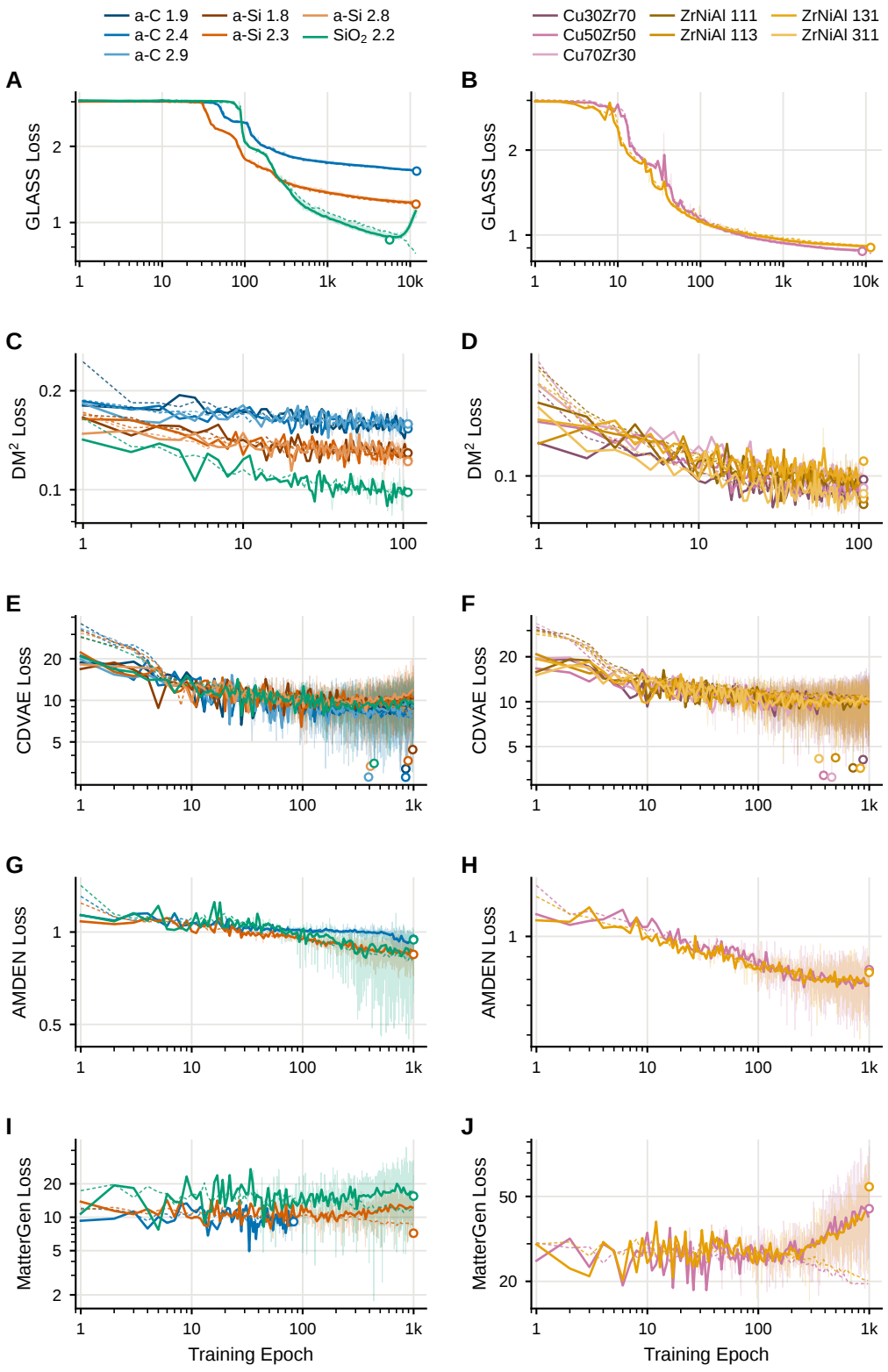


Figure 7: **Training convergence.** Validation loss against epoch for (A, B) GLASS, (C, D) DM^2, (E, F) CDVAE, (G, H) AMDEN, and (I, J) MatterGen; covalent systems and silica left, metallic right. Bold lines are the binned median of the per-epoch loss (faint), dashed lines the training loss, and circles the shipped checkpoint. GLASS, AMDEN, and MatterGen train one model per system; DM^2 and CDVAE train one per condition. AMDEN and MatterGen ship their final epoch rather than a validation-loss minimum, so their circles sit at the end of the curve, not its floor; MatterGen's carbon curve is the best (learning rate, seed) combination found by its fine-tuning sweep (Section C.8.3), the other four systems and AMDEN use each system's single fixed-budget run.

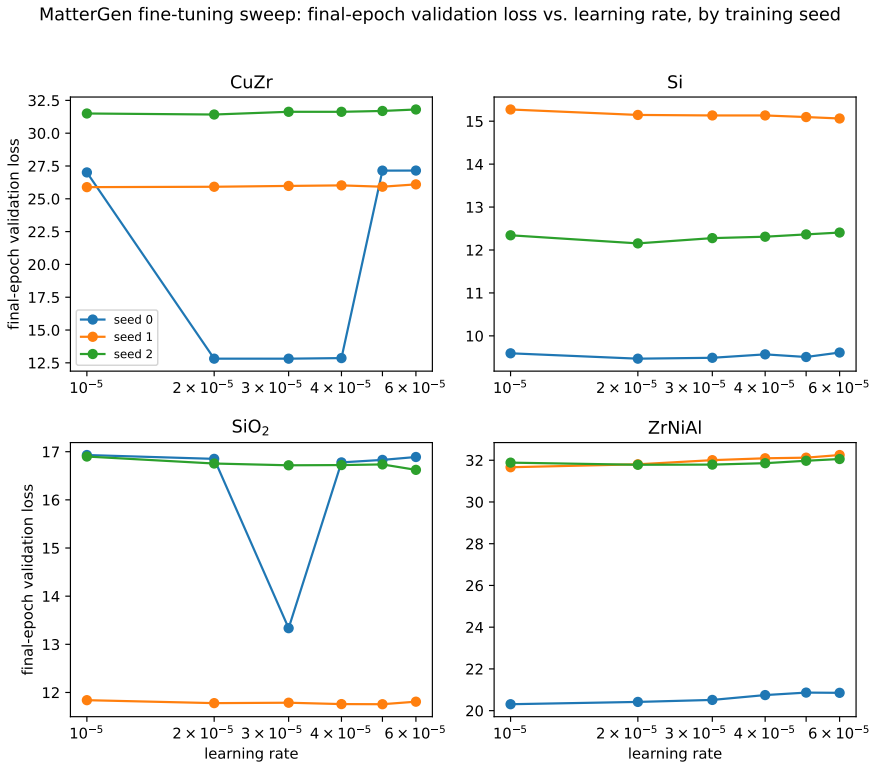


Figure 8: **MatterGen fine-tuning sweep.** Validation loss at the final trained epoch vs. learning rate, colored by training seed, one panel per system. Nearly flat lines within a seed indicate learning rate has little effect; vertical separation between seeds is the larger effect, though which seed is best varies by system. CuZr's seed 0 (panel 1) is the exception to the flat-line pattern: it jumps between two outcomes depending on learning rate rather than showing a smooth trend.

D PDF-GUIDANCE

D.1 PDF-GUIDANCE DETAILS

PDF guidance is an inference-time, gradient-based conditioning mechanism native to GLASS (Guo & Schwalbe-Koda, 2026). We evaluate its benefit for DM^2, AMDEN, and MatterGen. We swept its two hyperparameters, guidance strength ρ and smoothing σ, for two of the four models that support it. Guidance strength ($\rho \in \{10, 30, 100, 300, 1000, 3000, 10000\}$) and smoothing ($\sigma \in \{0.05, 0.10, 0.15, 0.20\}$) were swept per system and density, 28 settings per condition, against a held-out validation structure's $g(r)$, for MatterGen and AMDEN only (selection ranking: cutoff 8Å, 200 bins, per-point χ^2 smoothing 0.1; generation guidance itself: cutoff 8Å, 100 bins, the selected σ). DM^2 used a separate, smaller sweep over ρ alone, with no σ grid (Appendix D.3), and GLASS's guidance settings were not swept in this work at all, carried over unchanged from Guo & Schwalbe-Koda (2026) (Appendix D.2). The selected setting is the lowest observed mean PDF RMSE among evaluated grid points, not a proven global optimum, and several selections sit at the boundary of the explored grid (e.g. most AMDEN conditions select the largest evaluated $\rho = 10000$). Low accepted-count winners are annotated explicitly (e.g. MatterGen CuZr at 7.0 g/cm^3 selects its winning setting from only 2 accepted structures). MatterGen's amorphous carbon, silicon and silica settings were carried over from an earlier sweep that selected against test-split references rather than validation.

Table 9 lists every selected setting. AMDEN's 14 selections are uniformly validation-based and uniformly at full accepted count ($n = 10$).

Table 9: **Selected PDF guidance settings per condition (Section D.1).** n is the accepted-structure count the sweep's own RMSE ranking was computed from at that setting, not the final benchmark's accepted count (Tables 18–22/Figure 15). "–" marks carbon at 2.9 g/cm^3, whose sweep had zero accepted structures at every evaluated setting.

System / density	ρ	σ	n	PDF RMSE
AMDEN				
C 1.9	3000	0.10	10	0.1645
C 2.4	3000	0.10	10	0.1296
C 2.9	3000	0.10	10	0.1264
Si 1.8	10000	0.20	10	0.1543
Si 2.3	10000	0.15	10	0.1087
Si 2.8	1000	0.05	10	0.0586
SiO_2 2.2	3000	0.10	10	0.1265
CuZr 7.0	10000	0.10	10	0.0431
CuZr 7.4	10000	0.10	10	0.0538
CuZr 7.9	10000	0.05	10	0.0622
ZrNiAl 4.6	10000	0.05	10	0.0439
ZrNiAl 5.8	10000	0.05	10	0.0283
ZrNiAl 6.1	10000	0.05	10	0.0377
ZrNiAl 6.7	10000	0.05	10	0.0484
MatterGen				
CuZr 7.0	300	0.10	2	0.0288
CuZr 7.4	1000	0.15	10	0.0240
CuZr 7.9	3000	0.20	7	0.0227
ZrNiAl 4.6	30	0.05	1	0.0331
ZrNiAl 5.8	3000	0.20	10	0.0584
ZrNiAl 6.1	1000	0.20	10	0.0220
ZrNiAl 6.7	3000	0.20	10	0.0327
C 1.9	300	0.10	3	0.0386
C 2.4	1000	0.05	3	0.0352
C 2.9	1000	0.05	–	–
Si 1.8	10	0.05	3	0.0861
Si 2.3	1000	0.15	1	0.0730
Si 2.8	10000	0.20	3	0.0308
SiO_2 2.2	3000	0.15	3	0.0417

D.2 GLASS PDF-GUIDANCE DETAILS

GLASS-c adds a PDF term to the score at every predictor step. From the noisy positions $\mathbf{x}_t$ it forms the Tweedie estimate (Efron, 2011) of the clean structure, $\hat{\mathbf{x}}_0 = \mathbf{x}_t + \sigma(t)^2 s_\theta(\mathbf{x}_t, t)$, computes its partial pair distribution functions $g_{ab}(r)$ with a differentiable Gaussian-kernel histogram (width 0.15 Å, 100 bins to the 5 Å graph cutoff), and compares them with the mean over the condition's 35 training structures, $\bar{g}_{ab}(r)$:

$$s_{\text{PDF}} = -\frac{\rho}{\sum_{ab}\|g_{ab}(\hat{\mathbf{x}}_0) - \bar{g}_{ab}\|} \nabla_{\hat{\mathbf{x}}_0} \frac{1}{P} \sum_{ab} \|g_{ab}(\hat{\mathbf{x}}_0) - \bar{g}_{ab}\|^2, \tag{1}$$

with the network held fixed and P the number of species pairs. The normalization makes ρ a step size rather than a multiplier on an unbounded gradient, and the term enters the reverse-SDE predictor with the same $g^2(t)\,dt$ weight as the learned score; the Langevin corrector steps carry no PDF term. For a-C and silica, a Tersoff term $-\lambda(t)\nabla E(\hat{\mathbf{x}}_0)/N$ is added in both predictor and corrector, with $\lambda(t) = 0.30\,\text{sigmoid}(-20(t-0.5))$ so that it switches on in the low-noise half of the trajectory, clamped per atom at 10. Hyperparameters used to impose such guidance are identical to the best ones from Guo & Schwalbe-Koda (2026) (see also Section D.3.1).

D.3 DM^2 PDF-GUIDANCE DETAILS

DM^2-c uses the same guidance term, evaluated on DM^2's denoised estimate with the network held fixed, and applies it only during the 2,900 annealed steps as $\mathbf{x} \leftarrow \mathbf{x} + (\sigma_i^2 - \sigma_{i+1}^2)\, s_{\text{PDF}}$, the discrete analogue of $g^2(t)\,dt$; the 100 clean-up steps are unguided. The target is again the mean PDF of the 35 training structures (width 0.15 Å, 100 bins to 5 Å). Because DM^2 takes about twelve times as many steps as GLASS, GLASS's $\rho = 500$ does not carry over. We swept $\rho \in \{2000, 10000, 20000\}$ plus the unguided sampler, three structures per arm, at one condition per system (a-C 2.4, a-Si 2.3, silica 2.2, $Cu_{50}Zr_{50}$ 7.4, ZrNiAl 1:1:1 5.8), and chose on the $g(r)$ error and mean coordination against the training mean: $\rho = 10000$ for a-C, CuZr, and ZrNiAl, and 20000 for a-Si and silica.

D.3.1 SWEEP OF TRADE-OFF BETWEEN PRIOR AND GUIDANCE

Because the sweep chose ρ against the same training mean the guidance targets, we rescored it against the test structures and added $\rho \in \{10, 100, 500\}$ (Figure 9). Held-out scoring changes no choice: the ranking of the three swept values by $g(r)$ χ^2 is identical against training and test references in all five systems. Guidance is inert below $\rho \approx 2000$ (a-C: 0.46 unguided and 0.47, 0.45, 0.45 at $\rho = 10$, 100, 500, against 0.056 at the chosen 10000), and $\rho = 500$, GLASS's value, is a local worst case on silica, CuZr, and ZrNiAl (silica 1.10, more than twice the unguided 0.48). On the metallic systems, the coordination-outlier row prefers a different ρ than $g(r)$, but the gaps are within draw noise and every CuZr arm of the original sweep sits below MD's own leave-one-out floor, so we kept the $g(r)$ choice. The one tradeoff the sweep does show is a-C at $\rho = 20000$, whose coordination outliers (20 per 1,000 atoms) are 3.3 times MD's floor while its $g(r)$ is barely worse than at 10000; at three structures per arm we read this as suggestive.

D.4 MATTERGEN PDF-GUIDANCE DETAILS

MatterGen's guidance settings (Table 9) come from two different sweeps. CuZr and ZrNiAl were selected from the same validation-split sweep as the other models. Amorphous carbon, silicon, and silica were instead carried over from a parameter sweep scored against test-split references (Appendix D.1), and coverage in that study is uneven: carbon at 2.9 g/cm^3 has zero accepted structures at every one of the 28 evaluated settings, consistent with MatterGen's broader failure to reach that condition unguided (Section 2.3), and several other conditions (amorphous silicon in particular, and ZrNiAl 4.6 g/cm^3) have only a handful of the 28 settings represented rather than the full grid. Where data exists, the selected ρ varies widely by condition (from 10 to 10000) with no single dominant value, unlike AMDEN's consistent preference for the largest evaluated ρ (Appendix D.5).

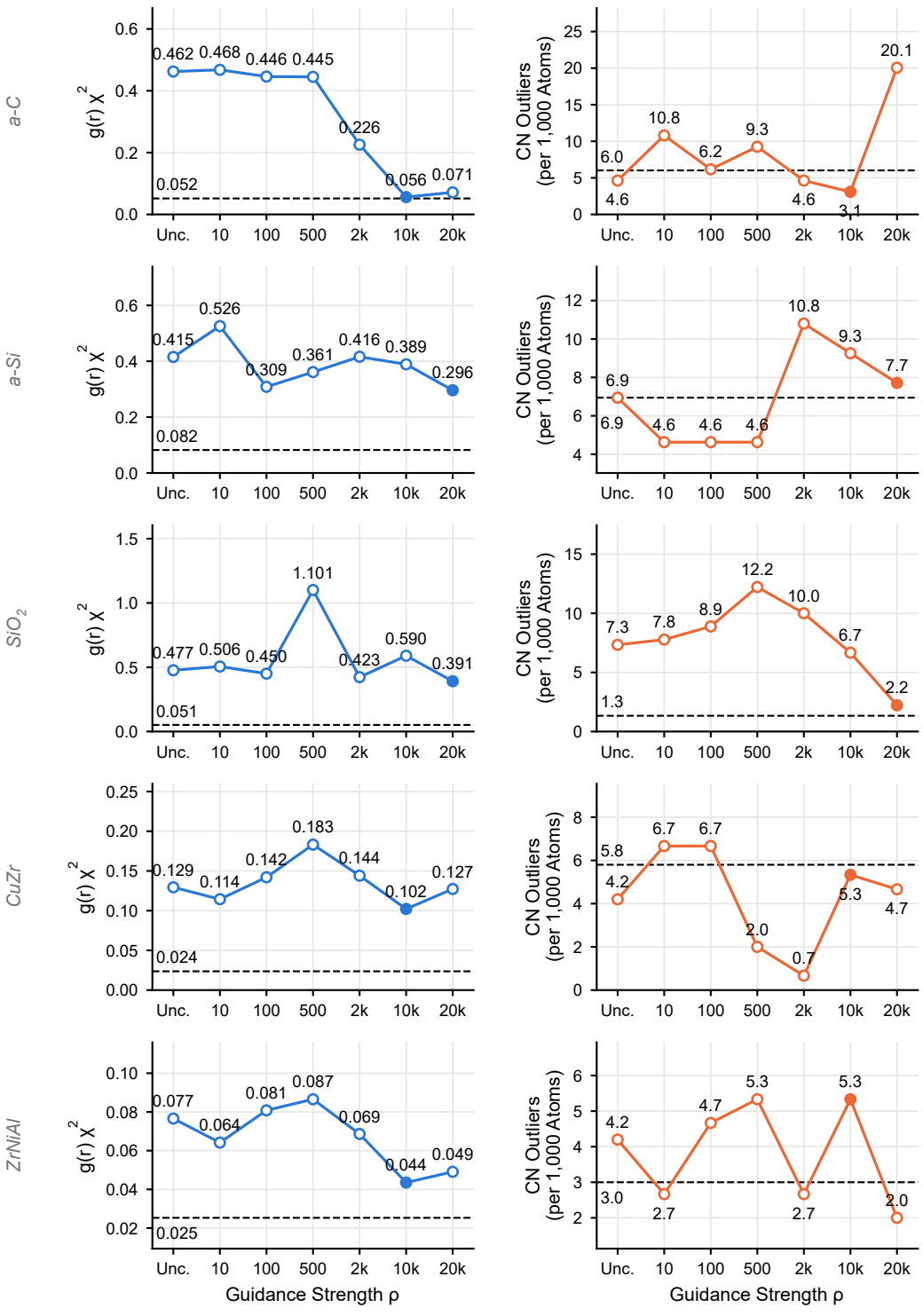


Figure 9: **DM^2 guidance-strength sweep, scored against the test structures.** $g(r)$ χ^2 (left) and coordination outliers per 1,000 atoms (right) against ρ, one row per system; three structures per arm. Filled markers are the chosen ρ, dashed lines MD's leave-one-out floor, and Unc. the unguided sampler.

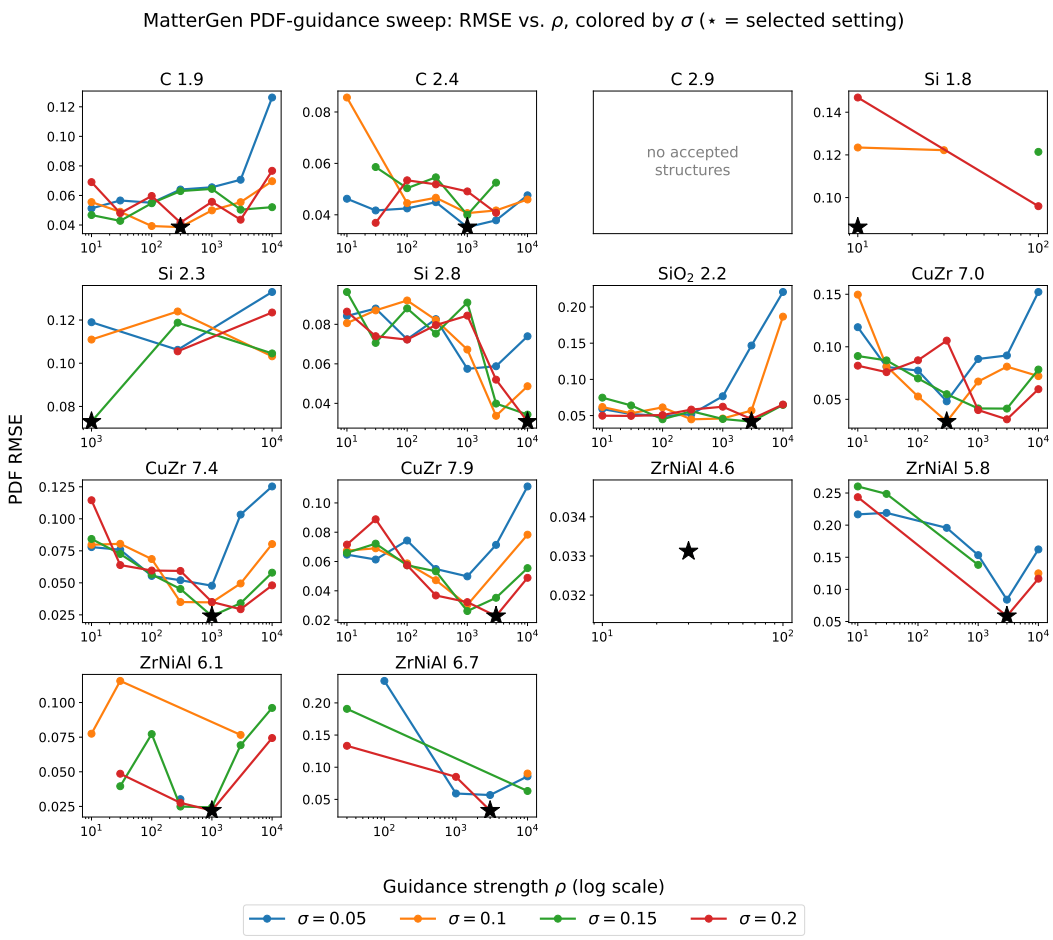


Figure 10: **MatterGen PDF-guidance sweep.** PDF RMSE vs. guidance strength ρ (log scale), colored by smoothing σ, one panel per condition. The selected setting (Table 9) is marked with a black star. Carbon 2.9 g/cm^3 had zero accepted structures at any setting; ZrNiAl 4.6 g/cm^3's pilot sweep evaluated only a single setting, so only that point is shown.

D.5 AMDEN PDF-Guidance Details

AMDEN's guidance settings (Table 9) come from the full validation-split sweep with complete 28-setting coverage at every one of the 14 conditions. The selected setting is the largest evaluated guidance strength ($\rho = 10000$) at 9 of 14 conditions, the next-largest evaluated value ($\rho = 3000$) at 4, and $\rho = 1000$ at the remaining one (amorphous silicon 2.8 g/cm^3), with smoothing σ at the smallest two evaluated values (0.05 or 0.10) everywhere except amorphous silicon at 1.8 and 2.3 g/cm^3, which favor larger σ (0.15–0.20). This consistent preference for strong guidance is unlike MatterGen's noisier, condition-dependent optimum (Appendix D.4), and several selections sitting at the boundary of the explored grid means a wider ρ range might move these optima further still.

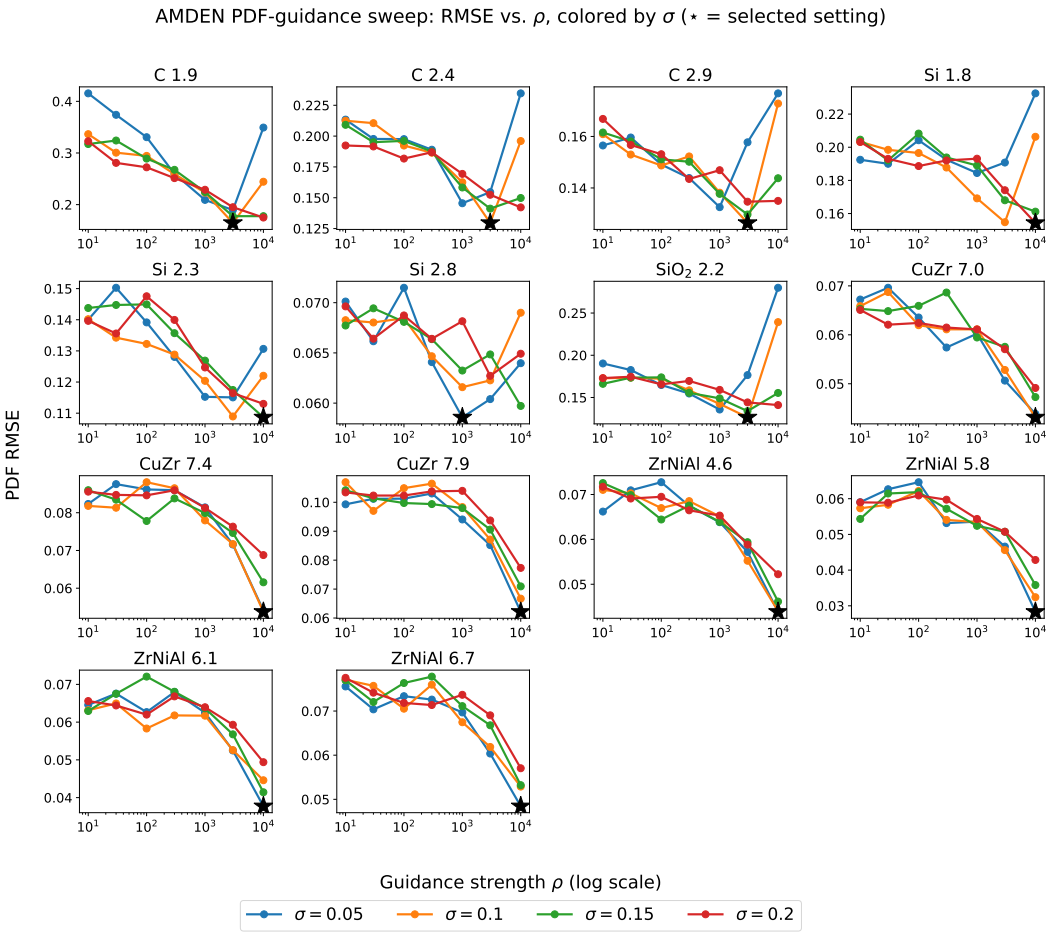


Figure 11: **AMDEN PDF-guidance sweep.** PDF RMSE vs. guidance strength ρ (log scale), colored by smoothing σ, one panel per condition (complete 28-setting coverage throughout). The selected setting (Table 9) is marked with a black star; in most panels it sits at or near the lowest-RMSE point at the largest evaluated ρ.

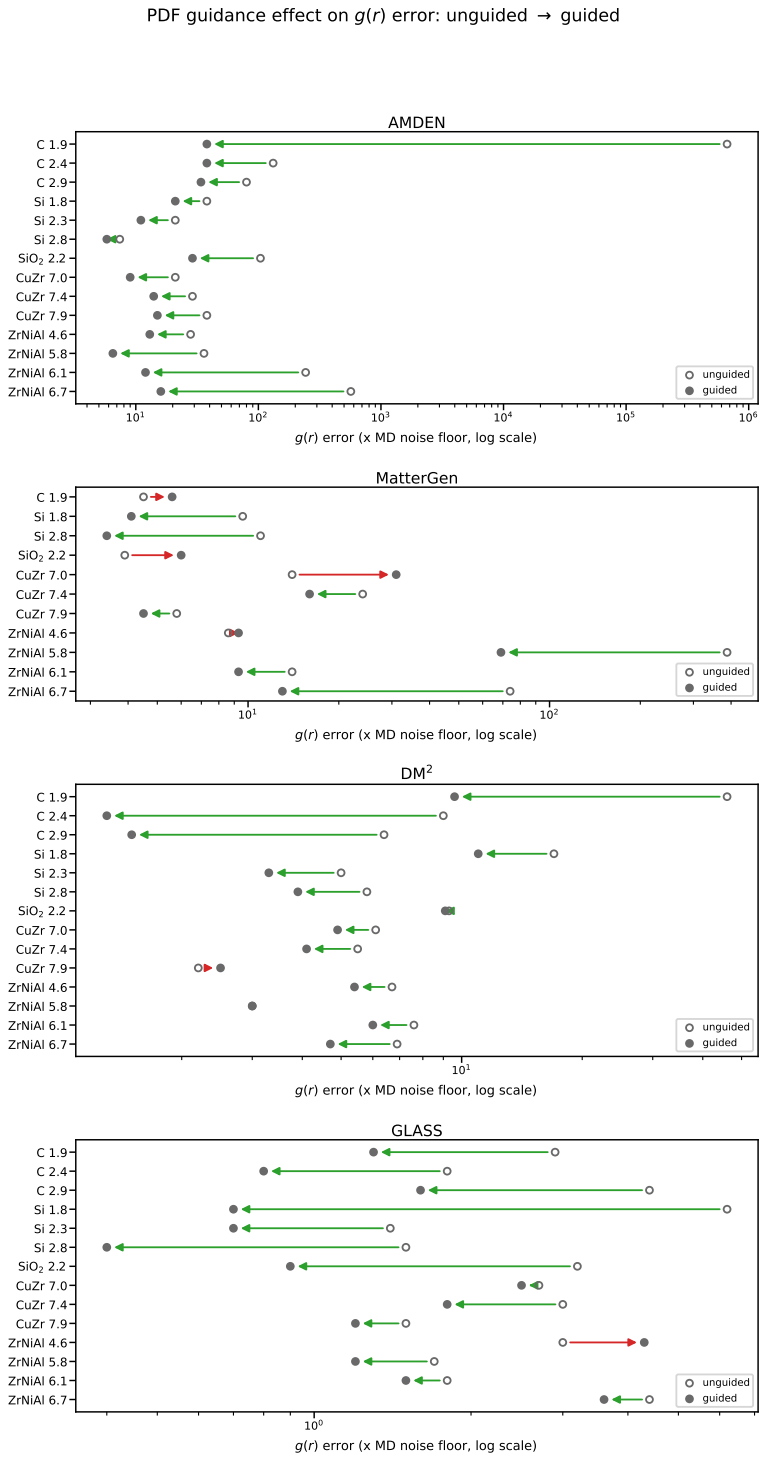


Figure 12: **PDF guidance's effect on $g(r)$ error, unguided to guided, log scale.** Hollow circles mark the unguided value, solid gray circles the guided value; the connecting arrow (pointing from unguided to guided) is green where guidance improves $g(r)$ error and red where it worsens it. From top to bottom: AMDEN, MatterGen, DM^2, and GLASS.

E FULL RESULTS

E.1 PER-GROUP COMPARISON FOR A-SI

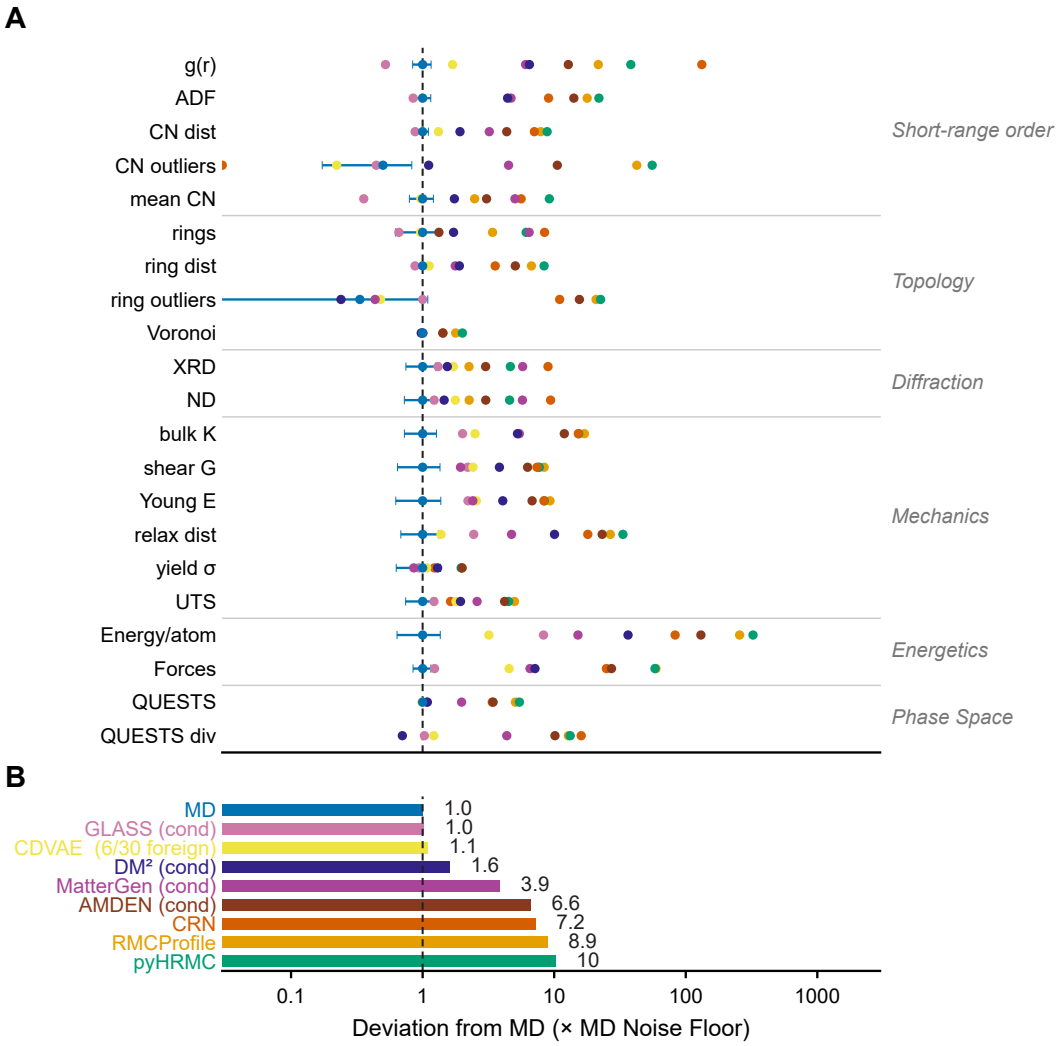


Figure 13: **Per-metric profiles grouped by family for amorphous silicon.** **(A)** Log-scale deviations from MD reference across 21 metric rows, grouped into six metric families. **(B)** Median deviation (Overall × MD) over the rows a method carries (k/N marked when short), sorted ascending. Note that ordering by overall error differs from the mean rank in Table 3 because places are assigned from pooled per-structure ranks.

Figure 13 groups the per-metric values by family. On the short-range order rows, the leading methods are inside the MD variability: on amorphous silicon at 2.3 g/cm^3 (Table 1), GLASS-c, CDVAE, and DM^2-c all sit between 0.3 and 1.0 times MD on the coordination distribution, mean coordination number, and Voronoi rows. On these rows, the differences between the top methods are smaller than MD's own leave-one-out deviation. The rows that separate them are energy per atom, forces, relaxation distance, and QUESTS divergence.

E.2 CN STUDY

Coordination depends on the neighbor rule, and for metals it depends on the rule more than on the structure. On the MD test structures, four of five common rules agree on a-Si at 2.3 g/cm^3 to within

0.22 (3.93 to 4.15; Voronoi gives 6.51), but on $Cu_{50}Zr_{50}$ they span 4.5 to 13.1 (Figure 14, top). We count neighbors within a cutoff at the first minimum of each structure's own $g(r)$, the same rule the ADF and the rings use, so every metric sees one definition of a bond and no method is favored by a global one. The cost of the rule is that a method with a diffuse $g(r)$, such as RMC, resolves its own minimum further out and is credited with a wider shell; we report that as a finding rather than pinning the cutoff to MD's.

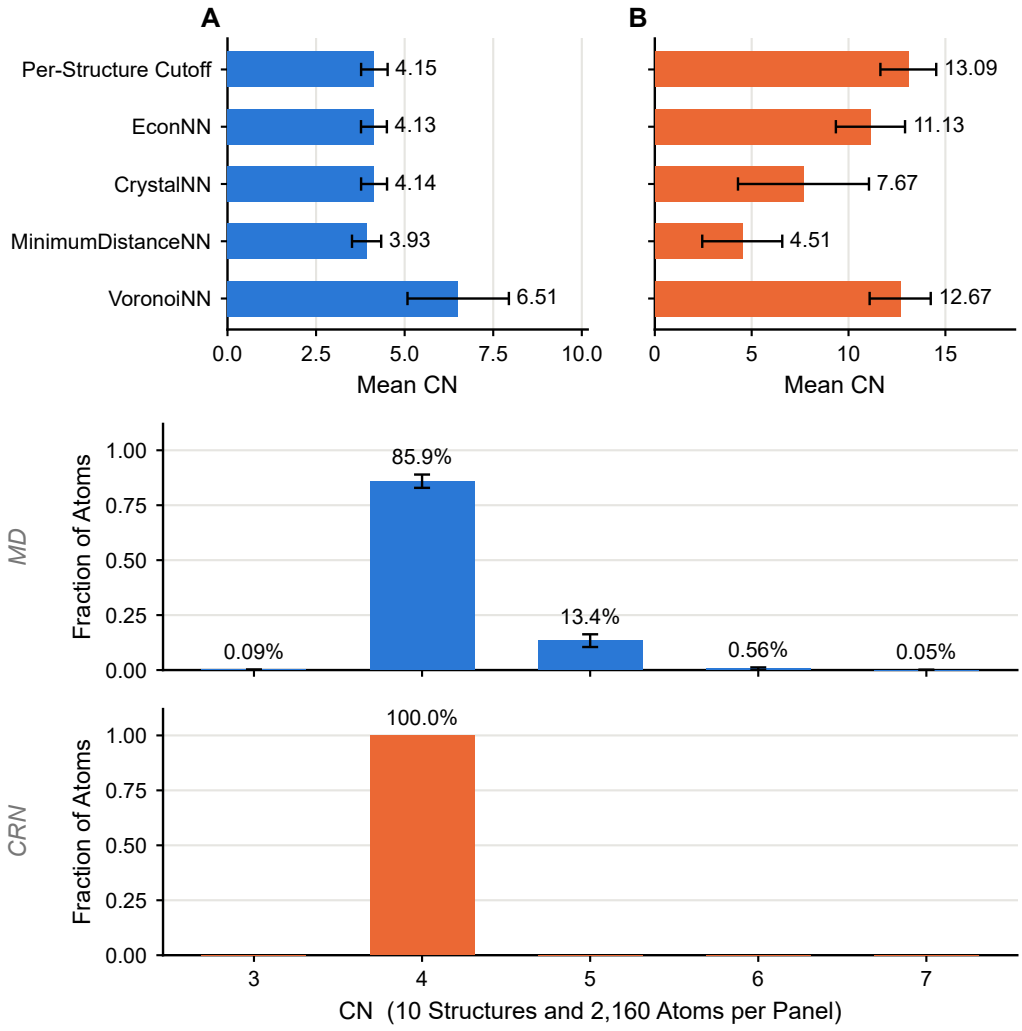


Figure 14: **Coordination rules.** Top: mean coordination of the MD test structures under five neighbor rules, (A) a-Si at 2.3 and (B) $Cu_{50}Zr_{50}$ at 7.4 g/cm^3 (per-structure cutoffs 2.94 and 3.82 Å); whiskers are the per-atom standard deviation. Bottom: per-atom coordination at a-Si 2.3 g/cm^3, MD under the benchmark cutoff against CRN's bond network.

E.3 Full Per-System Tables

Tables 10–14 extend the main tables with MD, every unconditional arm, and the two HRMC variants (HRMC-L, the liquid start; HRMC-w80, the 0.80 wall on CuZr). A cell is the same number in the main and full tables because it measures one method against MD; the mean rank is recomputed within the columns shown, so a method's place can differ between the two. A header marked k/N runs belongs to an arm generated at only some of the system's conditions. The per-condition tables are released with the benchmark.

1997 1996 1995 1994 1993 1992 1991 1990 1989 1988 1987 1986 1985 1984 1983 1982 1981 1980 1979 1978 1977 1976 1975 1974 1973 1972 1971 1970 1969 1968 1967 1966 1965 1964 1963 1962 1961 1960 1959 1958 1957 1956 1955 1954 1953 1952 1951 1950 1949 1948 1947 1946 1945 1944

Table 10: **Full metric profile, amorphous carbon.** Medians normalized by MD's leave-one-out noise floor, averaged over the system's conditions; ± is the bootstrap standard error; bold is the lowest value per row; * below one minus MD's bootstrap band; ‡ half or more of the draws failed; *fail* no evaluable structure. Mean rank is recomputed over the columns shown.

Metric	MD (n=10)	RMC (n=10)	HRMC (n=10)	HRMC-L (n=10)	GLASS-c (n=10)	GLASS-u (n=10)	DM2-c (n=10)	DM2-u (n=10)	CDVAE (n=10)	MatterGen-c (n=10 · 2/3 runs)	MatterGen-u (n=10 · 1/3 runs)	AMDEN-c (n=10)	AMDEN-u (n=10)
g(r)	1.0 ±0.15	33 ±0.3	225 ±52	24 ±0.5	**1.2 ±0.2**	2.9 ±0.6	5.1 ±1.0	18 ±4.1	3.0 ±211	4.3 ±0.9	3.9 ±0.8	38 ±2.4	3855 ±40421
ADF	1.0 ±0.20	97 ±1.0	102 ±3.4	18 ±1.0	3.7 ±0.3	**2.7 ±0.3**	23 ±2.0	30 ±3.3	2.9 ±5.6	3.5 ±0.5	3.2 ±0.4	77 ±1.2	128 ±10.0
CN dist	1.0 ±0.22	18 ±0.5	22 ±0.9	5.2 ±0.4	1.5 ±0.2	**1.3 ±0.3**	4.6 ±2.4	11 ±4.0	1.6 ±1.4	3.9 ±0.2	3.5 ±0.6	14 ±0.4	28 ±1.8
CN outliers	0.3 ±0.44	70 ±2.9	121 ±12	5.0 ±1.5	1.3 ±0.3	1.0 ±0.3	2.5 ±18	38 ±31	0.3 ±7.6	3.8 ±0.5	**0.0 ±0.0**	37 ±2.9	163 ±20
mean CN	1.0 ±0.39	6.2 ±0.3	56 ±29	6.8 ±0.8	**1.6 ±0.3**	1.6 ±0.6	8.8 ±23	57 ±39	1.6 ±6.3	4.6 ±0.6	5.0 ±1.6	18 ±3.2	165 ±24
rings	1.0 ±0.37	4.3 ±0.3	11 ±4.7	3.9 ±0.6	1.1 ±0.3	**0.9 ±0.3**	2.3 ±1.7	5.0 ±2.9	1.7 ±0.8	2.3 ±0.3	2.4 ±0.8	1.7 ±0.7	14 ±2.3
ring dist	1.0 ±0.12	8.5 ±0.1	9.5 ±0.3	2.0 ±0.1	**1.2 ±0.2**	1.4 ±0.1	2.7 ±0.2	3.6 ±0.3	1.5 ±0.4	1.3 ±0.1	1.7 ±0.2	5.9 ±0.1	7.6 ±0.2
ring outliers	0.3 ±0.12	118 ±2.2	127 ±5.5	6.0 ±0.9	0.9 ±0.2	**0.9 ±0.2**	14 ±3.1	13 ±5.9	1.0 ±5.6	1.7 ±0.4	2.9 ±1.4	69 ±2.2	64 ±19
Voronoi	1.0 ±0.04	1.9 ±0.0	2.1 ±0.0	1.0 ±0.0	0.9 ±0.0*	0.9 ±0.0*	0.9 ±0.0*	0.9 ±0.0*	0.9 ±0.0*	**0.9 ±0.0***	1.0 ±0.0*	1.5 ±0.0	1.5 ±0.0
XRD	1.0 ±0.39	1.1 ±0.3	11 ±1.5	3.1 ±0.4	**0.8 ±0.3**	1.2 ±0.4	1.1 ±0.5	1.9 ±0.4	2.2 ±8.3	2.8 ±0.8	2.2 ±0.6	5.4 ±0.7	11 ±1.1
ND	1.0 ±0.38	1.8 ±0.3	11 ±1.5	3.1 ±0.4	**0.9 ±0.3**	1.4 ±0.4	1.1 ±0.5	2.6 ±0.6	2.3 ±19	2.9 ±0.9	1.7 ±0.6	5.8 ±1.0	18 ±2.0
bulk K	1.0 ±0.40	fail	11‡	4.3 ±0.7	2.9 ±0.7	2.3 ±0.6	5.4 ±0.7‡	6.7 ±0.8‡	**2.0 ±0.3**	8.2 ±1.4	2.3 ±0.6	16‡	24 ±4.5‡
shear G	1.0 ±0.36	fail	16‡	5.7 ±0.8	5.0 ±0.8	4.8 ±1.0	8.6 ±1.2‡	9.7 ±0.5‡	**4.5 ±0.7**	12 ±1.7	5.1 ±1.1	17‡	22 ±2.5‡
Young E	1.0 ±0.39	fail	16‡	6.2 ±0.8	4.9 ±0.9	4.7 ±1.0	9.3 ±1.3‡	11 ±0.5‡	**4.5 ±0.7**	13 ±1.9	5.8 ±1.4	18‡	23 ±3.0‡
relax dist	1.0 ±0.54	fail	49‡	9.9 ±0.7	**2.6 ±0.5**	2.9 ±0.5	6.7 ±1.1‡	9.4 ±0.7‡	2.9 ±0.7	4.1 ±1.4	2.7 ±0.7	43‡	46 ±3.2‡
relaxed n [converged/total]	30/30	0/30	1/30	20/30	24/30	28/30	21/30	15/30	21/30	18/20	10/10	2/30	3/30
yield σ	1.0 ±0.37	fail	9.6 ±0.6‡	1.2 ±0.4	1.0 ±0.3	**0.7 ±0.4**	2.2 ±0.7	2.8 ±0.7	1.4 ±1.1	2.0 ±0.5	1.3 ±0.4	7.1 ±0.8‡	9.0 ±0.6‡
UTS	1.0 ±0.37	fail	2.7 ±0.3‡	1.2 ±0.3	**0.5 ±0.2***	0.6 ±0.2*	0.9 ±0.3	1.4 ±0.3	0.9 ±1.4	2.0 ±0.5	1.5 ±0.3	2.6 ±0.2‡	2.6 ±0.2‡
Energy/atom	1.0 ±0.31	282 ±1.9	344 ±9.9	45 ±1.8	9.1 ±0.6	8.2 ±0.7	62 ±10	106 ±19	**7.6 ±1050**	14 ±0.8	11 ±1.3	223 ±16	954 ±147
Forces	1.0 ±0.15	80 ±0.9	96 ±1.5	42 ±1.3	4.4 ±0.5	4.5 ±0.5	36 ±1.4	44 ±1.7	10.0 ±7.8	**3.0 ±0.4**	3.5 ±0.2	73 ±1.6	87 ±1.2
QUESTS	1.0	4.9	5.1	3.9	1.3	**1.3**	1.7	1.8	1.7	1.8	1.5	4.8	4.9
QUESTS div	0.0	18	18	15	3.3	**3.0**	4.7	7.5	6.3	7.8	6.0	18	18
foreign n [count/total]	0/30	0/30	0/30	0/30	0/30	0/30	0/30	0/30	7/30	0/20	0/10	0/30	0/30
Overall (×MD)	1.0	14 (45/63)	25 (51/63)	4.7	**1.4**	1.6	3.7	5.8	1.9	3.4	2.7	16 (51/63)	26 (51/63)
Mean rank	–	8.7	10.1	5.9	2.3	2.1	4.7	5.7	4.0	4.9	4.2	8.4	9.7

2051 2050 2049 2048 2047 2046 2045 2044 2043 2042 2041 2040 2039 2038 2037 2036 2035 2034 2033 2032 2031 2030 2029 2028 2027 2026 2025 2024 2023 2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 2005 2004 2003 2002 2001 2000 1999 1998

Table 11: **Full metric profile, amorphous silicon.** Conventions as in Table 10.

Metric	MD (n=10)	RMC (n=10)	HRMC (n=10)	CRN (n=10)	GLASS-c (n=10)	GLASS-u (n=10)	DM2-c (n=10)	DM2-u (n=10)	CDVAE (n=10)	MatterGen-c (n=2)	MatterGen-u (n=2 · 2/3 runs)	AMDEN-c (n=10)	AMDEN-u (n=10)
g(r)	1.0 ±0.16	22 ±0.2	38 ±0.3	132 ±3.6	**0.5 ±0.1***	3.5 ±0.9	6.5 ±0.7	9.1 ±1.1	1.7 ±0.5	6.1 ±1.7	9.7 ±2.2	13 ±1.0	23 ±1.4
ADF	1.0 ±0.15	18 ±0.3	22 ±0.3	9.1 ±0.2	**0.8 ±0.2**	1.3 ±0.2	4.4 ±0.3	5.3 ±0.4	1.0 ±0.2	4.7 ±0.6	4.1 ±0.3	14 ±0.5	13 ±0.5
CN dist	1.0 ±0.11	7.9 ±0.3	8.8 ±0.2	7.1 ±0.2	**0.9 ±0.1***	1.6 ±0.2	1.9 ±0.3	2.4 ±0.3	1.3 ±0.2	3.2 ±0.4	3.2 ±0.3	4.4 ±0.3	4.5 ±0.2
CN outliers	0.5 ±0.33	42 ±3.0	56 ±2.9	**0.0 ±0.0**	0.4 ±0.4	0.2 ±0.2	1.1 ±0.5	1.9 ±0.5	0.2 ±0.3	4.5 ±2.1	3.5 ±2.3	11 ±1.4	14 ±1.4
mean CN	1.0 ±0.21	2.5 ±0.1	9.2 ±0.3	5.6 ±0.1	**0.4 ±0.2***	1.4 ±0.3	1.7 ±0.4	3.2 ±0.5	1.0 ±0.2	5.0 ±1.0	4.9 ±0.7	3.1 ±0.9	2.1 ±0.6
rings	1.0 ±0.38	3.4 ±0.3	6.1 ±0.4	8.4 ±0.4	**0.7 ±0.3**	1.4 ±0.5	1.7 ±0.5	2.5 ±0.5	1.0 ±0.3	6.4 ±1.7	8.2 ±1.5	1.3 ±0.7	4.5 ±0.8
ring dist	1.0 ±0.10	6.7 ±0.1	8.4 ±0.2	3.6 ±0.1	**0.9 ±0.1***	1.1 ±0.1	1.9 ±0.1	2.3 ±0.1	1.1 ±0.1	1.8 ±0.1	1.9 ±0.2	5.1 ±0.2	5.7 ±0.2
ring outliers	0.3 ±0.76	21 ±2.8	23 ±2.6	11 ±1.2	1.0 ±0.4	0.5 ±0.5	0.2 ±0.4	0.5 ±0.3	0.5 ±0.3	0.4 ±0.8	**0.1 ±0.2**	16 ±2.3	23 ±3.3
Voronoi	1.0 ±0.05	1.8 ±0.0	2.0 ±0.0	1.4 ±0.0	1.0 ±0.0	**1.0 ±0.0**	1.0 ±0.0	1.0 ±0.0	1.0 ±0.1	1.0 ±0.0	1.0 ±0.0	1.4 ±0.0	1.4 ±0.0
XRD	1.0 ±0.25	2.3 ±0.4	4.6 ±0.4	9.0 ±0.4	**1.3 ±0.2**	1.7 ±0.3	1.5 ±0.3	1.6 ±0.4	1.7 ±0.7	5.8 ±1.7	4.3 ±0.8	3.0 ±0.5	4.8 ±0.7
ND	1.0 ±0.27	2.3 ±0.4	4.6 ±0.3	9.4 ±0.4	**1.2 ±0.2**	1.6 ±0.3	1.5 ±0.2	1.5 ±0.4	1.8 ±0.7	5.7 ±1.8	5.2 ±1.0	3.0 ±0.4	4.8 ±0.7
bulk K	1.0 ±0.27	17 ±1.1	15 ±69	15 ±0.7	2.0 ±0.6	**1.8 ±0.9**	5.3 ±0.9	4.4 ±0.9	2.5 ±0.6	5.4 ±2.6	7.5 ±0.8	12 ±1.3	21 ±4.0
shear G	1.0 ±0.36	8.4 ±0.8	7.7 ±0.9	7.4 ±0.4	2.2 ±0.6	**1.6 ±0.5**	3.8 ±0.6	3.7 ±0.5	2.4 ±0.5	1.9 ±1.5	3.5 ±0.2	6.3 ±0.7	9.0 ±1.1
Young E	1.0 ±0.37	9.3 ±0.9	8.4 ±1.0	8.4 ±0.4	2.2 ±0.6	**1.9 ±0.5**	4.1 ±0.7	3.9 ±0.5	2.6 ±0.5	2.4 ±1.3	3.8 ±0.2	6.8 ±0.7	9.9 ±1.3
relax dist	1.0 ±0.32	27 ±0.7	33 ±1.2	18 ±0.4	2.4 ±0.4	2.2 ±0.5	10 ±0.7	11 ±1.0	**1.4 ±0.4**	4.7 ±1.2	3.1 ±2.0	23 ±1.1	23 ±1.3
relaxed n [converged/total]	30/30	30/30	30/30	30/30	30/30	30/30	30/30	30/30	24/30	17/17	12/12	30/30	30/30
yield σ	1.0 ±0.37	2.0 ±0.2	2.0 ±0.3	1.2 ±0.3	0.9 ±0.3	**0.8 ±0.4**	1.3 ±0.4	1.5 ±0.4	1.1 ±0.5	0.9 ±0.3	2.1 ±0.3	2.0 ±0.3	2.0 ±0.2
UTS	1.0 ±0.26	5.0 ±0.3	4.5 ±0.4	1.6 ±0.4	**1.2 ±0.4**	1.8 ±0.3	1.9 ±0.4	2.0 ±0.5	1.8 ±0.6	2.6 ±0.8	2.5 ±0.4	4.2 ±0.3	4.3 ±0.6
Energy/atom	1.0 ±0.36	257 ±4.7	325 ±4.0	83 ±4.7	8.3 ±0.9	9.5 ±1.0	36 ±1.4	42 ±2.1	**3.2 ±0.8**	15 ±2.4	17 ±2.5	130 ±3.7	143 ±4.3
Forces	1.0 ±0.15	59 ±0.8	58 ±0.8	25 ±1.0	**1.2 ±0.3**	1.8 ±0.3	7.1 ±0.7	8.1 ±1.1	4.5 ±0.5	6.6 ±0.6	7.4 ±0.4	27 ±1.3	30 ±1.5
QUESTS	1.0	5.1	5.4	3.4	1.1	**0.9**	1.1	1.3	1.0	2.0	2.1	3.4	4.0
QUESTS div	0.0	13	13	16	1.0	1.3	**0.7**	1.0	1.2	4.4	1.6	10	11
foreign n [count/total]	0/30	0/30	0/30	0/30	0/30	0/30	0/30	0/30	6/30	0/17	0/12	0/30	0/30
Overall (×MD)	1.0	8.9	10	7.2	**1.0**	1.3	1.6	1.9	1.1	3.9	3.7	6.6	7.8
Mean rank	–	9.5	10.3	8.1	2.5	3.0	4.3	5.5	3.8	5.9	6.0	7.9	9.2

Table 12: **Full metric profile, silica.** Conventions as in Table 10.

Metric	MD (n=10)	RMC (n=10)	HRMC (n=10)	HRMC-L (n=10)	CRN (n=10)	GLASS-c (n=10)	GLASS-u (n=10)	DM2-c (n=10)	DM2-u (n=10)	CDVAE (n=10)	MatterGen-c (n=10)	MatterGen-u (n=10)	AMDEN-c (n=10)	AMDEN-u (n=10)
g(r)	1.0 ±0.11	50 ±0.7	131 ±37	84 ±4.7	41 ±1.6	**0.9 ±0.2**	3.3 ±0.5	9.9 ±0.9	8.5 ±1.3	4.4 ±0.8	6.9 ±0.7	3.2 ±1.5	31 ±2.4	79 ±13
ADF	1.0 ±0.24	35 ±0.4	30 ±1.2	13 ±0.3	7.9 ±0.1	**1.4 ±0.1**	1.8 ±0.2	7.3 ±0.6	8.0 ±0.5	4.6 ±0.6	4.4 ±0.6	1.8 ±0.3	17 ±1.1	18 ±1.1
CN dist	1.0 ±0.19	11 ±0.2	11 ±0.8	6.7 ±0.2	7.4 ±0.0	**0.9 ±0.2**	1.5 ±0.1	2.7 ±0.4	3.5 ±0.1	2.3 ±0.3	4.5 ±0.3	1.1 ±0.2	8.4 ±4.5	26 ±4.1
CN outliers	0.0 ±0.37	56 ±3.7	64 ±16	8.5 ±1.2	**0.0 ±0.0**	0.5 ±0.4	0.0 ±0.4	1.5 ±0.6	1.5 ±0.5	0.0 ±0.1	4.0 ±1.3	0.0 ±0.1	26 ±70	300 ±68
mean CN	1.0 ±0.50	4.9 ±0.4	3.1 ±12	8.3 ±1.0	9.9 ±0.0	2.0 ±0.3	4.0 ±0.6	13 ±1.1	14 ±1.2	**1.6 ±1.3**	8.5 ±1.8	3.4 ±1.1	16 ±44	187 ±39
rings	1.0 ±0.35	9.0 ±0.7	13 ±4.2	5.0 ±0.9	5.5 ±1.0	2.3 ±0.5	3.0 ±0.7	14 ±1.4	18 ±1.4	**0.9 ±1.2**	20 ±1.7	4.2 ±0.3	15 ±12	52 ±7.5
ring dist	1.0 ±0.08	5.6 ±0.1	5.5 ±0.1	2.5 ±0.1	2.7 ±0.0	**1.1 ±0.1**	1.3 ±0.1	1.3 ±0.1	1.4 ±0.1	1.1 ±0.2	1.9 ±0.2	1.1 ±0.2	2.9 ±0.1	2.9 ±0.1
ring outliers	0.0 ±0.12	0.7 ±2.1	4.7 ±2.1	1.7 ±0.4	**0.0 ±0.0**	0.0 ±0.2	0.0 ±0.3	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	3.5 ±1.4	0.0 ±0.1	7.5 ±3.9	0.0 ±0.4
Voronoi	1.0 ±0.03	1.1 ±0.0	1.0 ±0.0	**0.7 ±0.0***	1.5 ±0.0	0.9 ±0.0*	1.0 ±0.0	1.0 ±0.0	0.9 ±0.0*	1.0 ±0.1	0.8 ±0.1*	0.9 ±0.0*	0.8 ±0.0*	0.7 ±0.1*
XRD	1.0 ±0.39	3.6 ±0.3	8.4 ±0.8	8.2 ±0.5	7.1 ±0.4	1.7 ±0.6	2.3 ±0.4	1.7 ±0.5	**1.7 ±0.5**	4.0 ±1.4	2.9 ±0.5	1.8 ±0.4	5.7 ±1.3	3.7 ±0.9
ND	1.0 ±0.44	7.3 ±0.4	3.3 ±0.6	4.6 ±0.9	7.9 ±0.5	**1.7 ±0.6**	2.3 ±0.5	1.8 ±0.5	2.3 ±0.8	3.3 ±1.6	7.1 ±1.4	2.7 ±0.7	9.9 ±2.0	8.5 ±1.6
bulk K	1.0 ±0.21	11‡	fail	9.1 ±0.3‡	1.6 ±0.1	2.5 ±0.4‡	**1.4 ±0.4**	2.5 ±0.6‡	1.5 ±0.6	1.5 ±0.6	9.5 ±0.9‡	6.7 ±1.9‡	9.4 ±0.4‡	10 ±0.7
shear G	1.0 ±0.53	29‡	fail	22 ±1.8‡	**2.7 ±0.4**	4.5 ±0.5‡	4.2 ±1.2	4.7 ±0.7‡	4.8 ±1.1	3.7 ±0.9	22 ±2.6‡	10 ±2.1‡	23 ±1.8‡	22 ±2.4
Young E	1.0 ±0.64	31‡	fail	23 ±1.5‡	**3.0 ±0.4**	4.8 ±0.5‡	4.4 ±1.2	4.7 ±0.6‡	4.7 ±1.2	3.8 ±1.1	24 ±2.7‡	12 ±2.8‡	25 ±1.7‡	24 ±2.3
relax dist	1.0 ±0.71	39‡	fail	24 ±1.1‡	1.3 ±0.6	**1.0 ±0.2‡**	1.9 ±0.5	4.0 ±0.5‡	5.5 ±1.2	2.0 ±0.7	20 ±4.6‡	9.5 ±3.8‡	23 ±1.9‡	24 ±2.4
relaxed n [converged/total]	5/10	1/10	0/10	3/10	10/10	4/10	8/10	5/10	7/10	7/10	4/10	4/10	5/10	7/10
yield σ	1.0 ±0.17	2.2 ±0.1	fail	1.7 ±0.3	3.3 ±0.7	0.9 ±0.2	1.3 ±0.3	1.0 ±0.2	**0.6 ±0.2***	1.1 ±2.2	1.6 ±0.2	1.1 ±0.2	1.7 ±0.4	2.0 ±0.1
UTS	1.0 ±0.36	19 ±0.5	fail	14 ±0.4	4.2 ±0.7	1.7 ±0.5	**1.1 ±0.8**	3.3 ±1.2	2.5 ±0.5	2.2 ±1.3	7.0 ±1.0	2.6 ±0.3	10 ±1.4	12 ±1.1
Energy/atom	1.0 ±0.45	1021 ±12	920 ±29	654 ±12	16 ±0.4	4.4 ±1.0	**1.1 ±0.5**	42 ±3.1	48 ±3.6	6.4 ±0.7	103 ±9.8	4.7 ±1.7	579 ±72	980 ±81
Forces	1.0 ±0.24	143 ±1.6	137 ±3.4	99 ±3.1	16 ±0.0	6.0 ±1.1	4.7 ±0.7	23 ±1.3	26 ±1.7	**3.8 ±0.7**	24 ±2.0	3.8 ±0.5	110 ±3.9	132 ±3.2
QUESTS	1.0	4.6	4.7	4.5	**0.0**	0.9	0.8	1.3	1.4	1.0	2.8	1.3	4.0	4.4
QUESTS div	0.0	14	14	14	35	1.5	1.0	**0.2**	1.2	0.7	11	3.5	14	14
foreign n [count/total]	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	2/10	0/10	0/10	0/10	0/10
Overall (×MD)	1.0	11	11 (15/21)	8.5	4.2	1.5	**1.5**	2.7	2.5	2.0	7.0	2.7	14	18
Mean rank	–	11.0	11.6	9.1	6.5	3.3	3.4	5.3	5.5	4.5	7.8	4.2	9.1	9.8

Table 13: **Full metric profile, CuZr.** Conventions as in Table 10. HRMC appears only as the 0.80-wall arm (HRMC-w80).

Metric	MD (n=10)	RMC (n=10)	HRMC-w80 (n=10)	GLASS-c (n=10)	GLASS-u (n=10)	DM2-c (n=10)	DM2-u (n=10)	CDVAE (n=10)	MatterGen-c (n=10)	MatterGen-u (n=10)	AMDEN-c (n=10)	AMDEN-u (n=10)
g(r)	1.0 ±0.16	20 ±0.3	620 ±411	1.8 ±0.2	2.1 ±0.3	3.4 ±0.4	4.4 ±0.4	**1.7 ±0.2**	6.1 ±6.3	8.6 ±5.1	12 ±1.1	27 ±2.1
ADF	1.0 ±0.16	6.6 ±0.1	22 ±3.6	1.3 ±0.2	**1.2 ±0.2**	1.5 ±0.2	1.7 ±0.2	1.4 ±0.2	2.5 ±0.3	1.6 ±0.3	6.0 ±0.4	7.9 ±0.4
CN dist	1.0 ±0.08	3.3 ±0.1	6.9 ±3.2	1.1 ±0.1	1.1 ±0.1	**1.0 ±0.1**	1.1 ±0.1	1.1 ±0.1	1.6 ±0.2	1.2 ±0.1	1.9 ±0.2	1.9 ±0.2
CN outliers	1.0 ±0.49	**0.1 ±0.1***	169 ±79	0.7 ±0.2	0.8 ±0.2	0.7 ±0.3	0.6 ±0.3	0.9 ±0.2	1.0 ±0.4	0.8 ±0.3	0.7 ±0.3	1.2 ±0.3
mean CN	1.0 ±0.41	2.0 ±0.5	87 ±61	2.3 ±1.0	2.6 ±0.7	**1.9 ±0.7**	1.9 ±0.8	2.5 ±0.7	4.0 ±1.1	2.2 ±0.9	4.1 ±1.2	6.5 ±1.7
rings	1.0 ±0.24	2.5 ±0.4	32 ±12	1.2 ±0.5	1.6 ±0.3	1.2 ±0.4	**1.2 ±0.4**	1.4 ±0.5	2.2 ±0.5	1.3 ±0.5	2.4 ±1.0	4.3 ±1.4
ring dist	1.0 ±0.31	4.2 ±0.3	34 ±14	1.5 ±0.4	**1.1 ±0.3**	1.4 ±0.2	1.7 ±0.3	1.6 ±0.4	3.2 ±0.4	1.6 ±0.5	5.4 ±0.4	7.0 ±0.2
ring outliers	0.0 ±0.03	1.8 ±0.8	1047 ±627	0.0 ±0.1	**0.0 ±0.1**	0.0 ±0.0	0.0 ±0.1	0.0 ±0.0	0.3 ±0.2	0.0 ±0.1	1.4 ±0.6	3.1 ±1.0
Voronoi	1.0 ±0.05	2.0 ±0.1	6.3 ±0.6	**1.0 ±0.0**	1.0 ±0.1	1.0 ±0.1	1.0 ±0.0	1.0 ±0.1	1.5 ±0.1	1.0 ±0.1	2.6 ±0.1	2.8 ±0.1
XRD	1.0 ±0.15	10 ±0.4	50 ±20	1.3 ±0.2	**1.1 ±0.3**	1.3 ±0.2	1.2 ±0.2	2.1 ±0.4‡	5.3 ±1.8	5.0 ±1.5	6.2 ±0.8	7.7 ±0.8
ND	1.0 ±0.20	4.2 ±0.4	35 ±13	1.3 ±0.2	**0.9 ±0.3**	1.3 ±0.2	1.2 ±0.1	1.8 ±0.4‡	5.5 ±1.7	3.0 ±1.0	3.1 ±0.3	3.9 ±0.5
bulk K	1.0 ±0.53	3.7 ±0.5	2.7 ±0.5	1.3 ±0.5	2.0 ±0.8	1.5 ±0.4	**1.3 ±0.5**	1.8 ±0.5‡	5.5 ±2.1	4.2 ±1.7	1.7 ±0.5	2.1 ±0.8
shear G	1.0 ±0.30	2.1 ±0.5	1.6 ±0.5	1.3 ±0.3	**0.9 ±0.4**	2.0 ±0.5	1.8 ±0.6	2.7 ±0.8‡	1.6 ±0.5	2.0 ±0.4	1.5 ±0.7	1.5 ±0.7
Young E	1.0 ±0.28	2.1 ±0.5	1.6 ±0.5	1.3 ±0.3	**0.9 ±0.4**	2.0 ±0.5	1.8 ±0.6	2.7 ±0.8‡	1.6 ±0.5	2.0 ±0.5	1.5 ±0.7	1.5 ±0.7
relax dist	1.0 ±0.24	46 ±1.4	57 ±3.3	5.3 ±0.9	**3.7 ±0.9**	4.4 ±0.9	6.2 ±0.7	7.0 ±1.3‡	17 ±2.2	7.7 ±1.4	32 ±1.5	33 ±1.0
relaxed n [converged/total]	30/30	30/30	30/30	30/30	30/30	30/30	30/30	20/30	30/30	30/30	30/30	29/30
yield σ	1.0 ±0.23	1.5 ±0.6	1.6 ±0.4	**1.0 ±0.4**	1.7 ±0.4	1.6 ±0.5	1.5 ±0.4	1.5 ±0.5‡	1.7 ±0.4	1.4 ±0.4	1.6 ±0.5	1.6 ±0.7
UTS	1.0 ±0.28	2.6 ±0.4	2.6 ±0.5	**1.0 ±0.4**	1.5 ±0.4	1.5 ±0.5	1.5 ±0.3	2.4 ±0.6‡	2.5 ±0.4	2.3 ±0.7	3.3 ±0.4	3.2 ±0.5
Energy/atom	1.0 ±0.27	267 ±6.2	1117 ±447	8.8 ±0.8	9.3 ±0.8	2.9 ±0.6	**2.8 ±0.5**	16 ±6.7‡	53 ±25	49 ±22	83 ±7.7	83 ±5.5
Forces	1.0 ±0.11	79 ±0.4	99 ±18	**1.5 ±0.2**	1.6 ±0.3	20 ±0.4	20 ±0.4	7.4 ±0.6‡	17 ±0.9	22 ±0.3	1.7 ±0.4	2.4 ±0.4
QUESTS	1.0	1.7	2.4	1.0	1.0	0.8	**0.7**	0.8	1.1	0.9	1.4	1.6
QUESTS div	0.0	5.9	14	1.2	1.3	3.2	4.1	4.2	2.1	**0.9**	6.0	7.8
foreign n [count/total]	0/30	0/30	0/30	0/30	0/30	0/30	0/30	10/30	0/30	0/30	0/30	0/30
Overall (×MD)	1.0	3.1	26	1.2	**1.2**	1.3	1.3	1.7	2.2	1.7	2.3	3.3
Mean rank	–	8.3	9.6	3.1	3.5	3.8	4.1	5.7	6.6	5.6	7.5	8.3

Table 14: **Full metric profile, ZrNiAl.** Conventions as in Table 10.

Metric	MD (n=10)	RMC (n=10)	HRMC (n=10)	HRMC-L (n=10)	GLASS-c (n=10)	GLASS-u (n=10)	DM2-c (n=10)	DM2-u (n=10)	CDVAE (n=10)	MatterGen-c (n=1)	MatterGen-u (n=9)	AMDEN-c (n=6)	AMDEN-u (n=8)
g(r)	1.0 ±0.12	34 ±0.5	71 ±2.0	52 ±1.1	2.6 ±0.3	2.9 ±0.5	5.0 ±0.5	6.0 ±0.6	**2.3 ±0.5**	26 ±6.7	101 ±14	12 ±1.0	177 ±114
ADF	1.0 ±0.12	5.5 ±0.2	10 ±0.2	8.5 ±0.1	1.3 ±0.2	**1.3 ±0.2**	1.9 ±0.2	2.1 ±0.2	1.4 ±0.2	2.7 ±0.2	1.7 ±0.3	4.6 ±0.3	7.1 ±0.4
CN dist	1.0 ±0.10	2.6 ±0.1	2.2 ±0.1	2.1 ±0.1	1.3 ±0.2	1.4 ±0.1	1.7 ±0.2	1.7 ±0.3	**1.2 ±0.2**	1.6 ±0.1	1.5 ±0.2	1.8 ±0.2	1.9 ±0.2
CN outliers	1.0 ±0.33	**0.2 ±0.1***	2.3 ±0.3	1.5 ±0.4	0.9 ±0.2	0.9 ±0.3	1.4 ±0.3	1.4 ±0.3	1.5 ±0.4	1.1 ±0.6	1.1 ±0.3	1.7 ±0.4	2.2 ±0.5
mean CN	1.0 ±0.24	**0.4 ±0.2***	1.4 ±0.3	1.3 ±0.1	1.1 ±0.5	1.2 ±0.5	2.4 ±0.4	2.5 ±0.7	1.5 ±0.4	0.9 ±0.3	1.0 ±0.4	1.4 ±0.6	1.2 ±0.5
rings	1.0 ±0.32	1.3 ±0.3	5.5 ±0.4	3.9 ±0.3	1.1 ±0.3	1.1 ±0.4	2.1 ±0.3	2.1 ±0.6	1.2 ±0.3	**0.6 ±0.3**	1.3 ±0.4	1.9 ±0.6	2.2 ±0.8
ring dist	1.0 ±0.16	1.9 ±0.1	5.5 ±0.2	4.0 ±0.1	**0.9 ±0.2**	1.0 ±0.2	2.3 ±0.3	2.5 ±0.5	1.3 ±0.3	1.4 ±0.2	1.4 ±0.3	2.6 ±0.2	3.5 ±0.2
ring outliers	0.0 ±0.10	2.2 ±0.7	12 ±2.5	3.7 ±1.3	**0.0 ±0.1**	0.0 ±0.1	0.1 ±0.1	0.1 ±0.1	0.0 ±0.1	0.3 ±0.2	0.0 ±0.1	1.1 ±0.5	1.6 ±0.6
Voronoi	1.0 ±0.05	1.6 ±0.1	3.3 ±0.1	2.6 ±0.1	1.1 ±0.1	**1.0 ±0.1**	1.1 ±0.1	1.1 ±0.1	1.0 ±0.1	1.4 ±0.1	1.1 ±0.1	1.8 ±0.1	2.2 ±0.1
XRD	1.0 ±0.26	5.8 ±0.3	5.5 ±0.2	3.6 ±0.3	2.5 ±0.4	3.1 ±0.3	1.3 ±0.4	1.2 ±0.4	1.5 ±0.4	2.6 ±0.3	1.7 ±0.2	**0.9 ±0.2**	4.2 ±1.1
ND	1.0 ±0.27	3.6 ±0.2	4.0 ±0.1	3.1 ±0.3	2.0 ±0.3	2.7 ±0.3	1.2 ±0.3	1.0 ±0.3	1.3 ±0.4	2.9 ±0.3	2.2 ±0.5	**0.9 ±0.2**	4.2 ±1.2
bulk K	1.0 ±0.57	7.4 ±0.5	7.4 ±0.4	3.3 ±0.5	5.0 ±0.6	6.3 ±0.5	**1.3 ±0.4**	1.6 ±0.5	1.9 ±0.6‡	5.0 ±0.9	3.0 ±1.0	3.0 ±1.2	3.0 ±0.8
shear G	1.0 ±0.52	4.2 ±0.7	3.4 ±0.8	3.1 ±0.5	2.9 ±0.6	**2.5 ±0.7**	2.8 ±0.6	3.0 ±0.9	3.3 ±1.1‡	4.0 ±0.9	4.7 ±0.9	2.8 ±1.9	4.4 ±0.9
Young E	1.0 ±0.53	4.1 ±0.7	3.3 ±0.8	3.0 ±0.5	2.8 ±0.6	**2.4 ±0.7**	2.7 ±0.6	3.0 ±0.9	3.3 ±1.1‡	3.9 ±0.9	4.6 ±0.9	2.8 ±2.0	4.3 ±0.9
relax dist	1.0 ±0.27	58 ±2.1	59 ±2.0	46 ±1.8	10 ±1.8	**8.5 ±1.6**	13 ±1.9	11 ±2.1	9.7 ±1.7‡	33 ±1.6	15 ±3.1	39 ±1.6	40 ±2.5
relaxed n [converged/total]	40/40	39/40	40/40	40/40	40/40	40/40	40/40	40/40	29/40	27/27	39/39	36/36	38/38
yield σ	1.0 ±0.41	2.0 ±0.6	2.0 ±0.6	2.2 ±0.5	1.4 ±0.5	2.3 ±0.6	1.3 ±0.4	1.8 ±0.5	2.3 ±0.7‡	3.1 ±0.5	2.1 ±0.8	1.5 ±0.6	**1.3 ±0.6**
UTS	1.0 ±0.24	2.9 ±0.4	2.7 ±0.4	2.7 ±0.4	1.8 ±0.3	2.3 ±0.4	**1.2 ±0.3**	1.7 ±0.5	1.9 ±0.5‡	2.2 ±0.5	2.1 ±0.6	1.7 ±0.3	2.0 ±0.4
Energy/atom	1.0 ±0.29	279 ±4.5	230 ±3.5	155 ±2.1	28 ±0.9	33 ±0.7	**6.5 ±0.5**	7.1 ±0.4	15 ±5.6‡	40 ±12	58 ±11	25 ±8.9	132 ±37
Forces	1.0 ±0.09	74 ±0.7	46 ±0.9	51 ±0.9	**1.2 ±0.1**	1.4 ±0.2	16 ±0.2	16 ±0.2	4.3 ±0.5‡	11 ±0.9	16 ±0.3	3.2 ±0.4	3.0 ±0.3
QUESTS	1.0	1.8	2.1	1.9	1.0	1.1	**0.9**	0.9	0.9	1.3	1.2	1.4	2.1
QUESTS div	0.0	7.1	9.3	8.5	1.8	1.6	3.0	3.8	**0.8**	3.8	9.5	4.3	8.3
foreign n [count/total]	0/40	0/40	0/40	0/40	0/40	0/40	0/40	0/40	11/40	0/27	0/39	0/36	0/38
Overall (×MD)	1.0	3.5	4.7	3.4	**1.4**	1.6	1.8	1.9	1.5	2.2	1.9	2.1	3.0
Mean rank	–	8.8	10.0	8.9	4.0	4.8	4.6	5.4	5.0	6.6	5.8	6.2	8.0

E.4 VALIDITY TABLE

Table 15 counts, for every method, the scored structures that fail each validity gate; each structure is counted under the first gate it fails. Only three models fail any gate. CDVAE's 38 failures are 36 out-of-composition elements and 2 overlapping pairs; DM^2's 14 and AMDEN's 72 are all overlapping atoms, DM^2's all at a-C 1.9 g/cm^3 and AMDEN's all on a-C and silica. No structure fails the zero-ADF gate, because every structure that would trips the pair-distance gate first. Table 16 gives the same accounting per condition over every draw, including draws beyond the ten scored, for the methods whose pass rate varies.

Table 15: **Gate failures by method,** pooled over the 14 conditions, counted within the ten scored structures per condition. Each structure is counted under the first gate it fails: an element outside the training composition, a pair closer than 1.0 Å, an identically zero ADF. HRMC (all arms) pools the random-start, liquid-start, and 0.80-wall arms.

Method	Conditions	Scored	Foreign element	Pair < 1.0 Å	Zero ADF	Passing	Rate
MD	14	140	0	0	0	140	100%
RMC	14	140	0	0	0	140	100%
HRMC (all arms)	14	220	0	0	0	220	100%
CRN	4	40	0	0	0	40	100%
GLASS-c	14	140	0	0	0	140	100%
GLASS-u	14	140	0	0	0	140	100%
DM2-c	14	140	0	6	0	134	96%
DM2-u	14	140	0	8	0	132	94%
CDVAE	14	140	36	2	0	102	73%
MatterGen-c	13	104	0	0	0	104	100%
MatterGen-u	11	101	0	0	0	101	100%
AMDEN-c	14	136	0	33	0	103	76%
AMDEN-u	14	138	0	39	0	99	72%

Table 16: **Validity per condition,** valid over drawn structures, for every method whose pass rate varies. RMC, HRMC, CRN, and all variations of GLASS and MatterGen pass every gate on every structure and are omitted.

System	Condition	CDVAE	DM2-c	DM2-u	AMDEN-c	AMDEN-u
a-C	C_1.9	15/20 75%	16/40 40%	10/120 8%	0/10 0%	0/10 0%
	C_2.4	13/20 65%	10/10 100%	10/10 100%	2/10 20%	0/10 0%
	C_2.9	15/20 75%	10/10 100%	10/10 100%	5/10 50%	1/10 10%
a-Si	Si_1.8	14/20 70%	10/10 100%	10/10 100%	10/10 100%	10/10 100%
	Si_2.3	17/20 85%	10/10 100%	10/10 100%	10/10 100%	10/10 100%
	Si_2.8	17/20 85%	10/10 100%	10/10 100%	10/10 100%	10/10 100%
SiO2	SiO2_2.2	14/20 70%	10/10 100%	10/10 100%	0/10 0%	0/10 0%
CuZr	Cu30Zr70_7.0	17/20 85%	10/10 100%	10/10 100%	10/10 100%	10/10 100%
	Cu50Zr50_7.4	14/20 70%	10/10 100%	10/10 100%	10/10 100%	10/10 100%
	Cu70Zr30_7.9	10/20 50%	10/10 100%	10/10 100%	10/10 100%	10/10 100%
ZrNiAl	Zr1Ni1Al1_5.8	15/20 75%	10/10 100%	10/10 100%	10/10 100%	10/10 100%
	Zr1Ni1Al3_4.6	17/20 85%	10/10 100%	10/10 100%	6/6 100%	8/8 100%
	Zr1Ni3Al1_6.7	9/20 45%	10/10 100%	10/10 100%	10/10 100%	10/10 100%
	Zr3Ni1Al1_6.1	18/20 90%	10/10 100%	10/10 100%	10/10 100%	10/10 100%
	all conditions	205/280 73%	146/170 86%	140/250 56%	103/136 76%	99/138 72%

E.5 Attempted-Draw Density Coverage

E.5.1 Discussion of Unguided vs. Guided Density Coverage

Tables 18–22 and Figure 12 summarize accepted structures only. Figure 15 instead shows every attempted unguided draw, accepted or not, and makes the mechanism behind Section 2.3's coverage numbers directly visible. MatterGen's per-system distributions (top row) are each a single narrow mode that simply does not reach several target buckets — carbon's mode sits at ~ 1.7 g/cm^3 and never approaches 2.4 or 2.9; amorphous silicon's mode sits at ~ 3.0 and rarely reaches 1.8 or 2.3 — because unconditional MatterGen has no lever to move this mode at all. AMDEN's distributions (bottom row) are visibly multi-modal even without PDF guidance, one mode per target density, because the native density-conditioning property is active in both the "unguided" and PDF-guided arms (Section 2.3); the modes are cleanly separated for carbon, amorphous silicon and CuZr, and increasingly overlapping toward ZrNiAl's harder edge buckets (4.6, 6.7 g/cm^3), consistent with the hit-rate collapse reported for those conditions elsewhere in this appendix.

Figure 16 shows the same breakdown for every attempted PDF-guided draw. AMDEN's already-separated modes (Figure 15) sharpen further under guidance, each one narrowing around its target rather than changing shape. MatterGen's response is qualitatively different from what per-target guidance might suggest: rather than splitting into one narrow mode per target density, its guided draws largely remain one continuous, wider spread spanning much of the unguided range — carbon's guided draws span roughly 1.75–2.9 g/cm^3 with no separation visible between the 1.9, 2.4 and 2.9 targets, and amorphous silicon's guided draws stay concentrated near its unguided mode (~ 2.8–3.0 g/cm^3) rather than reaching down to 1.8 or 2.3. This is consistent with guidance nudging each draw toward its target without fully overriding MatterGen's own unconditional generative range, and with that paper's observation that guidance does not fully close MatterGen's coverage gap even where it helps.

E.5.2 Density-Control Mechanism

Figure 17 isolates which of AMDEN's two mechanisms actually drives density control: the ghost-atom padding itself, or the explicit density-conditioning property layered on top of it (Appendix A.8). With ghost atoms but no density property, generated density is nearly insensitive to the requested target and instead clusters around one value regardless of which target was asked for (2.12, 2.17, and 2.08 g/cm^3 at nominal 1.8, 2.3, and 2.8, respectively), with a spread of 0.15–0.20 g/cm^3 per target. Adding the density property closes this gap: means move to 1.86, 2.39, and 2.88 g/cm^3, within 0.09 g/cm^3 of each target, and the per-target spread falls by roughly an order of magnitude at the two higher targets (0.03 and 0.02 g/cm^3). The ghost-atom mechanism alone is therefore not what steers density toward a target. Instead, it supplies the capacity for atom count to vary, but only the density property converts that capacity into targeted control.

E.6 Incremental Effect of PDF Guidance

Tables 18–22 isolate what PDF guidance adds. Each pairs the unguided and guided arm of the four generative models that have both; the cells are the full-table values, and the mean rank is recomputed among these eight studies. Guidance improves DM^2 and AMDEN on all five systems and GLASS on four, the exception being amorphous carbon (1.93 against 1.81 unguided). It worsens MatterGen on four of five, most on silica (6.00 against 3.48); amorphous silicon is the exception (4.46 against 4.74), on as few as two structures per arm at 1.8 g/cm^3.

Before the per-metric breakdown, Table 17 isolates the *relaxed n* column alone, per model and condition: the denominator is the number of accepted draws contributing to that arm's cell (so it doubles as a coverage count), and the numerator is how many of those further survived relaxation for the mechanical-property rows. Reading down a model's pair of columns shows where guidance changes not just structural quality but how many structures reach a scoreable state at all. MatterGen's coverage gaps are visible directly: no accepted draws at amorphous carbon 2.4 g/cm^3 unguided (filled only once guided) and zero accepted in either arm at 2.9 g/cm^3. Guidance lowers the relaxed fraction at silica for every model except MatterGen, and at amorphous carbon 1.9 g/cm^3 for MatterGen and GLASS but not AMDEN or DM^2; almost every other cell is saturated at full or near-full survival

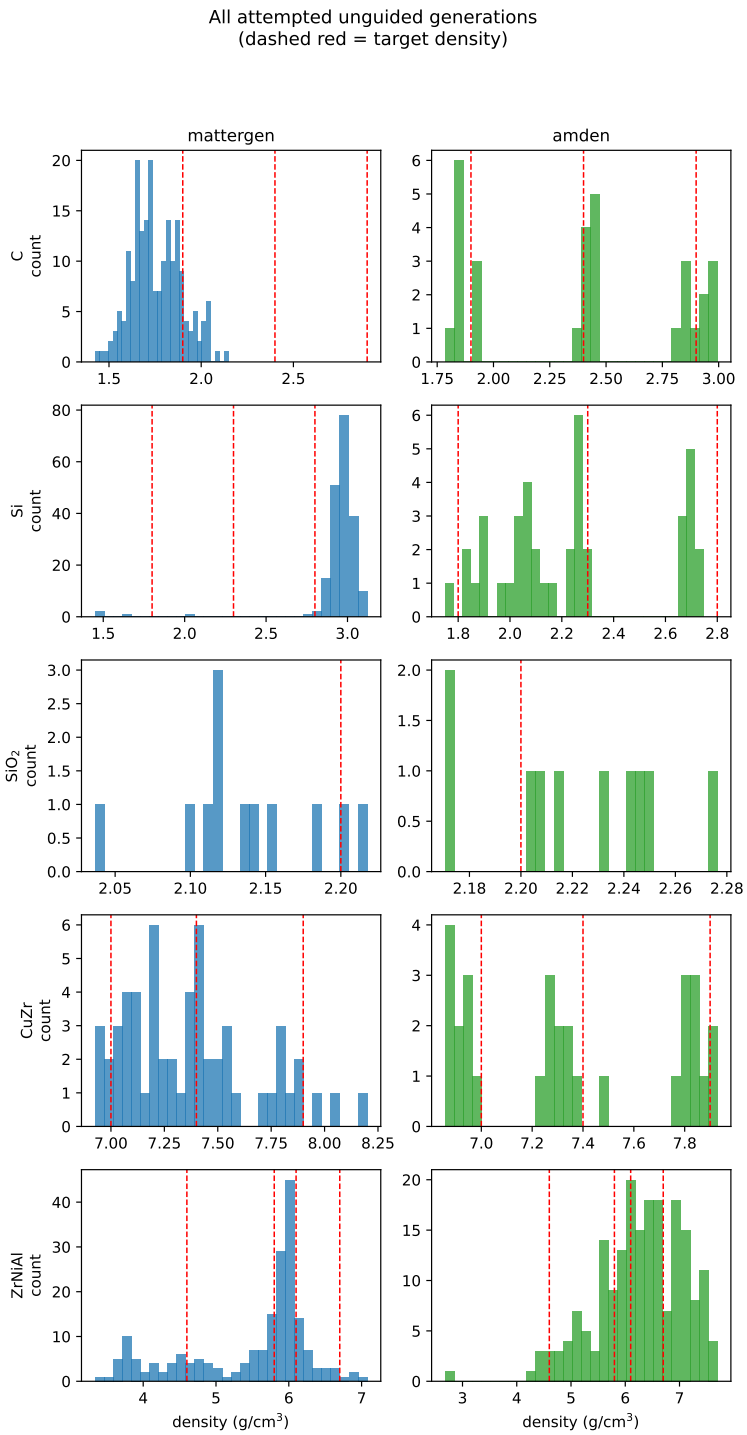


Figure 15: **Density of every attempted unguided generation**, MatterGen (left column) and AMDEN (right column), one row per system. Dashed red lines mark the target density.

in both arms, so this axis mostly discriminates the same handful of hard conditions the per-metric tables already flag.

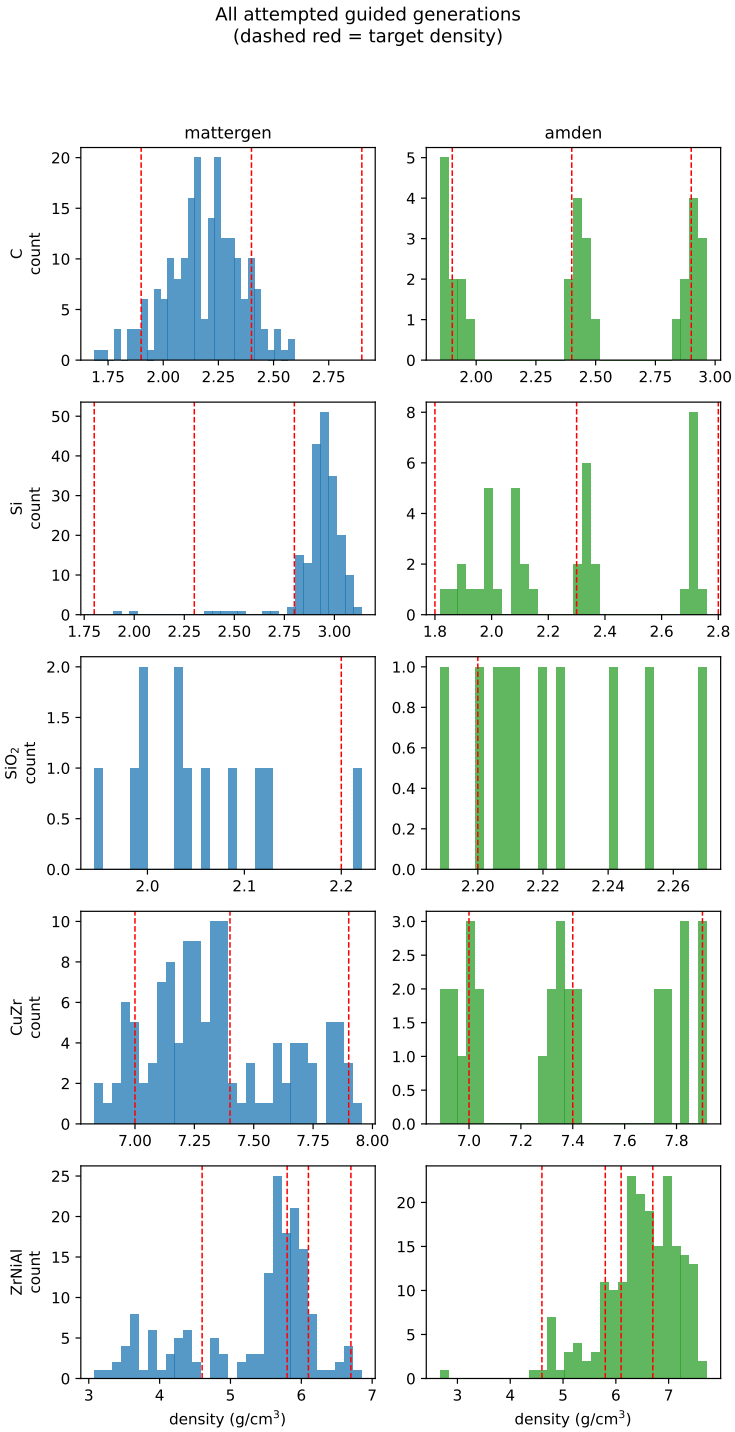


Figure 16: **Density of every attempted PDF-guided generation**, MatterGen (left column) and AMDEN (right column), one row per system. Dashed red lines mark the target density.

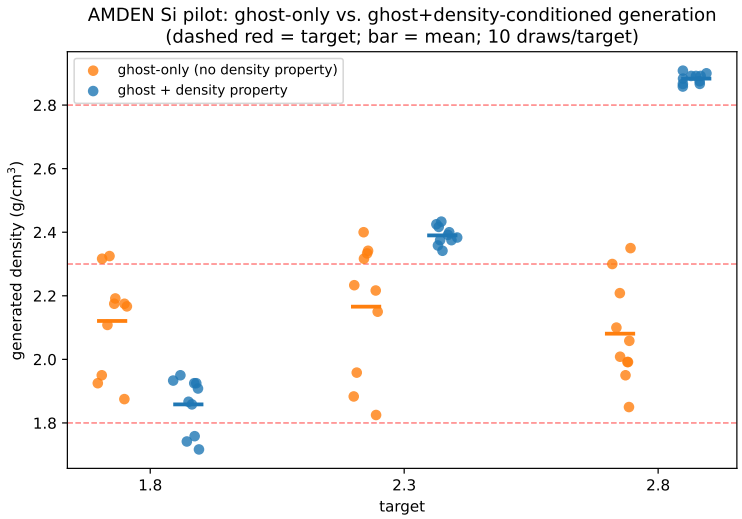


Figure 17: **AMDEN density-control mechanism, amorphous silicon pilot.** Generated density per draw (10 draws/target) at three nominal targets, ghost-only configuration (orange) versus ghost+density-property (blue). Dashed red lines mark the targets; horizontal bars mark the per-target, per-configuration mean.

Table 17: **Accepted-draw and relaxed-structure coverage, unguided (U) versus PDF-guided (C).** Cells are $n_{\text{relaxed}}/n_{\text{accepted}}$: the denominator is the accepted-draw count feeding that arm's condition (a coverage count in its own right), the numerator how many of those further survived relaxation for the mechanical-property rows of Tables 18–22. "–" marks an arm with no accepted draws recorded (not the same as $0/0$, generated but none relaxed). Shading marks the guided arm's relaxed fraction relative to unguided: green improves, red worsens; unshaded cells are unchanged or not comparable.

System	ρ	AMDEN U	AMDEN C	MatterGen U	MatterGen C	GLASS U	GLASS C	DM^2 U	DM^2 C
a-C	1.9	0/10	0/10	10/10	9/10	10/10	9/10	1/10	5/10
	2.4	0/10	0/10	–	9/10	8/10	8/10	9/10	7/10
	2.9	3/10	2/10	–	–	10/10	7/10	5/10	9/10
a-Si	1.8	10/10	10/10	2/2	2/2	10/10	10/10	10/10	10/10
	2.3	10/10	10/10	–	5/5	10/10	10/10	10/10	10/10
	2.8	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
SiO_2	2.2	7/10	5/10	4/10	4/10	8/10	4/10	7/10	5/10
CuZr	7.0	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
	7.4	9/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
	7.9	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
ZrNiAl	4.6	8/8	6/6	10/10	6/6	10/10	10/10	10/10	10/10
	5.8	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
	6.1	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
	6.7	10/10	10/10	9/9	1/1	10/10	10/10	10/10	10/10

Table 18: **Incremental effect of PDF guidance, amorphous carbon.** Unguided (-u) and PDF-guided (-c) arms of GLASS, DM^2, MatterGen, and AMDEN. Cells equal those of Table 10; the mean rank is recomputed over these eight columns. Markers as in Table 10; k/N runs marks an arm generated at only some of the system's conditions.

Metric	GLASS-u (n=10)	GLASS-c (n=10)	DM2-u (n=10)	DM2-c (n=10)	MatterGen-u (n=10 · 1/3 runs)	MatterGen-c (n=10 · 2/3 runs)	AMDEN-u (n=10)	AMDEN-c (n=10)
g(r)	2.9 ±0.6	**1.2 ±0.2**	18 ±4.1	5.1 ±1.0	3.9 ±0.8	4.3 ±0.9	3855 ±40421	38 ±2.4
ADF	**2.7 ±0.3**	3.7 ±0.3	30 ±3.3	23 ±2.0	3.2 ±0.4	3.5 ±0.5	128 ±10.0	77 ±1.2
CN dist	**1.3 ±0.3**	1.5 ±0.2	11 ±4.0	4.6 ±2.4	3.5 ±0.6	3.9 ±0.2	28 ±1.8	14 ±0.4
CN outliers	1.0 ±0.3	1.3 ±0.3	38 ±31	2.5 ±18	**0.0 ±0.0**	3.8 ±0.5	163 ±20	37 ±2.9
mean CN	1.6 ±0.6	**1.6 ±0.3**	57 ±39	8.8 ±23	5.0 ±1.6	4.6 ±0.6	165 ±24	18 ±3.2
rings	**0.9 ±0.3**	1.1 ±0.3	5.0 ±2.9	2.3 ±1.7	2.4 ±0.8	2.3 ±0.3	14 ±2.3	1.7 ±0.7
ring dist	1.4 ±0.1	**1.2 ±0.2**	3.6 ±0.3	2.7 ±0.2	1.7 ±0.2	1.3 ±0.1	7.6 ±0.2	5.9 ±0.1
ring outliers	**0.9 ±0.2**	0.9 ±0.2	13 ±5.9	14 ±3.1	2.9 ±1.4	1.7 ±0.4	64 ±19	69 ±2.2
Voronoi	0.9 ±0.0*	0.9 ±0.0*	0.9 ±0.0*	0.9 ±0.0*	1.0 ±0.0*	**0.9 ±0.0***	1.5 ±0.0	1.5 ±0.0
XRD	1.2 ±0.4	**0.8 ±0.3**	1.9 ±0.4	1.1 ±0.5	2.2 ±0.6	2.8 ±0.8	11 ±1.1	5.4 ±0.7
ND	1.4 ±0.4	**0.9 ±0.3**	2.6 ±0.6	1.1 ±0.5	1.7 ±0.6	2.9 ±0.9	18 ±2.0	5.8 ±1.0
bulk K	2.3 ±0.6	2.9 ±0.7	6.7 ±0.8‡	5.4 ±0.7‡	**2.3 ±0.6**	8.2 ±1.4	24 ±4.5‡	16‡
shear G	**4.8 ±1.0**	5.0 ±0.8	9.7 ±0.5‡	8.6 ±1.2‡	5.1 ±1.1	12 ±1.7	22 ±2.5‡	17‡
Young E	**4.7 ±1.0**	4.9 ±0.9	11 ±0.5‡	9.3 ±1.3‡	5.8 ±1.4	13 ±1.9	23 ±3.0‡	18‡
relax dist	2.9 ±0.5	**2.6 ±0.5**	9.4 ±0.7‡	6.7 ±1.1‡	2.7 ±0.7	4.1 ±1.4	46 ±3.2‡	43‡
relaxed n [converged/total]	28/30	24/30	15/30	21/30	10/10	18/20	3/30	2/30
yield σ	**0.7 ±0.4**	1.0 ±0.3	2.8 ±0.7	2.2 ±0.7	1.3 ±0.4	2.0 ±0.5	9.0 ±0.6‡	7.1 ±0.8‡
UTS	0.6 ±0.2*	**0.5 ±0.2***	1.4 ±0.3	0.9 ±0.3	1.5 ±0.3	2.0 ±0.5	2.6 ±0.2‡	2.6 ±0.2‡
Energy/atom	**8.2 ±0.7**	9.1 ±0.6	106 ±19	62 ±10	11 ±1.3	14 ±0.8	954 ±147	223 ±16
Forces	4.5 ±0.5	4.4 ±0.5	44 ±1.7	36 ±1.4	3.5 ±0.2	**3.0 ±0.4**	87 ±1.2	73 ±1.6
QUESTS	**1.3**	1.3	1.8	1.7	1.5	1.8	4.9	4.8
QUESTS div	**3.0**	3.3	7.5	4.7	6.0	7.8	18	18
foreign n [count/total]	0/30	0/30	0/30	0/30	0/10	0/20	0/30	0/30
Overall (×MD)	1.6	**1.4**	5.8	3.7	2.7	3.4	26 (51/63)	16 (51/63)
Mean rank	1.8	1.9	4.3	3.6	3.4	4.0	6.7	6.1

Table 19: **Incremental effect of PDF guidance, amorphous silicon.** Conventions as in Table 18; cells equal those of Table 11.

Metric	GLASS-u (n=10)	GLASS-c (n=10)	DM2-u (n=10)	DM2-c (n=10)	MatterGen-u (n=2 · 2/3 runs)	MatterGen-c (n=2)	AMDEN-u (n=10)	AMDEN-c (n=10)
g(r)	3.5 ±0.9	**0.5 ±0.1***	9.1 ±1.1	6.5 ±0.7	9.7 ±2.2	6.1 ±1.7	23 ±1.4	13 ±1.0
ADF	1.3 ±0.2	**0.8 ±0.2**	5.3 ±0.4	4.4 ±0.3	4.1 ±0.3	4.7 ±0.6	13 ±0.5	14 ±0.5
CN dist	1.6 ±0.2	**0.9 ±0.1***	2.4 ±0.3	1.9 ±0.3	3.2 ±0.3	3.2 ±0.4	4.5 ±0.2	4.4 ±0.3
CN outliers	**0.2 ±0.2**	0.4 ±0.4	1.9 ±0.5	1.1 ±0.5	3.5 ±2.3	4.5 ±2.1	14 ±1.4	11 ±1.4
mean CN	1.4 ±0.3	**0.4 ±0.2***	3.2 ±0.5	1.7 ±0.4	4.9 ±0.7	5.0 ±1.0	2.1 ±0.6	3.1 ±0.9
rings	1.4 ±0.5	**0.7 ±0.3**	2.5 ±0.5	1.7 ±0.5	8.2 ±1.5	6.4 ±1.7	4.5 ±0.8	1.3 ±0.7
ring dist	1.1 ±0.1	**0.9 ±0.1***	2.3 ±0.1	1.9 ±0.1	1.9 ±0.2	1.8 ±0.1	5.7 ±0.2	5.1 ±0.2
ring outliers	0.5 ±0.5	1.0 ±0.4	0.5 ±0.3	0.2 ±0.4	**0.1 ±0.2**	0.4 ±0.8	23 ±3.3	16 ±2.3
Voronoi	**1.0 ±0.0**	1.0 ±0.0	1.0 ±0.0	1.0 ±0.0	1.0 ±0.0	1.0 ±0.0	1.4 ±0.0	1.4 ±0.0
XRD	1.7 ±0.3	**1.3 ±0.2**	1.6 ±0.4	1.5 ±0.3	4.3 ±0.8	5.8 ±1.7	4.8 ±0.7	3.0 ±0.5
ND	1.6 ±0.3	**1.2 ±0.2**	1.5 ±0.4	1.5 ±0.2	5.2 ±1.0	5.7 ±1.8	4.8 ±0.7	3.0 ±0.4
bulk K	**1.8 ±0.9**	2.0 ±0.6	4.4 ±0.9	5.3 ±0.9	7.5 ±0.8	5.4 ±2.6	21 ±4.0	12 ±1.3
shear G	**1.6 ±0.5**	2.2 ±0.6	3.7 ±0.5	3.8 ±0.6	3.5 ±0.2	1.9 ±1.5	9.0 ±1.1	6.3 ±0.7
Young E	**1.9 ±0.5**	2.2 ±0.6	3.9 ±0.5	4.1 ±0.7	3.8 ±0.2	2.4 ±1.3	9.9 ±1.3	6.8 ±0.7
relax dist	**2.2 ±0.5**	2.4 ±0.4	11 ±1.0	10 ±0.7	3.1 ±2.0	4.7 ±1.2	23 ±1.3	23 ±1.1
relaxed n [converged/total]	30/30	30/30	30/30	30/30	12/12	17/17	30/30	30/30
yield σ	**0.8 ±0.4**	0.9 ±0.3	1.5 ±0.4	1.3 ±0.4	2.1 ±0.3	0.9 ±0.3	2.0 ±0.2	2.0 ±0.3
UTS	1.8 ±0.3	**1.2 ±0.4**	2.0 ±0.5	1.9 ±0.4	2.5 ±0.4	2.6 ±0.8	4.3 ±0.6	4.2 ±0.3
Energy/atom	9.5 ±1.0	**8.3 ±0.9**	42 ±2.1	36 ±1.4	17 ±2.5	15 ±2.4	143 ±4.3	130 ±3.7
Forces	1.8 ±0.3	**1.2 ±0.3**	8.1 ±1.1	7.1 ±0.7	7.4 ±0.4	6.6 ±0.6	30 ±1.5	27 ±1.3
QUESTS	**0.9**	1.1	1.3	1.1	2.1	2.0	4.0	3.4
QUESTS div	1.3	1.0	1.0	**0.7**	1.6	4.4	11	10
foreign n [count/total]	0/30	0/30	0/30	0/30	0/12	0/17	0/30	0/30
Overall (×MD)	1.3	**1.0**	1.9	1.6	3.7	3.9	7.8	6.6
Mean rank	2.6	2.0	4.4	3.5	4.7	4.5	7.0	6.2

Table 20: **Incremental effect of PDF guidance, silica.** Conventions as in Table 18; cells equal those of Table 12.

Metric	GLASS-u (n=10)	GLASS-c (n=10)	DM2-u (n=10)	DM2-c (n=10)	MatterGen-u (n=10)	MatterGen-c (n=10)	AMDEN-u (n=10)	AMDEN-c (n=10)
g(r)	3.3 ±0.5	**0.9 ±0.2**	8.5 ±1.3	9.9 ±0.9	3.2 ±1.5	6.9 ±0.7	79 ±13	31 ±2.4
ADF	1.8 ±0.2	**1.4 ±0.1**	8.0 ±0.5	7.3 ±0.6	1.8 ±0.3	4.4 ±0.6	18 ±1.1	17 ±1.1
CN dist	1.5 ±0.1	**0.9 ±0.2**	3.5 ±0.1	2.7 ±0.4	1.1 ±0.2	4.5 ±0.3	26 ±4.1	8.4 ±4.5
CN outliers	**0.0 ±0.4**	0.5 ±0.4	1.5 ±0.5	1.5 ±0.6	0.0 ±0.1	4.0 ±1.3	300 ±68	26 ±70
mean CN	4.0 ±0.6	**2.0 ±0.3**	14 ±1.2	13 ±1.1	3.4 ±1.1	8.5 ±1.8	187 ±39	16 ±44
rings	3.0 ±0.7	**2.3 ±0.5**	18 ±1.4	14 ±1.4	4.2 ±0.3	20 ±1.7	52 ±7.5	15 ±12
ring dist	1.3 ±0.1	**1.1 ±0.1**	1.4 ±0.1	1.3 ±0.1	1.1 ±0.2	1.9 ±0.2	2.9 ±0.1	2.9 ±0.1
ring outliers	0.0 ±0.3	**0.0 ±0.2**	0.0 ±0.0	0.0 ±0.0	0.0 ±0.1	3.5 ±1.4	0.0 ±0.4	7.5 ±3.9
Voronoi	1.0 ±0.0	0.9 ±0.0*	0.9 ±0.0*	1.0 ±0.0	0.9 ±0.0*	0.8 ±0.1*	**0.7 ±0.1***	0.8 ±0.0*
XRD	2.3 ±0.4	1.7 ±0.6	**1.7 ±0.5**	1.7 ±0.5	1.8 ±0.4	2.9 ±0.5	3.7 ±0.9	5.7 ±1.3
ND	2.3 ±0.5	**1.7 ±0.6**	2.3 ±0.8	1.8 ±0.5	2.7 ±0.7	7.1 ±1.4	8.5 ±1.6	9.9 ±2.0
bulk K	**1.4 ±0.4**	2.5 ±0.4‡	1.5 ±0.6	2.5 ±0.6‡	6.7 ±1.9‡	9.5 ±0.9‡	10 ±0.7	9.4 ±0.4‡
shear G	**4.2 ±1.2**	4.5 ±0.5‡	4.8 ±1.1	4.7 ±0.7‡	10 ±2.1‡	22 ±2.6‡	22 ±2.4	23 ±1.8‡
Young E	**4.4 ±1.2**	4.8 ±0.5‡	4.7 ±1.2	4.7 ±0.6‡	12 ±2.8‡	24 ±2.7‡	24 ±2.3	25 ±1.7‡
relax dist	1.9 ±0.5	**1.0 ±0.2‡**	5.5 ±1.2	4.0 ±0.5‡	9.5 ±3.8‡	20 ±4.6‡	24 ±2.4	23 ±1.9‡
relaxed n [converged/total]	8/10	4/10	7/10	5/10	4/10	4/10	7/10	5/10
yield σ	1.3 ±0.3	0.9 ±0.2	**0.6 ±0.2***	1.0 ±0.2	1.1 ±0.2	1.6 ±0.2	2.0 ±0.1	1.7 ±0.4
UTS	**1.1 ±0.8**	1.7 ±0.5	2.5 ±0.5	3.3 ±1.2	2.6 ±0.3	7.0 ±1.0	12 ±1.1	10 ±1.4
Energy/atom	**1.1 ±0.5**	4.4 ±1.0	48 ±3.6	42 ±3.1	4.7 ±1.7	103 ±9.8	980 ±81	579 ±72
Forces	4.7 ±0.7	6.0 ±1.1	26 ±1.7	23 ±1.3	**3.8 ±0.5**	24 ±2.0	132 ±3.2	110 ±3.9
QUESTS	**0.8**	0.9	1.4	1.3	1.3	2.8	4.4	4.0
QUESTS div	1.0	1.5	1.2	**0.2**	3.5	11	14	14
foreign n [count/total]	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10
Overall (×MD)	**1.5**	1.5	2.5	2.7	2.7	7.0	18	14
Mean rank	2.8	2.4	3.9	3.8	3.5	6.0	7.0	6.7

Table 21: **Incremental effect of PDF guidance, CuZr.** Conventions as in Table 18; cells equal those of Table 13.

Metric	GLASS-u (n=10)	GLASS-c (n=10)	DM2-u (n=10)	DM2-c (n=10)	MatterGen-u (n=10)	MatterGen-c (n=10)	AMDEN-u (n=10)	AMDEN-c (n=10)
g(r)	2.1 ±0.3	**1.8 ±0.2**	4.4 ±0.4	3.4 ±0.4	8.6 ±5.1	6.1 ±6.3	27 ±2.1	12 ±1.1
ADF	**1.2 ±0.2**	1.3 ±0.2	1.7 ±0.2	1.5 ±0.2	1.6 ±0.3	2.5 ±0.3	7.9 ±0.4	6.0 ±0.4
CN dist	1.1 ±0.1	1.1 ±0.1	1.1 ±0.1	**1.0 ±0.1**	1.2 ±0.1	1.6 ±0.2	1.9 ±0.2	1.9 ±0.2
CN outliers	0.8 ±0.2	0.7 ±0.2	**0.6 ±0.3**	0.7 ±0.3	0.8 ±0.3	1.0 ±0.4	1.2 ±0.3	0.7 ±0.3
mean CN	2.6 ±0.7	2.3 ±1.0	1.9 ±0.8	**1.9 ±0.7**	2.2 ±0.9	4.0 ±1.1	6.5 ±1.7	4.1 ±1.2
rings	1.6 ±0.3	1.2 ±0.5	**1.2 ±0.4**	1.2 ±0.4	1.3 ±0.5	2.2 ±0.5	4.3 ±1.4	2.4 ±1.0
ring dist	**1.1 ±0.3**	1.5 ±0.4	1.7 ±0.3	1.4 ±0.2	1.6 ±0.5	3.2 ±0.4	7.0 ±0.2	5.4 ±0.4
ring outliers	**0.0 ±0.1**	0.0 ±0.1	0.0 ±0.1	0.0 ±0.0	0.0 ±0.1	0.3 ±0.2	3.1 ±1.0	1.4 ±0.6
Voronoi	1.0 ±0.1	**1.0 ±0.0**	1.0 ±0.0	1.0 ±0.1	1.0 ±0.1	1.5 ±0.1	2.8 ±0.1	2.6 ±0.1
XRD	**1.1 ±0.3**	1.3 ±0.2	1.2 ±0.2	1.3 ±0.2	5.0 ±1.5	5.3 ±1.8	7.7 ±0.8	6.2 ±0.8
ND	**0.9 ±0.3**	1.3 ±0.2	1.2 ±0.1	1.3 ±0.2	3.0 ±1.0	5.5 ±1.7	3.9 ±0.5	3.1 ±0.3
bulk K	2.0 ±0.8	1.3 ±0.5	**1.3 ±0.5**	1.5 ±0.4	4.2 ±1.7	5.5 ±2.1	2.1 ±0.8	1.7 ±0.5
shear G	**0.9 ±0.4**	1.3 ±0.3	1.8 ±0.6	2.0 ±0.5	2.0 ±0.4	1.6 ±0.5	1.5 ±0.7	1.5 ±0.7
Young E	**0.9 ±0.4**	1.3 ±0.3	1.8 ±0.6	2.0 ±0.5	2.0 ±0.5	1.6 ±0.5	1.5 ±0.7	1.5 ±0.7
relax dist	**3.7 ±0.9**	5.3 ±0.9	6.2 ±0.7	4.4 ±0.9	7.7 ±1.4	17 ±2.2	33 ±1.0	32 ±1.5
relaxed n [converged/total]	30/30	30/30	30/30	30/30	30/30	30/30	29/30	30/30
yield σ	1.7 ±0.4	**1.0 ±0.4**	1.5 ±0.4	1.6 ±0.5	1.4 ±0.4	1.7 ±0.4	1.6 ±0.7	1.6 ±0.5
UTS	1.5 ±0.4	**1.0 ±0.4**	1.5 ±0.3	1.5 ±0.5	2.3 ±0.7	2.5 ±0.4	3.2 ±0.5	3.3 ±0.4
Energy/atom	9.3 ±0.8	8.8 ±0.8	**2.8 ±0.5**	2.9 ±0.6	49 ±22	53 ±25	83 ±5.5	83 ±7.7
Forces	1.6 ±0.3	**1.5 ±0.2**	20 ±0.4	20 ±0.4	22 ±0.3	17 ±0.9	2.4 ±0.4	1.7 ±0.4
QUESTS	1.0	1.0	**0.7**	0.8	0.9	1.1	1.6	1.4
QUESTS div	1.3	1.2	4.1	3.2	**0.9**	2.1	7.8	6.0
foreign n [count/total]	0/30	0/30	0/30	0/30	0/30	0/30	0/30	0/30
Overall (×MD)	**1.2**	1.2	1.3	1.3	1.7	2.2	3.3	2.3
Mean rank	2.9	2.7	3.5	3.2	4.7	5.6	7.0	6.2

Table 22: **Incremental effect of PDF guidance, ZrNiAl.** Conventions as in Table 18; cells equal those of Table 14.

Metric	GLASS-u (n=10)	GLASS-c (n=10)	DM2-u (n=10)	DM2-c (n=10)	MatterGen-u (n=9)	MatterGen-c (n=1)	AMDEN-u (n=8)	AMDEN-c (n=6)
g(r)	2.9 ±0.5	**2.6 ±0.3**	6.0 ±0.6	5.0 ±0.5	101 ±14	26 ±6.7	177 ±114	12 ±1.0
ADF	**1.3 ±0.2**	1.3 ±0.2	2.1 ±0.2	1.9 ±0.2	1.7 ±0.3	2.7 ±0.2	7.1 ±0.4	4.6 ±0.3
CN dist	1.4 ±0.1	**1.3 ±0.2**	1.7 ±0.3	1.7 ±0.2	1.5 ±0.2	1.6 ±0.1	1.9 ±0.2	1.8 ±0.2
CN outliers	0.9 ±0.3	**0.9 ±0.2**	1.4 ±0.3	1.4 ±0.3	1.1 ±0.3	1.1 ±0.6	2.2 ±0.5	1.7 ±0.4
mean CN	1.2 ±0.5	1.1 ±0.5	2.5 ±0.7	2.4 ±0.4	1.0 ±0.4	**0.9 ±0.3**	1.2 ±0.5	1.4 ±0.6
rings	1.1 ±0.4	1.1 ±0.3	2.1 ±0.6	2.1 ±0.3	1.3 ±0.4	**0.6 ±0.3**	2.2 ±0.8	1.9 ±0.6
ring dist	1.0 ±0.2	**0.9 ±0.2**	2.5 ±0.5	2.3 ±0.3	1.4 ±0.3	1.4 ±0.2	3.5 ±0.2	2.6 ±0.2
ring outliers	0.0 ±0.1	**0.0 ±0.1**	0.1 ±0.1	0.1 ±0.1	0.0 ±0.1	0.3 ±0.2	1.6 ±0.6	1.1 ±0.5
Voronoi	**1.0 ±0.1**	1.1 ±0.1	1.1 ±0.1	1.1 ±0.1	1.1 ±0.1	1.4 ±0.1	2.2 ±0.1	1.8 ±0.1
XRD	3.1 ±0.3	2.5 ±0.4	1.2 ±0.4	1.3 ±0.4	1.7 ±0.2	2.6 ±0.3	4.2 ±1.1	**0.9 ±0.2**
ND	2.7 ±0.3	2.0 ±0.3	1.0 ±0.3	1.2 ±0.3	2.2 ±0.5	2.9 ±0.3	4.2 ±1.2	**0.9 ±0.2**
bulk K	6.3 ±0.5	5.0 ±0.6	1.6 ±0.5	**1.3 ±0.4**	3.0 ±1.0	5.0 ±0.9	3.0 ±0.8	3.0 ±1.2
shear G	**2.5 ±0.7**	2.9 ±0.6	3.0 ±0.9	2.8 ±0.6	4.7 ±0.9	4.0 ±0.9	4.4 ±0.9	2.8 ±1.9
Young E	**2.4 ±0.7**	2.8 ±0.6	3.0 ±0.9	2.7 ±0.6	4.6 ±0.9	3.9 ±0.9	4.3 ±0.9	2.8 ±2.0
relax dist	**8.5 ±1.6**	10 ±1.8	11 ±2.1	13 ±1.9	15 ±3.1	33 ±1.6	40 ±2.5	39 ±1.6
relaxed n [converged/total]	40/40	40/40	40/40	40/40	39/39	27/27	38/38	36/36
yield σ	2.3 ±0.6	1.4 ±0.5	1.8 ±0.5	1.3 ±0.4	2.1 ±0.8	3.1 ±0.5	**1.3 ±0.6**	1.5 ±0.6
UTS	2.3 ±0.4	1.8 ±0.3	1.7 ±0.5	**1.2 ±0.3**	2.1 ±0.6	2.2 ±0.5	2.0 ±0.4	1.7 ±0.3
Energy/atom	33 ±0.7	28 ±0.9	7.1 ±0.4	**6.5 ±0.5**	58 ±11	40 ±12	132 ±37	25 ±8.9
Forces	1.4 ±0.2	**1.2 ±0.1**	16 ±0.2	16 ±0.2	16 ±0.3	11 ±0.9	3.0 ±0.3	3.2 ±0.4
QUESTS	1.1	1.0	0.9	**0.9**	1.2	1.3	2.1	1.4
QUESTS div	**1.6**	1.8	3.8	3.0	9.5	3.8	8.3	4.3
foreign n [count/total]	0/40	0/40	0/40	0/40	0/39	0/27	0/38	0/36
Overall (×MD)	1.6	**1.4**	1.9	1.8	1.9	2.2	3.0	2.1
Mean rank	3.8	3.2	4.3	3.8	4.4	5.2	6.2	5.1