CRYSTALJTF: JACOBIAN-PULLED GRADIENT GUIDANCE WITH ATOM-ADAPTIVE GRADIENT SCALING FOR CRYSTAL DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models offer a promising route to crystal discovery, but generating thermodynamically stable and novel structures remains challenging. Integrating physical signals from pretrained interatomic potentials into sampling trajectories could improve generation quality without additional training. Effective feedback, however, must address two challenges: the coordinate-space mismatch between predicted-endpoint forces and current sampling coordinates, and differences in local response scales across atoms in periodic systems. We propose **CrystalJTF**, a plug-and-play, training-free guidance framework for crystal diffusion models. Point Q-JTF pulls endpoint forces back to the current sampling coordinates through the predicted clean endpoint map's Jacobian transpose. Relative-BB uses sampling history to adapt atom-wise relative update scales, while a shared periodic geometry cap limits displacement magnitudes. CrystalJTF freezes both the generator and potential and retains native lattice and element-type sampling rules, supporting composition-conditioned and unconditional generation. Across six benchmark settings involving CrysLLM-Gen, DiffCSP, and MatterGen, CrystalJTF consistently improves the fractions of stable or metastable, unique, and novel structures (SUN/MSUN). Under CHGNet 0.3.0 evaluation, SUN and MSUN increase by an average of 1.28 and 4.78 percentage points on MP-20, respectively, while MatterGen's MSUN increases by 13.8 percentage points on Alex-MP-20. Sensitivity analyses across evaluation-potential versions and matching tolerances further support the robustness of these results. An anonymized reference implementation is available at https://anonymous.4open.science/r/CrystalJTF-605C/.

1 INTRODUCTION

The structure and composition of crystalline materials determine their potential in energy storage, catalysis, and electronic devices. Exploring this vast design space requires candidates that are both structurally novel and physically stable. Diffusion models offer a promising route by learning from known crystal distributions (Xie et al., 2022; Jiao et al., 2023; Zeni et al., 2025), yet common discovery workflows separate generation from physical validation. Relaxation and filtering operate as post-processing: they can improve or discard candidates, but cannot feed physical signals back into the sampling trajectories that produced them. Incorporating physical feedback during sampling could instead steer exploration toward better candidates while retaining subsequent validation. Because such steering can alter the balance between stability and exploration, energy reduction alone is insufficient to assess its value. Joint novelty–stability metrics are needed to evaluate whether guidance yields physically viable structures distinct from known materials.

Pretrained generators provide structural priors, while interatomic potentials provide energy and force evaluations (Batatia et al., 2022; Deng et al., 2023). Existing approaches connect generation and physical feedback by training additional components or adapting the generator. For example, DAO trains a time-dependent energy predictor for composition-conditioned crystal structure prediction (Wu et al., 2026), while Chemeleon2 uses reinforcement-learning post-training with multi-objective rewards (Park & Walsh, 2026). We pursue a complementary, practically useful direction: *can an off-the-shelf potential guide a pretrained crystal generator during sampling, without updat-*

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.

ing either model's parameters or training an auxiliary predictor? A training-free interface would reuse learned structural priors and retain native lattice and element-type sampling, while making physical feedback a plug-and-play component across generators, avoiding the cost of additional training tailored to each framework.

This interface faces two coupled challenges. First, noisy intermediate structures can fall outside the potential's training distribution, making direct force evaluation unreliable. Evaluating a predicted clean endpoint mitigates this problem, but its forces must be transported through the endpoint map to the current sampling coordinates. Second, a uniform step size ignores differences in atomic response scales, while unconstrained atom-wise updates can violate periodic geometry. Effective **guidance** must therefore combine *force transport between coordinate spaces* with *relative atom-wise scaling under a shared periodic geometric constraint*.

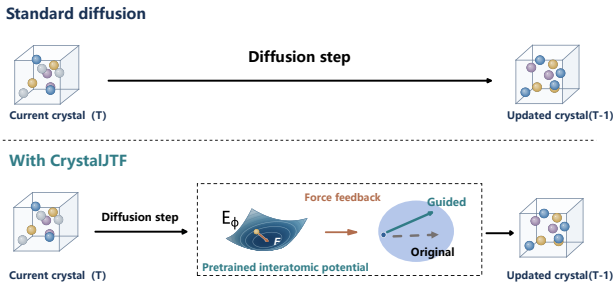


Figure 1: Standard and CrystalJTF-guided sampling. The generator and potential remain frozen. At each guided transition, the base sampler first produces a candidate at $t-1$. CrystalJTF then pulls back forces from a predicted clean endpoint to correct the candidate coordinates before sampling continues. Lattice and element-type updates remain governed by the base sampler's native rules.

To address these challenges, we propose **CrystalJTF**, a training-free, plug-and-play framework that guides crystal diffusion models with an interatomic potential (Figure 1). CrystalJTF modifies only atomic coordinates at each sampling step and keeps the atom count fixed. It leaves composition and lattice updates to the base generator, so it works with both composition-conditioned generation and unconditional generation in which the generator samples element identities. Specifically, CrystalJTF has two components. First, **Predicted-Endpoint Jacobian-Transpose Force Guidance (Point Q-JTF)** uses the endpoint map's transpose Jacobian to pull endpoint forces back to the current coordinates. Then, **Atom-Wise Relative Barzilai–Borwein Scaling (Relative-BB)** uses sampling history to allocate relative update scales across atoms. A *periodic geometry cap* limits the shared step.

We evaluate CrystalJTF on CrysLLMGen (Khastagir et al., 2025), DiffCSP (Jiao et al., 2023), and MatterGen (Zeni et al., 2025). On MP-20 with CHGNet 0.3.0 evaluation, SUN and MSUN improve by an average of 1.28 and 4.78 percentage points across the three generators. These joint quality gains arise through model-dependent changes in stability and novelty, supporting the use of physical guidance across heterogeneous generators without an explicit novelty reward.

Contributions. Our contributions are threefold. First, we formulate a unified coordinate-endpoint interface for training-free, plug-and-play physical guidance that connects frozen pretrained generators to off-the-shelf interatomic potentials during sampling, supporting both conditional and unconditional generation. Second, Point Q-JTF and Relative-BB jointly determine the correction direction and relative atom-wise scales under a shared periodic geometry cap; we analyze the correction's local first-order descent property under stated assumptions. Third, paired evaluations of three generators, a two-factor ablation, and checks across CHGNet versions and matching tolerances assess transferability, module contributions, and dependence on evaluation settings.

2 Related Work

Generative models for periodic crystals. Crystal generators incorporate periodicity and symmetry in different ways. CDVAE and DiffCSP use diffusion models to generate periodic structures and jointly model fractional coordinates and lattices (Xie et al., 2022; Jiao et al., 2023). FlowMM formulates joint generation over atom types, coordinates, and lattices as Riemannian flow matching (Miller et al., 2024), while MatterGen scales joint diffusion and supports property-conditioned design (Zeni et al., 2025). A complementary line represents crystals as text: CrystaLLM and FlowLLM use crystal language models (Antunes et al., 2024; Sriram et al., 2024), and CrysLLMGen combines LLM-based discrete prediction with equivariant diffusion refinement (Khastagir et al., 2025).

Physics- and force-guided structure generation. Physical objectives have been introduced through auxiliary energy predictors and policy optimization. DAO trains a time-dependent energy predictor for composition-conditioned crystal structure prediction (Wu et al., 2026). OMatG-IRL applies inference-time policy-gradient updates to the velocity field of the OMatG flow model, using terminal rewards to optimize composition- conditioned crystal structure prediction. Its de novo experiments use a hybrid setting that reinforces continuous coordinate and lattice dynamics while keeping discrete species generation fixed (Höllmer & Martiniani, 2026). Chemeleon2 uses reinforcement-learning post-training with multi-objective rewards for materials generation (Park & Walsh, 2026). For molecular conformation generation, Elign uses a pretrained machine-learning force field to provide force and energy rewards during policy post-training, so no force-field calls are needed at inference (Li et al., 2026).

Training-free guidance for diffusion models. Training-free methods such as DPS, Universal Guidance, and FreeDoM control pretrained diffusion models through external objectives, including gradients propagated from denoised estimates to the current state (Chung et al., 2023; Bansal et al., 2023; Yu et al., 2023). MPGD further considers compatibility with the learned data manifold (He et al., 2024). For crystals, de Lambilly et al. (2026) guide MatterGen without finetuning, including endpoint-energy guidance with MatterSim. We study atom-adaptive scaling and a periodic geometry cap for joint novelty–stability gains across three generators.

Interatomic potentials and adaptive optimization. Equivariant machine-learning interatomic potentials provide differentiable energies and forces. MACE constructs force fields with higher-order equivariant messages (Batatia et al., 2022), while CHGNet learns a charge-informed crystal potential from Materials Project trajectories (Deng et al., 2023). These potentials are commonly used for relaxation and post-generation screening. For step selection, the classical Barzilai–Borwein method estimates an inverse-curvature scale from consecutive iterates and gradients (Barzilai & Borwein, 1988). CrystalJTF brings these ingredients together for periodic crystal diffusion: it pulls forces from a frozen potential evaluated at a predicted clean endpoint back to the current coordinates, then applies relative atom-wise scaling under periodic geometric constraints. The same coordinate-level interface can be used with pretrained generators that expose a differentiable endpoint map, without auxiliary predictor training or inference-time policy optimization.

3 Background and Problem Setup

Crystal states. A periodic crystal is represented by fractional coordinates $X \in [0,1)^{N\times 3}$, a lattice $L \in \mathbb{R}^{3\times 3}$, and atomic species $A = (a_1, \ldots, a_N)$. At diffusion time τ, Cartesian coordinates are $R_\tau = X_\tau L_\tau$, and A_τ denotes the generator's current element state, which may be soft or masked. Guidance modifies only coordinates; lattice and element updates follow the base sampler. The generator and potential remain frozen.

Endpoint energy. Noisy intermediate structures may be unsuitable for evaluating an interatomic potential. Following denoised-estimate guidance (Chung et al., 2023), we evaluate a pretrained potential E_ϕ on the generator's predicted clean coordinates Q_τ. The Cartesian endpoint map and induced energy are

$$\begin{aligned} \mathcal{Y}_\tau(R) &= Q_\tau(RL_\tau^{-1}; L_\tau, A_\tau)L_\tau, \\ \mathcal{E}_\tau(R) &= E_\phi(\mathcal{Y}_\tau(R), L_\tau, \widehat{A}_\tau), \qquad F_\tau = -\nabla_{\widehat{R}} E_\phi, \end{aligned} \tag{1}$$

where $\widehat{R} = \mathcal{Y}_\tau(R_\tau)$ and $\widehat{A}_\tau$ is the physical element assignment selected for the potential. Coordinate derivatives hold τ, L_τ, A_τ, and $\widehat{A}_\tau$ fixed, without differentiating through discrete label selection. Generator-specific maps and assignments appear in Appendix B.

Guidance objective. The endpoint force F_τ and the sampler update act in different coordinate spaces. Applying F_τ directly to R_τ ignores the Jacobian of $\mathcal{Y}_\tau$ and need not decrease $\mathcal{E}_\tau$. Moreover, a uniform step ignores differences in atomic curvature and local geometry. We seek a coordinate increment with local descent and zero net translation,

$$\langle \nabla_R \mathcal{E}_\tau(R_\tau), \Delta R_\tau \rangle < 0, \qquad \mathbf{1}^\top \Delta R_\tau = 0, \tag{2}$$

whenever a nonzero feasible descent direction exists. The update is mapped back through $X_\tau^+ = \mathrm{wrap}(X_\tau + \Delta R_\tau L_\tau^{-1})$ and checked for periodic geometric validity. This requires endpoint-force transport to sampling coordinates and atom-wise scaling under local displacement constraints.

4 METHOD

CrystalJTF separates coordinate correction into endpoint-force transport, atom-wise relative scaling, and global displacement control (Figure 2). Point Q-JTF connects endpoint physics to the current state, Relative-BB allocates relative update scales across atoms, and the periodic geometry cap sets their shared amplitude. This separates physical response from allowable structural perturbation.

The correction follows the native transition, including its scheduled noise, at the resulting state's response time τ. Lattice and element-type updates retain their native rules. Algorithm 1 summarizes the integration; Appendix A specifies numerical thresholds and fallback rules.

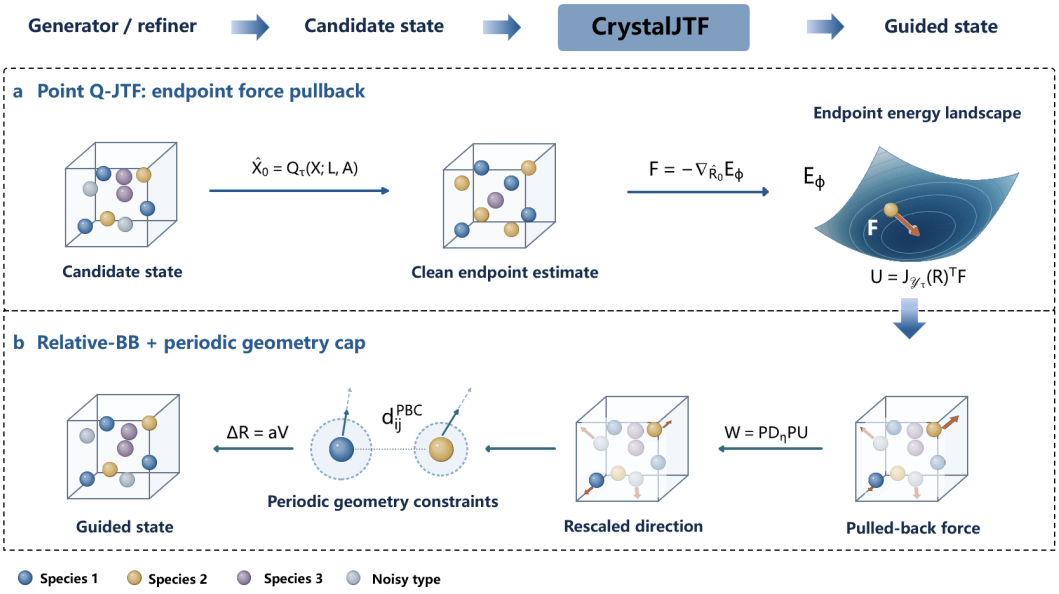


Figure 2: Overview of the CrystalJTF framework. (a) Point Q-JTF evaluates the interatomic force at the predicted clean endpoint and maps it back to the current coordinates through the endpoint-map Jacobian transpose. (b) Relative-BB assigns atom-wise update scales, while the periodic geometry cap limits the shared step. The base model retains control of lattice and species sampling.

4.1 ENDPOINT-FORCE TRANSPORT (POINT Q-JTF)

Endpoint forces act on predicted clean coordinates, whereas guidance updates the current state. Its perturbations affect endpoint energy through the generator, potentially moving several endpoint atoms. The pullback accounts for this response, which direct endpoint-force updates ignore.

Using $\mathcal{Y}_\tau$ and $\mathcal{E}_\tau$ from Section 3, we evaluate $F_\tau = -\nabla_{\widehat{R}} E_\phi$ at $\widehat{R} = \mathcal{Y}_\tau(R_\tau)$. The chain rule gives

$$U_\tau = J_{\mathcal{Y}_\tau}(R_\tau)^\top F_\tau = -\nabla_R \mathcal{E}_\tau(R_\tau). \tag{3}$$

Thus $\langle \nabla_R \mathcal{E}_\tau, U_\tau \rangle = -\|U_\tau\|_F^2 < 0$ for nonzero U_τ: the transported force is a local descent direction in the sampling coordinates. A single reverse-mode differentiation computes $J^\top F$ with the endpoint force held fixed, without constructing the full Jacobian or differentiating through the potential.

The transported direction need not align with the raw endpoint force. Its descent property concerns the induced endpoint energy at fixed τ, lattice, and endpoint species, not energies across diffusion times or the energy of the noisy intermediate crystal.

4.2 ATOM-WISE RELATIVE SCALING (RELATIVE-BB)

A shared scalar step cannot separately calibrate different atomic responses. We use historical displacements and gradient changes to estimate atom-wise scales without forming a Hessian. Because endpoint objectives vary across diffusion times, we use these estimates as relative calibration, not fixed-objective curvature; the geometry cap sets the final amplitude.

Let k index guidance opportunities, $P = I_N - \mathbf{1}\mathbf{1}^\top/N$ remove rigid translation, and $G_k = -PU_k$. Consecutive pre-guidance states define the periodic secant and gradient difference

$$S_k = P\ \mathrm{MIC}_{L_k}((X_k - X_{k-1})L_k), \qquad \Delta G_k = G_k - G_{k-1}. \tag{4}$$

The minimum-image convention (MIC) evaluates displacements in the current lattice. Adapting BB1 (Barzilai & Borwein, 1988), we retain relative inverse-curvature estimates rather than their time-dependent absolute scale:

$$b_{k,i} = \frac{\|S_{k,i}\|_2^2}{S_{k,i}^\top \Delta G_{k,i}}, \quad \eta_{k,i} = \begin{cases} \mathrm{clip}(b_{k,i}/m_k, \kappa^{-1}, \kappa), & i \in \mathcal{I}_k, \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

Here $\mathcal{I}_k$ contains atoms with nondegenerate secants and finite positive BB estimates, and $m_k = \mathrm{median}_{i\in\mathcal{I}_k}\, b_{k,i}$. Median normalization removes the shared multiplicative scale; clipping with $\kappa = 2$ bounds relative weights. Invalid estimates receive unit weights, as do all atoms when $\mathcal{I}_k$ is empty.

Atom-wise weighting can reintroduce translation after the original direction is centered. We therefore project both before and after applying $D_{\eta_k} = \mathrm{diag}(\eta_{k,1}, \ldots, \eta_{k,N})$: the first projection removes translation, and the second restores zero net translation after relative rescaling of the atoms.

$$W_k = PD_{\eta_k}PU_k, \qquad V_k = \frac{W_k}{\max(\|W_k\|_F, \epsilon_n)}. \tag{6}$$

For a locally differentiable endpoint energy at fixed time, lattice, and species, positive weights and $P = P^\top = P^2$ give

$$\langle \nabla_R \mathcal{E}_\tau, W_k \rangle_F = -\sum_{i=1}^{N} \eta_{k,i} \|(PU_k)_i\|_2^2 < 0 \quad \text{if } PU_k \neq 0. \tag{7}$$

Positive scaling within the projected subspace thus retains local descent without net translation. This is a directional statement, not a finite-step energy guarantee (Appendix A.5). History is reset when endpoint species change, so secants do not span different element assignments.

4.3 PERIODIC GEOMETRY CAP

Relative scaling leaves the global amplitude undetermined. An atomic bound limits motion relative to local geometry, while a neighbor-relative bound controls deformation from unequal neighboring displacements. A shared amplitude satisfies both bounds, followed by a periodic validity check.

Let $\mathcal{N}_6(i)$ contain the six nearest periodic neighbors of atom i and $\ell_i = \mathrm{median}_{j\in\mathcal{N}_6(i)}\, d_{ij}^{\mathrm{PBC}}$. We jointly limit atomic and neighbor-relative displacements:

$$a_k = \min\left\{ \min_i \frac{\rho\ell_i}{\|V_{k,i}\|_2}, \min_{\substack{i \\ j\in\mathcal{N}_6(i)}} \frac{\rho d_{ij}^{\mathrm{PBC}}}{\|V_{k,j} - V_{k,i}\|_2} \right\}, \tag{8}$$

where the pair minimum ranges over each atom i and its neighbors, and ratios with numerically zero denominators are omitted. We use $\rho = 0.15$. The fractional-coordinate proposal is

$$X_k^{\mathrm{trial}} = \mathrm{wrap}(X_k + a_k V_k L_k^{-1}). \tag{9}$$

It is accepted only if its minimum periodic distance is at least 0.5 Å; otherwise the native candidate is retained. This final check remains necessary because a finite neighbor set does not cover every possible close pair. The method uses no energy line search, and the geometry cap does not guarantee energy decrease after a finite step.

The first valid state initializes history without a correction. Guidance modifies only coordinates and retains the base sampler's lattice and element-type update rules, allowing the same interface to

support conditional and unconditional generators. Numerical thresholds and history-reset rules are specified in Appendix A.

Each active opportunity requires an endpoint-force evaluation and a reverse pass through the endpoint map. Only one previous coordinate–gradient pair is stored per sample, requiring $O(N)$ history. After neighbor-list construction, scaling and cap evaluation are linear in the retained atoms and neighbor edges; no dense coordinate Jacobian is stored.

Algorithm 1 CrystalJTF-guided sampling (one crystal)

Require: Frozen generator and potential, guidance interval $\mathcal{T}_g$, κ, ρ
Initialize state $\mathcal{C}$ and history $H \leftarrow \varnothing$
for each native denoising transition **do**
  $(X, L, A), \tau \leftarrow \mathrm{BaseStep}(\mathcal{C})$; $\mathcal{C} \leftarrow (X, L, A)$
  if $\tau \notin \mathcal{T}_g$ **then**
    continue
  end if
  Estimate $(\widehat{R}, \widehat{A})$; reset H on species change; evaluate F
  On invalid endpoint, force, or geometry: clear H and **continue**
  Compute U by Eq. (3); set $H_{\mathrm{next}} \leftarrow (X, -PU)$
  if $H \neq \varnothing$ and direction is valid **then**
    Compute V, a, and X^{trial} by Eqs. (6), (8), and (9)
    Accept valid X^{trial}; otherwise clear H_{next}
  end if
  $H \leftarrow H_{\mathrm{next}}$; $\mathcal{C} \leftarrow (X, L, A)$
end for
return $\mathcal{C}$

5 Experiments

5.1 Experimental setup

Generators and benchmarks. We evaluate CrystalJTF on three pretrained crystal generators with different architectures and sampling mechanisms: CrysLLMGen (Khastagir et al., 2025), DiffCSP (Jiao et al., 2023), and MatterGen (Zeni et al., 2025). The complete evaluation matrix contains six generator–benchmark settings: CrysLLMGen and DiffCSP on both MP-20 and MPTS-52 (Baird et al., 2024), MatterGen on MP-20, and MatterGen with its official Alex-MP-20 reference (Zeni et al., 2025). We sample 10,000 structures per arm for the CrysLLMGen and DiffCSP settings. Given MatterGen's large parameter count and longer sampling time in our setup, we generate 1,000 structures per arm for the MatterGen settings. All comparisons are made within a setting; the unequal sample counts are therefore not used for cross-generator count comparisons.

Paired inference protocol. For every setting, Baseline and CrystalJTF use the same frozen generator checkpoint, upstream conditioning, atom-count draw, and initial stochastic state whenever these quantities are exposed by the native sampler. The sample identifier is preserved through generation, relaxation, and scoring. CrystalJTF changes only the inference-time coordinate update: the lattice and element-type processes remain those of the original generator, and no generator parameter is retrained or fine-tuned. This paired design isolates the effect of the Jacobian-transported direction and atom-adaptive step calibration from changes in the underlying model or sampled inputs. Table 5 lists endpoint maps, checkpoints, guidance schedules, and evaluation references for each generator.

Relaxation and thermodynamic evaluation. Every generated structure is relaxed before thermodynamic scoring, including both atomic positions and the unit cell, using FIRE (Bitzek et al., 2006) with a force tolerance of $0.1\,\mathrm{eV}/\mathrm{\AA}$ and at most 500 optimization steps. To assess the robustness of our experimental conclusions to the evaluation-potential version, we repeat the complete relaxation and scoring pipeline with CHGNet (Deng et al., 2023) versions 0.2.0 and 0.3.0. The two versions serve only as evaluation potentials; evaluating the same generated samples with both versions does not constitute independent generation trials. Signed E_{hull} is computed against Materials Project (Jain et al., 2013) phase-diagram entries. Missing thermodynamic evaluations are reported as `thermo_unknown` and are not relabeled as unstable.

Metrics and reporting. Novelty (N) compares each relaxed structure with the corresponding training or reference set, while uniqueness (U) is recomputed within each generated arm. Both use Pymatgen (Ong et al., 2013) `StructureMatcher` with $(\mathrm{ltol}, \mathrm{stol}, \mathrm{angle_tol}) = (0.2, 0.3, 5^\circ)$, primitive-cell reduction, lattice scaling, and species-aware comparison. Strict stability (S) uses $E_{\mathrm{hull}} < 0\,\mathrm{eV/atom}$, and metastability (MS) uses $E_{\mathrm{hull}} < 0.1\,\mathrm{eV/atom}$. SUN and MSUN are the corresponding stable/metastable, unique, and novel intersections. Percentages use the fixed requested sample count as denominator rather than only successfully relaxed or thermodynamically evaluated structures. In Table 1, Δ denotes the paired CrystalJTF gain in percentage points.

5.2 Overall performance and cross-model generalization

Table 1: Generation performance across generators and benchmarks. Metrics are evaluated after relaxation (%, higher is better). Blue rows report Δ, the paired gain in percentage points.

CrysLLMGen		MP-20						MPTS-52					
CHGNet	Method	N	U	S	MS	**SUN**	**MSUN**	N	U	S	MS	**SUN**	**MSUN**
0.2.0	Baseline	82.76	77.14	11.60	52.93	5.53	30.59	93.28	93.20	7.22	38.00	5.71	30.34
	CrystalJTF	83.71	77.64	14.20	59.34	7.15	36.17	93.25	93.37	8.41	44.09	6.72	35.71
	Δ	+0.95	+0.50	+2.60	+6.41	+1.62	+5.58	-0.03	+0.17	+1.19	+6.09	+1.01	+5.37
0.3.0	Baseline	82.73	77.16	15.10	65.91	6.77	37.08	93.22	93.13	7.43	38.89	5.13	30.34
	CrystalJTF	83.53	77.54	17.08	70.31	8.28	41.82	93.17	93.27	8.72	44.93	6.46	35.96
	Δ	+0.80	+0.38	+1.98	+4.40	+1.51	+4.74	-0.05	+0.14	+1.29	+6.04	+1.33	+5.62
DiffCSP		MP-20						MPTS-52					
CHGNet	Method	N	U	S	MS	**SUN**	**MSUN**	N	U	S	MS	**SUN**	**MSUN**
0.2.0	Baseline	90.53	97.09	10.38	53.69	8.50	44.83	53.67	55.78	3.49	15.41	2.96	12.54
	CrystalJTF	91.41	97.15	11.89	59.53	10.02	51.36	54.47	55.78	4.04	16.68	3.59	14.19
	Δ	+0.88	+0.06	+1.51	+5.84	+1.52	+6.53	+0.80	+0.00	+0.55	+1.27	+0.63	+1.65
0.3.0	Baseline	90.52	97.10	11.71	58.61	8.86	48.22	54.58	56.82	3.74	16.12	2.99	12.93
	CrystalJTF	91.33	97.22	13.00	64.00	10.29	54.52	55.24	56.70	4.13	17.05	3.46	14.36
	Δ	+0.81	+0.12	+1.29	+5.39	+1.43	+6.30	+0.66	-0.12	+0.39	+0.93	+0.47	+1.43
MatterGen		MP-20						Alex-MP-20					
CHGNet	Method	N	U	S	MS	**SUN**	**MSUN**	N	U	S	MS	**SUN**	**MSUN**
0.2.0	Baseline	92.00	99.80	10.30	56.70	9.40	51.00	61.40	100.00	19.80	75.70	9.90	45.00
	CrystalJTF	95.40	100.00	9.90	56.70	9.50	53.60	82.30	99.90	18.00	72.90	13.10	59.40
	Δ	+3.40	+0.20	-0.40	+0.00	+0.10	+2.60	+20.90	-0.10	-1.80	-2.80	+3.20	+14.40
0.3.0	Baseline	92.20	99.90	11.10	60.70	9.10	53.50	61.40	100.00	21.40	87.00	10.30	50.60
	CrystalJTF	95.50	100.00	11.50	60.90	10.00	56.80	83.00	99.90	17.30	80.80	11.40	64.40
	Δ	+3.30	+0.10	+0.40	+0.20	+0.90	+3.30	+21.60	-0.10	-4.10	-6.20	+1.10	+13.80

Table 1 shows positive ΔSUN and ΔMSUN for all six generator–benchmark settings under both evaluation potentials. On the common MP-20 benchmark with CHGNet 0.3.0, CrystalJTF improves SUN by 1.51, 1.43, and 0.90 percentage points on CrysLLMGen, DiffCSP, and MatterGen, respectively, for an average gain of 1.28 points. The corresponding MSUN gains are 4.74, 6.30, and 3.30 points, averaging 4.78 points. CrystalJTF therefore transfers across substantially different generator architectures without retraining their parameters.

For CrysLLMGen and DiffCSP, the improvement is primarily stability driven: S and MS increase while shifts in N and U remain comparatively small. The same pattern generalizes beyond MP-20. On MPTS-52 under CHGNet 0.2.0/0.3.0, CrystalJTF raises SUN by 1.01/1.33 points and MSUN by 5.37/5.62 points for CrysLLMGen, and raises SUN by 0.63/0.47 points and MSUN by 1.65/1.43 points for DiffCSP. The larger MSUN gains indicate that guidance moves more samples into the broader near-stable region before all of them cross the stricter stability threshold. MatterGen exhibits a complementary, novelty-driven regime, which we analyze in Section 5.3. Importantly, the sign of every SUN and MSUN gain is unchanged across CHGNet versions, supporting robustness to the choice of evaluation potential.

Energy profiles. Figure 4 in Appendix C compares the post-relaxation energy distributions of Baseline and CrystalJTF. CrystalJTF lowers median E_{hull} from 0.0480 to 0.0383 eV/atom for Crys-LLMGen and from 0.0749 to 0.0655 eV/atom for DiffCSP, supporting the stability-driven improvements in Table 1.

5.3 MATCHER-ROBUST NOVELTY GAINS ON ALEX-MP-20

MatterGen on Alex-MP-20 exhibits the largest novelty increase in Table 1. To test whether this gain is caused by structures lying just across the default matching boundary, we rescore the same 1,000 paired, CHGNet 0.3.0-relaxed samples using strict, default, and loose `StructureMatcher` tolerances. Only N/U matching changes in this test; the thermodynamic labels and E_{hull} values are held fixed. Table 2 reports the joint and component metrics under all three matching rules.

Table 2: Sensitivity to structure-matching tolerances on MatterGen–Alex-MP-20. We evaluate 1,000 paired samples with CHGNet 0.3.0. Values are percentages; Δ is CrystalJTF minus Baseline. Strict, Default, and Loose (`ltol`, `stol`, `angle_tol`) are $(0.10, 0.20, 3^\circ)$, $(0.20, 0.30, 5^\circ)$, and $(0.30, 0.50, 10^\circ)$, respectively.

Matcher	Method	N	U	S	MS	**SUN**	**MSUN**
Strict	Baseline	63.6	100.0	21.4	87.0	11.1	52.5
	CrystalJTF	85.0	99.9	17.3	80.8	12.0	66.2
	Δ	+21.4	-0.1	-4.1	-6.2	+0.9	+13.7
Default	Baseline	61.4	100.0	21.4	87.0	10.3	50.6
	CrystalJTF	83.0	99.9	17.3	80.8	11.4	64.4
	Δ	+21.6	-0.1	-4.1	-6.2	+1.1	+13.8
Loose	Baseline	57.5	99.9	21.4	87.0	9.8	46.9
	CrystalJTF	76.6	99.9	17.3	80.8	10.5	58.8
	Δ	+19.1	+0.0	-4.1	-6.2	+0.7	+11.9

As expected, the absolute novelty rate decreases when the matcher is made more permissive. However, the paired novelty gain remains large at 21.4, 21.6, and 19.1 points under the strict, default, and loose rules. This persistence supports robustness to the tested matching tolerances. Uniqueness remains essentially saturated, while MSUN improves by 11.9–13.8 points and the SUN gain remains positive under every rule. In this MatterGen–Alex-MP-20 setting, higher novelty accompanies improvements in SUN and MSUN despite lower S and MS rates. These joint gains indicate a higher yield of unique, novel candidates meeting the respective stability thresholds, not a uniform improvement in all component metrics.

5.4 ABLATING DIRECTION TRANSPORT AND ATOM-WISE STEP CALIBRATION

We isolate the two components of CrystalJTF on unconditional DiffCSP–MP-20. All five configurations use the same atom-count schedule and 10,000-sample order, followed by the same post-generation relaxation and scoring protocol. The unguided DiffCSP serves as the baseline, while the four guided variants form a 2×2 design over the guidance direction (direct application of the endpoint force without pullback, or Point Q-JTF) and the step rule (uniform or Relative-BB). Table 3 reports all five configurations; the last two columns give gains over the baseline.

The ablation supports three conclusions. First, force guidance itself is useful: applying the endpoint force without pullback with a uniform step improves SUN/MSUN over the unguided baseline by 0.69/2.58 points. Second, Point Q-JTF consistently improves the direction. Replacing this endpoint-force direction with Point Q-JTF raises SUN/MSUN by 0.46/0.66 points under uniform steps and by 0.63/3.57 points under Relative-BB. Third, Relative-BB depends on direction quality. It provides only 0.20/0.38 points when applied without pullback, but 0.37/3.29 points after Point Q-JTF corrects the direction. The corresponding difference-in-differences interaction is positive, reaching 2.91 points for MSUN. Thus, the two modules are complementary rather than redundant: the transported direction makes atom-wise step calibration substantially more effective, and the complete

CrystalJTF configuration achieves the best S, MS, N, SUN, and MSUN among all five configurations while maintaining the uniqueness rate.

Table 3: Ablation of force transport and atom-wise scaling on DiffCSP–MP-20. Unconditional generation uses 10,000 paired samples. Values are percentages, and Δ denotes the gain over the unguided baseline in percentage points.

Method	N	U	S	MS	**SUN**	**MSUN**	ΔSUN	ΔMSUN
Baseline	90.53	97.09	10.38	53.69	8.50	44.83	–	–
Endpoint force (no pullback) + Uniform	91.22	97.13	10.94	55.68	9.19	47.41	+0.69	+2.58
Point Q-JTF + Uniform	91.03	97.15	11.50	56.52	9.65	48.07	+1.15	+3.24
Endpoint force (no pullback) + Relative-BB	91.09	97.02	11.14	56.15	9.39	47.79	+0.89	+2.96
Point Q-JTF + Relative-BB	91.41	97.15	11.89	59.53	10.02	51.36	+1.52	+6.53

5.5 Case study

Figure 3 compares paired CrysLLMGen outputs for K_2RbTlF_6 from the same initial candidate and diffusion seed. After CHGNet 0.3.0 relaxation, Baseline has signed $E_{\text{hull}} = 0.5648$ eV/atom, whereas CrystalJTF reaches -0.0088 eV/atom, crossing the evaluation's stability threshold. CHGNet 0.2.0 independently yields the same transition (0.5829 versus -0.0080 eV/atom). This selected strong-improvement case illustrates the contrast between paired sampling branches, not direct optimization of the final Baseline structure. Stability is assessed using surrogate potentials and a shared reference, without DFT confirmation. Appendix C.3 describes the selection criteria and shared rendering settings for this paired comparison.

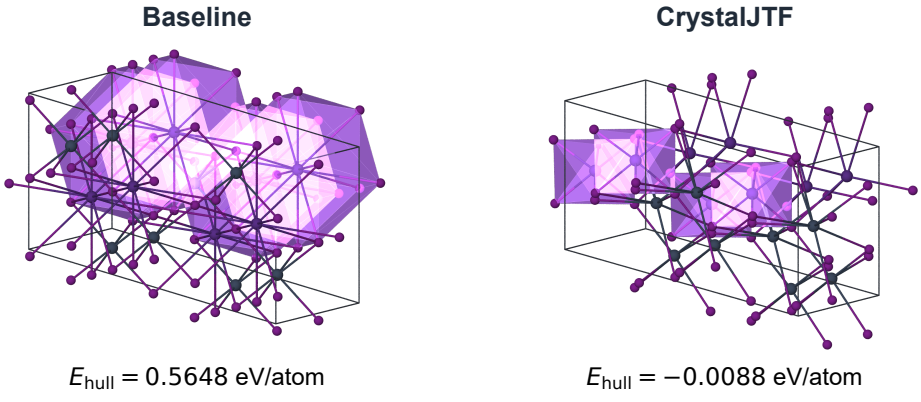


Figure 3: Paired structural comparison of K_2RbTlF_6 generated with and without CrystalJTF. After CHGNet 0.3.0 relaxation, CrystalJTF meets the stability criterion, while Baseline remains unstable. Shading shows Tl–F coordination polyhedra; both views use identical rendering rules.

6 Conclusion

We introduced CrystalJTF, a training-free, plug-and-play framework combining endpoint force pullback with atom-adaptive scaling while preserving native lattice and element-type sampling. Across three generators and multiple datasets in their respective evaluation settings, CrystalJTF improves SUN and MSUN. Ablation and robustness studies support component complementarity and consistent gains. These results highlight reusable inference-time physical feedback as a means to improve the joint novelty and stability of generated crystals. Future work will investigate integrating physical constraints into generator training, allowing explicit physical feedback to shape the learned structural prior and complement the inference-time guidance studied here. We will also explore physical guidance in latent spaces, studying how endpoint-based feedback can be transported through learned representations of crystal structure. These extensions will require careful treatment of periodic symmetry and geometric validity, alongside evaluation of how training-time and latent-space constraints affect the balance between thermodynamic stability and structural novelty.

AI USE STATEMENT

We used generative AI tools to assist with developing and refining research ideas, discussing the methodological framework, deriving equations, designing experiments, implementing and debugging code, running experimental workflows, analyzing and interpreting results, and drafting, translating, and revising the manuscript. AI tools also assisted with literature searches and figure preparation. The authors take responsibility for the final manuscript, implementation, experimental results, and scientific claims, including all AI-assisted content.

ETHICS STATEMENT

This work studies computational crystal generation. Stability estimates based on machine-learning potentials do not establish experimental synthesizability or safety. Generated candidates require further validation before practical use, and potential hazards associated with their constituent elements should be assessed before synthesis.

REPRODUCIBILITY STATEMENT

The method and experimental setup are described in the main text. Appendices A and B provide implementation details, generator-specific configurations, and evaluation protocols. Appendices C and D describe the energy-profile analysis and sample-size sensitivity experiments, respectively. An anonymized compact reference implementation, including the core guidance components, generator-specific endpoint adapters, and a minimal example, is available at `https://anonymous.4open.science/r/CrystalJTF-605C/`.

REFERENCES

Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, 2024. doi: 10.1038/s41467-024-54639-7.

Sterling G. Baird, Hasan M. Sayeed, Joseph Montoya, and Taylor D. Sparks. matbench-genmetrics: A Python library for benchmarking crystal structure generative models using time-based splits of Materials Project structures. *Journal of Open Source Software*, 9(97):5618, 2024. doi: 10.21105/joss.05618.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 843–852, 2023. URL `https://openaccess.thecvf.com/content/CVPR2023W/GCV/html/Bansal_Universal_Guidance_for_Diffusion_Models_CVPRW_2023_paper.html`.

Jonathan Barzilai and Jonathan M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988. doi: 10.1093/imanum/8.1.141.

Ilyes Batatia, David P. Kovács, Gregor Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11423–11436, 2022. doi: 10.52202/068431-0830.

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

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=OnD9zGAGT0k`.

Auguste de Lambilly, Vladimir Baturin, David Portehault, Guillaume Lambard, Nataliya Sokolovska, Florence d'Alché-Buc, and Jean-Claude Crivello. Finetuning-free diffusion model with adaptive constraint guidance for inorganic crystal structure generation. *arXiv preprint arXiv:2604.13354*, 2026. URL https://arxiv.org/abs/2604.13354.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023. doi: 10.1038/s42256-023-00716-3.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=o3BxOLoxm1.

Philipp Höllmer and Stefano Martiniani. Open materials generation with inference-time reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=xfHppnGXaH.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. doi: 10.1063/1.4812323.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17464–17497, 2023. doi: 10.52202/075280-0767.

Subhojyoti Khastagir, Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. LLM meets diffusion: A hybrid framework for crystal material generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 169190–169215, 2025. doi: 10.52202/085713-5635.

Yunyang Li, Lin Huang, Luojia Xia, Wenhe Zhang, and Mark Gerstein. Elign: Equivariant diffusion model alignment from foundational machine learned force fields. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=7u1EZZcwqX.

Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M. Wood. FlowMM: Generating materials with Riemannian flow matching. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35664–35686, 2024. URL https://proceedings.mlr.press/v235/miller24a.html.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source Python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. doi: 10.1016/j.commatsci.2012.10.028.

Hyunsoo Park and Aron Walsh. Guiding generative models to uncover diverse and novel crystals via reinforcement learning. *Nature Machine Intelligence*, 8(7):1087–1099, 2026. doi: 10.1038/s42256-026-01262-4.

Anuroop Sriram, Benjamin Kurt Miller, Ricky T. Q. Chen, and Brandon M. Wood. FlowLLM: Flow matching for material generation with large language models as base distributions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 46025–46046, 2024. doi: 10.52202/079017-1464.

Liming Wu, Wenbing Huang, Rui Jiao, Jianxing Huang, Liwei Liu, Yipeng Zhou, Hao Sun, Yang Liu, Fuchun Sun, Yuxiang Ren, and Ji-Rong Wen. Siamese foundation models for crystal structure prediction. *Nature Communications*, 17(1):6103, 2026. doi: 10.1038/s41467-026-72362-3.

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. URL https://openreview.net/forum?id=03RLpj-tc_.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23174–23184, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yu_FreeDoM_Training-Free_Energy-Guided_Conditional_Diffusion_Model_ICCV_2023_paper.html.

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

A Complete Method and Implementation Details

A.1 Coordinate, endpoint, and indexing conventions

For a structure with N atoms, fractional coordinates are represented by $X \in [0,1)^{N\times 3}$ and Cartesian coordinates by $R = XL$, where L is the current lattice matrix. The coordinate guidance never modifies L or the current element state A. All Cartesian displacements are converted back to fractional coordinates using right multiplication by L^{-1} and wrapped into the unit cell. Here A suppresses the response-time index of A_τ; $\widehat{A}$ denotes the physical endpoint element assignment supplied to the potential. Both are held fixed during coordinate differentiation.

Let τ denote the response time of the denoiser. For the score-like coordinate output $p_\theta(X, L, A, \tau)$ used by the base generator, the clean-coordinate estimator is

$$Q_\tau(X; L, A) = \begin{cases} X, & \tau = 0, \\ X - c_\tau p_\theta(X, L, A, \tau), & \tau > 0, \end{cases} \qquad c_\tau = \sigma_\tau^2 \sqrt{\nu_\tau}, \tag{10}$$

where σ_τ and ν_τ are inherited from the coordinate-noise schedule of the base model; specifically, $\nu_\tau := \texttt{sigmas_norm}[\tau]$ in the implementation. The Cartesian endpoint map used for force transport is $\mathcal{Y}_\tau(R) = Q_\tau(RL^{-1}; L, A)L$.

We distinguish diffusion time τ from the guidance-history index k. Diffusion time selects the denoiser and noise level, whereas k indexes successive states at which valid force guidance is attempted for the same sample. The Relative-BB secant uses consecutive history states $k-1$ and k; it does not assume that their integer diffusion-time difference is one.

A.2 Complete CrystalJTF sampling algorithm

A.2.1 Inputs

CrystalJTF receives a coordinate candidate X_k, lattice L_k, atomic species A, response time τ, endpoint estimator Q_τ, differentiable potential E_ϕ, Relative-BB clipping factor κ, geometric fraction ρ, and minimum-distance threshold d_{valid}. Each sample stores at most one previous pair (X_{k-1}, G_{k-1}).

A.2.2 Unified Update

For each guidance opportunity, the implementation performs the following operations in order.

1. Predict the fractional endpoint $\widehat{X}_{0|\tau} = Q_\tau(X_k; L_k, A)$ and reconstruct $\widehat{R}_{0|\tau} = \widehat{X}_{0|\tau} L_k$.
2. Evaluate the endpoint force $F_k = -\nabla_{\widehat{R}_{0|\tau}} E_\phi(\widehat{R}_{0|\tau}, L_k, \widehat{A})$.
3. Reparameterize the current candidate by Cartesian coordinates and compute the vector–Jacobian product
$$U_k = J_{\mathcal{Y}_\tau}(R_k)^\top F_k \tag{11}$$
with one reverse-mode automatic-differentiation pass. The endpoint force is treated as a fixed cotangent, so this step does not differentiate through the potential model.
4. Remove global translation using $P = I_N - N^{-1}\mathbf{1}\mathbf{1}^\top$, set $\widetilde{U}_k = PU_k$, and define the energy-gradient convention $G_k = -\widetilde{U}_k$.
5. If no valid history exists, store (X_k, G_k) and leave X_k unchanged. Otherwise, compute the minimum-image Cartesian secant
$$S_k = P\ \mathrm{MIC}_{L_k}((X_k - X_{k-1})L_k)\,, \qquad \Delta G_k = G_k - G_{k-1}. \tag{12}$$
6. For each atom, calculate a BB1 value only when the local secant has positive curvature. Normalize valid values by their median, clip the resulting relative weights to $[\kappa^{-1}, \kappa]$, and assign unit weight to invalid atoms. With D_{η_k} denoting the diagonal matrix of weights, construct
$$W_k = PD_{\eta_k}\widetilde{U}_k, \qquad V_k = \frac{W_k}{\max(\|W_k\|_F, \epsilon_n)}, \qquad \epsilon_n = 10^{-6}. \tag{13}$$

7. Build the six-nearest-neighbor periodic geometry. Let ℓ_i be the median of the six periodic neighbor distances of atom i. Compute the geometry-controlled amplitude

$$a_k = \min\left\{ \min_{\|V_{k,i}\|_2 > \epsilon_v} \frac{\rho \ell_i}{\|V_{k,i}\|_2}, \min_{\substack{j \in \mathcal{N}_6(i) \\ \|V_{k,j} - V_{k,i}\|_2 > \epsilon_v}} \frac{\rho d_{ij}^{\mathrm{PBC}}}{\|V_{k,j} - V_{k,i}\|_2} \right\}. \tag{14}$$

8. Form and wrap the candidate

$$X_k^{\mathrm{trial}} = \mathrm{wrap}\big(X_k + a_k V_k L_k^{-1}\big). \tag{15}$$

   Accept it only if the periodic geometry is finite and $d_{\min}^{\mathrm{PBC}}(X_k^{\mathrm{trial}}, L_k) \geq d_{\mathrm{valid}}$; otherwise the current coordinates remain unchanged.

9. Commit the current history pair (X_k, G_k) only if the current state and direction remain valid and the proposal has not been rejected. On rejection, clear the history and do not store the current pair.

The production configuration neither tests multiple candidate amplitudes nor selects proposals by post-update energy. The periodic geometry cap alone determines the global amplitude.

A.3 Degenerate cases and fallback rules

Explicit fallbacks prevent failed guidance from partially modifying a sample trajectory.

- **Non-finite endpoint or force.** Clear the affected sample's history and skip its update.
- **Invalid current geometry.** If the current structure has non-finite geometry, an ill-conditioned lattice, fewer than six recoverable periodic neighbors, or a minimum periodic distance below d_{valid}, the update is skipped and its history is cleared.
- **History initialization.** The first valid state stores (X_k, G_k) without a Relative-BB update.
- **Weak transported force.** After translation removal, an ℓ_1 norm no greater than 10^{-3} causes the update to be skipped while retaining the current state as the next history anchor.
- **Invalid local curvature.** Atom i receives a BB value only if $\|S_{k,i}\|_2^2 > 10^{-16}$, $S_{k,i}^\top \Delta G_{k,i} > 0$, and the resulting ratio is finite and positive. Otherwise $\eta_{k,i} = 1$.
- **No positive-curvature atoms.** The median normalizer and all atom weights default to one.
- **Degenerate weighted direction.** Non-finite centered weighted directions or Frobenius norms at most 10^{-12} trigger fallback to the centered JTF direction with unit atom weights.
- **Zero geometric amplitude.** Atom and pair speeds no greater than 10^{-12} are omitted from the corresponding geometric ratios. If no positive finite bound remains, skip displacement.
- **Invalid proposal.** A wrapped proposal with minimum periodic distance below d_{valid} is rejected atomically; no partial coordinate update is committed, and its BB history is cleared.

A.4 Hyperparameters and computational cost

Table 4 lists the fixed controller parameters used in the main experiments.

We set $\rho = 0.15$, limiting each atomic displacement to at most 15% of its local reference distance ℓ_i, and each retained neighbor-relative displacement to at most 15% of the corresponding periodic distance d_{ij}^{PBC}. These are upper bounds, not prescribed displacements for every atom. The parameter controls the allowed geometric perturbation, rather than selecting a step by energy decrease. We use the same value across all six generator–benchmark settings and observe positive SUN and MSUN gains in each setting under both evaluation-potential versions (Table 1), supporting the transferability of this default setting across generators and benchmarks.

For each active guidance opportunity, CrystalJTF requires one endpoint-force evaluation by the surrogate potential and one reverse pass through the endpoint map. The latter computes a VJP and therefore avoids the quadratic memory cost of materializing the full coordinate Jacobian. Relative-BB uses $O(N)$ stored history per sample. Once the local neighbor list is built, atom-wise weighting and geometry-cap calculations scale linearly with the numbers of atoms and retained neighbor edges.

Table 4: CrystalJTF hyperparameters and numerical thresholds.

Quantity	Symbol	Value
Relative-BB clipping factor	κ	2
Periodic geometry fraction	ρ	0.15
Periodic neighbors per atom	$\lvert\mathcal{N}_6(i)\rvert$	6
Minimum valid distance	d_{valid}	0.5 Å
Weak-direction threshold	–	10^{-3} (ℓ_1 norm)
Secant-squared threshold	–	10^{-16}
Direction-norm threshold	–	10^{-12}
Normalization denominator floor	ϵ_n	10^{-6}
Geometry-speed threshold	ϵ_v	10^{-12}
Energy candidate selection	–	disabled

A.5 Local descent of the coordinate correction

Fix one guidance opportunity and suppress its index. The response time, lattice, generator element state, and selected physical endpoint assignment are held constant. Assume the Cartesian endpoint map $\mathcal{Y}$ and the potential are differentiable locally on the selected periodic coordinate branch. In particular, the derivative does not pass through discrete element selection or a change of periodic branch. With $\mathcal{E}(R) = E_\phi(\mathcal{Y}(R), L, \widehat{A})$ and $F = -\nabla_{\widehat{R}} E_\phi$, the chain rule yields $U = J_{\mathcal{Y}}(R)^\top F = -\nabla_R \mathcal{E}(R)$.

Let $P = I_N - \mathbf{1}\mathbf{1}^\top/N$ and $D_\eta = \operatorname{diag}(\eta_1, \ldots, \eta_N)$, where every weight is finite and strictly positive. Relative-BB satisfies this condition through positive clipping bounds and unit-weight fallbacks. For $W = PD_\eta PU$, symmetry and idempotence of P give

$$\begin{aligned} \langle \nabla_R \mathcal{E}, W \rangle_F &= -\langle PU, D_\eta PU \rangle_F \\ &= -\sum_{i=1}^{N} \eta_i \|(PU)_i\|_2^2 < 0 \qquad \text{if } PU \neq 0. \end{aligned} \tag{16}$$

Moreover, $\mathbf{1}^\top W = 0$. Dividing by the positive normalization denominator in Eq. (6) and multiplying by any positive finite step preserves both zero net translation and the negative directional derivative. If $PU = 0$, the zero-translation subspace has no strict descent direction at first order.

This result concerns the direction at the current state, not the accuracy of the historical BB curvature estimate. Under the local differentiability assumptions, sufficiently small positive steps decrease $\mathcal{E}$. The geometry cap is not an energy line search and does not ensure such a step size; neither finite-step energy decrease nor monotonic energy along the diffusion trajectory follows. Skipped or rejected corrections leave the native candidate unchanged and make no strict-descent claim.

B Generator-Specific Integration and Evaluation Protocols

Table 5 records the complete interface between CrystalJTF and each frozen generator. Short SHA-256 prefixes identify the exact checkpoints used in the experiments. The same MACE checkpoint is used for inference-time guidance in every setting.

Table 5: Generator-specific integration and evaluation settings. The Uniform and Relative-BB ablation arms differ only in their atom-wise relative weights; all other safeguards and evaluation rules are shared.

Configuration	CrysLLMGen	DiffCSP	MatterGen
Generator and endpoint estimator			
Generator checkpoint	MP-20: Llama-3-8B + 5-epoch LoRA (`0b3f45c2`) and continuous refiner `model_final.pt` (`fe88044b`). MPTS-52: Llama LoRA `final` and refiner `last.pt` (`f67bf4cf`).	MP-20: official `diffusion_w_type last.ckpt` (`82c0b980`). MPTS-52: dataset-specific unconditional `diffusion_w_type last.ckpt` (`8ce2c3d1`).	MP-20: official `mp_20_base` checkpoint (`ffb80e44`). Alex-MP-20: official `mattergen_base` checkpoint (`81668ee1`), MatterGen 1.0.3.
Q_τ construction	Local periodic Tweedie endpoint from the coordinate refiner, $Q_\tau(X) = X - \sigma_\tau^2\sqrt{\nu_\tau}\, p_\theta(X,\tau)$; $Q_0(X) = X$.	The same local periodic Tweedie form, evaluated through the frozen DiffCSP decoder at the current coordinate state.	Local unwrapped score estimate, $Q_t(X) = X + \mathrm{Var}_t[X]\, s_\theta(X,t)$, using MatterGen's official marginal variance, including its atom-count adjustment.
Species and lattice	Species proposed by the LLM are fixed during refinement. CrystalJTF updates coordinates only; the native lattice trajectory is unchanged.	Native soft element and lattice diffusion are retained. Clean-label argmax species are used only for MACE; BB history is reset when these labels change. No lattice guidance is added.	Revealed species are retained and masked species use the clean real-element argmax only for MACE; label changes reset BB history. Native species and lattice sampling are otherwise unchanged.
Guidance and safety controller			
Guidance potential	MACE `2023-12-10-mace-128-L0_energy_epoch-249.model`, MACE-Torch 0.3.13, SHA-256 `2ddb079c`. Forces are evaluated on Q_τ and pulled back only through the coordinate endpoint map.		
Schedule and frequency	Last 200 of 1,000 denoising steps, applied at every step ($\tau = 200, \ldots, 1$).	Last 200 of 1,000 denoising steps, applied at every step ($\tau = 200, \ldots, 1$).	Last 200 of 1,000 denoising steps, applied at every step (continuous time $t \le 0.2$, including the endpoint).
Controller parameters	$\kappa = 2$, $\rho = 0.15$, six periodic neighbors per atom, and $d_{\mathrm{valid}} = 0.5$ Å. Positive atom-wise BB1 values are median-normalized and clipped to $[1/\kappa, \kappa]$; invalid curvature receives unit weight.		
Uniform vs. Relative-BB	Both remove rigid translation, use one joint Frobenius normalization, apply the same periodic geometry cap, wrap the candidate, and perform the same validity check. Uniform fixes every relative weight to one; Relative-BB changes only the atom-wise relative weights.		
Skip and fallback rules	Skip on non-finite endpoint/force, invalid or ill-conditioned geometry, weak or degenerate transported direction, non-positive finite geometry bound, or an invalid proposal. The first valid opportunity initializes history without a BB displacement; changing clean species clears the affected history; a rejected proposal leaves the native sampler state unchanged.		
Post-relaxation evaluation			
Novelty reference	MP-20 training split or MPTS-52 training split, matched to the generator checkpoint.	MP-20 training split or MPTS-52 training split, matched to the generator checkpoint.	MP-20 training split for `mp_20_base`; the official MP2020-corrected Alex-MP reference (`c722f72c`) for `mattergen_base`.
Relaxation and energy	CHGNet 0.2.0 and 0.3.0 are used independently as evaluation potentials. FIRE relaxes coordinates and cell to $0.1\,\mathrm{eV/Å}$ or 500 steps. Signed E_{hull} is computed from a shared frozen Materials Project phase-diagram cache; unavailable entries are `thermo_unknown`, never reassigned as unstable.		

Continued on next page

Table 5: Generator-specific integration and evaluation settings (continued).

Configuration	CrysLLMGen	DiffCSP	MatterGen
Metric denominator	All N, U, S, MS, SUN, and MSUN rates use the fixed requested sample count, including relaxation failures and thermodynamic unknowns, as the denominator.		

C Energy-Profile Analysis and Visualization

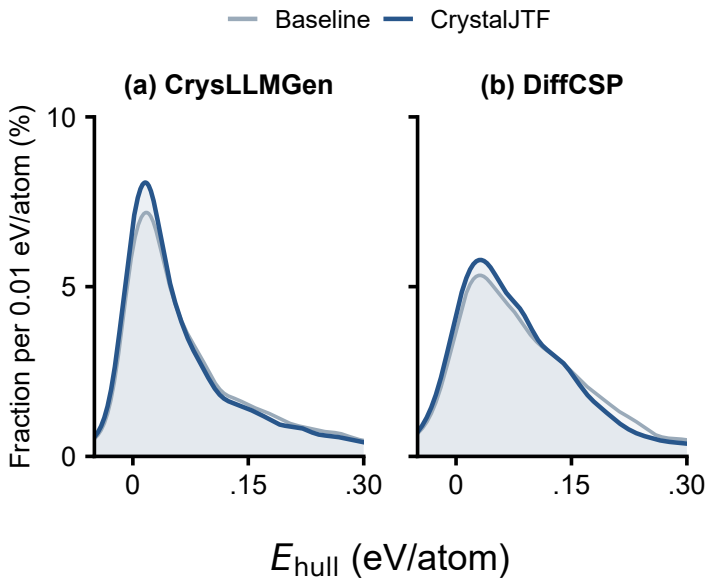


Figure 4: Post-relaxation energy distributions on MP-20. CHGNet 0.3.0 is used for (a) CrysLLMGen and (b) DiffCSP. Curves show percentages of all 10,000 generated structures per 0.01 eV/atom. Plotting details are provided in this appendix.

Figure 4 uses the MP-20 outputs after CHGNet 0.3.0 relaxation and the Materials Project thermodynamic evaluation described in Section 5.1. Each arm contains 10,000 structures. Energy counts are 9,520/9,520 for CrysLLMGen and 9,782/9,815 for DiffCSP (Baseline/CrystalJTF).

C.1 Smoothed Energy Distributions

Panels (a,b) use Gaussian kernel density estimates with a common pooled robust Silverman bandwidth for both arms of each model: approximately 0.01136 eV/atom for CrysLLMGen and 0.01232 eV/atom for DiffCSP. The curves are scaled to percentages of the full generated population per 0.01 eV/atom interval. Unavailable energies remain in the fixed 10,000-sample denominator but do not contribute to the density curves. Medians reported in the text are computed from the available raw energies, without smoothing. All finite evaluated energies contribute to the calculations, including those outside the displayed interval $[-0.05, 0.30]$ eV/atom. Signed negative energies are retained under the shared reference convention.

C.2 Interpretation

The shaded regions show areas under the density curves, not confidence intervals. The figure describes marginal energy distributions and does not establish statistical significance or an energy decrease for every paired structure. Its scope is CrysLLMGen and DiffCSP on MP-20; the joint novelty and stability metrics for all generators remain reported in Table 1.

C.3 PAIRED CRYSTAL CASE STUDY

Figure 3 shows a selected CrysLLMGen pair for K_2RbTlF_6 (sample 7877), with the same initial candidate, composition, atom count, and diffusion seed. After CHGNet 0.3.0 relaxation, the signed E_{hull} decreases from 0.5648 eV/atom for Baseline to -0.0088 eV/atom for CrystalJTF, crossing our stability threshold. Independent relaxation and scoring with CHGNet 0.2.0 gives the same qualitative transition, from 0.5829 to -0.0080 eV/atom. Both arms converge under the prescribed relaxation settings in both versions.

OVITO renders the CIF structures in Figure 3 with a common camera and length scale. Shaded regions show Tl–F coordination polyhedra; connections use the same 3.6 Å metal–F distance cutoff in both arms and do not denote bond orders. A $2 \times 2 \times 1$ display supercell and boundary-neighbor images complete the coordination environments; each original cell has 10 atoms.

For this illustration, we screened held-out sample IDs 1000–9999 for converged pairs with Baseline $E_{\text{hull}} > 0.1$ eV/atom and guided $E_{\text{hull}} < 0$ under both CHGNet versions. Eight pairs met these criteria. Among cases with at most 12 atoms per cell, we selected the largest reduction under CHGNet 0.3.0. This is therefore a strong-improvement example rather than a representative estimate of average performance. The two structures are outputs of paired sampling branches, not a direct relaxation path from the Baseline structure to the guided structure. Stability refers to the surrogate-potential and shared-reference evaluation; this case lacks independent DFT validation.

D SAMPLE-SIZE SENSITIVITY OF UNIQUENESS-BASED METRICS

Crystal-generation studies commonly report uniqueness-dependent metrics at different generation budgets. This makes cross-paper comparisons difficult: uniqueness is defined relative to the other structures in the evaluated set, so its value can change even when the generator and evaluation protocol are unchanged. We quantify this effect for CrysLLMGen–MP-20 by comparing paired Baseline and CrystalJTF evaluations with 1,000 and 10,000 requested samples. Both budgets use the same generator and guidance checkpoints, CHGNet 0.3.0 relaxation, novelty reference, structure matcher, thermodynamic thresholds, and fixed-denominator post-relaxation protocol described in Appendix B. At the 1,000-sample budget, all structures are relaxed successfully; 950 structures in each arm have available thermodynamic references, while the same 50 structures remain `thermo_unknown` in both arms.

Table 6: Sensitivity to sample size on CrysLLMGen–MP-20. Evaluation uses CHGNet 0.3.0. All entries are percentages; blue Δ rows give paired CrystalJTF gains in percentage points at the same sample budget.

Samples	Method	N	U	S	MS	**SUN**	**MSUN**
1,000	Baseline	82.20	96.70	14.70	66.50	9.50	47.50
	CrystalJTF	83.30	96.40	17.40	71.50	11.50	53.20
	Δ	+1.10	-0.30	+2.70	+5.00	+2.00	+5.70
10,000	Baseline	82.73	77.16	15.10	65.91	6.77	37.08
	CrystalJTF	83.53	77.54	17.08	70.31	8.28	41.82
	Δ	+0.80	+0.38	+1.98	+4.40	+1.51	+4.74

D.1 CRYSTALJTF REMAINS EFFECTIVE ACROSS SAMPLING BUDGETS

At 1,000 samples, CrystalJTF improves SUN/MSUN by 2.00/5.70 percentage points; at 10,000 samples, it retains gains of 1.51/4.74 points. The corresponding improvements in S and MS are also positive at both budgets. Thus, the main conclusion is not specific to the smaller evaluation: the guidance benefit persists when the generation budget is increased by an order of magnitude.

D.2 Sample Count as an Evaluation Confounder

The absolute gap between the two budgets is substantial even though the energy-related rates remain similar. For Baseline, S changes only from 14.70 to 15.10 and MS from 66.50 to 65.91, whereas U falls from 96.70 to 77.16. CrystalJTF shows the same pattern: S changes from 17.40 to 17.08 and MS from 71.50 to 70.31, while U falls from 96.40 to 77.54. Consequently, Baseline SUN/MSUN decrease from 9.50/47.50 to 6.77/37.08, and CrystalJTF SUN/MSUN decrease from 11.50/53.20 to 8.28/41.82. This behavior follows from the set-relative definition of uniqueness: as more structures are generated, additional within-set duplicates reduce U and therefore the joint SUN/MSUN rates, without implying worse thermodynamic quality.

This comparison exposes an often-overlooked source of variation in reported crystal-generation results. Uniqueness-dependent metrics computed with different requested sample counts are not directly comparable. We therefore recommend that studies report the exact generation budget and metric denominator, and that method comparisons use matched sample counts. The paired gains within each row block of Table 6 are the appropriate quantities for evaluating CrystalJTF; the absolute 1k and 10k values should not be interpreted as a like-for-like ranking across budgets.