EP-Flow: Disordered Crystal Structure Prediction without Site-Level Annotations

Qiuliang Liu[1,2,†], **Liming Wu**[3,†], **Qi Li**[1,2], **Zhonglong Peng**[1,2], **Chang Chen**[1,2]
Xiaolong Chen[1,2], **Wenbing Huang**[3,*], **Shifeng Jin**[1,2,*]

[1] Institute of Physics, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Gaoling School of Artificial Intelligence, Renmin University of China

[†] These authors contributed equally to this work.
[*] Corresponding authors.

Abstract

Generative models have made rapid progress in ordered crystal structure prediction, yet many functional materials are intrinsically disordered, with substitutional mixing, vacancies, or interstitial species controlling their properties. Existing crystal generators either assume deterministic site occupations or require site-level disorder annotations, which are often unavailable when the chemical formula is the primary input. We formulate disordered crystal structure prediction through an Occupancy Distribution Matrix (ODM), a continuous site-by-species representation that unifies ordered crystals, solid solutions, vacancy disorder, and interstitial occupancy. A valid ODM must satisfy coupled site-wise occupancy, mass-conservation, and non-negativity constraints, placing each sample on a formula-dependent transportation polytope. We propose Entropic Polytope Flow (EP-Flow), a marginal-constrained flow matching framework that canonicalizes heterogeneous polytopes into a shared double-centered space, learns a marginal-preserving flow, and recovers feasible occupancies through a Sinkhorn inverse map. By jointly generating occupancies, fractional coordinates, and lattice parameters, EP-Flow achieves state-of-the-art performance on formula-conditioned disordered CSP benchmarks derived from COD and MPDS, substantially outperforming adapted ordered-crystal generators. Analyses further show that EP-Flow recovers sparse and chemically meaningful local disorder patterns rather than merely matching global composition statistics.The project repository is available at https://github.com/LongK-LQL/EP-Flow.

1 Introduction

Crystalline materials underpin technologies ranging from semiconductors and batteries to solid electrolytes, catalysts, magnets, and superconductors. Recent machine learning advances have accelerated Crystal Structure Prediction (CSP) and generative materials design, with models such as CDVAE [1], GNoME [2], DiffCSP [3], FlowMM [4] and MatterGen [5] demonstrating remarkable progress. However, most existing methods are built around ordered structures, often derived from DFT calculations at 0 K, where each crystallographic site is occupied by a single deterministic element. Real materials frequently exhibit crystallographic disorder under synthesis or operating conditions, including substitutional mixing, vacancies, and interstitial species. This mismatch is substantial rather than marginal: nearly 47% of verified ICSD structures [6] contain some form of disorder. Such disorder can govern ionic transport, thermal stability, magnetism, and electronic behavior, motivating generative models that explicitly handle disordered crystals.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Disordered CSP is not merely ordered CSP with soft labels. In conventional ordered CSP, once a cell size or formula-unit convention is specified, the target composition fixes the species to be placed, and each generated site is assigned one deterministic species. In disordered crystals, the model must also infer how stoichiometric mass is allocated over crystallographic sites: multiple species may share a site, a site may be partially vacant, or an interstitial position may be fractionally occupied. For common substitutional solid solutions, the site budget is usually implied by the formula. However, for vacancy and interstitial disorder, the reported composition is insufficient to specify all crystallographic sites. To resolve this, we extend the stoichiometric vector $\boldsymbol{c}$ by incorporating a vacancy pseudo-species (Vac). This ensures every site is fully accounted for, satisfying $\sum_{j=1}^{M} c_j = N$, where N is the total site count, M is the number of candidate species (including Vac), and c_j is the target count for species j. This yields an Occupancy Distribution Matrix (ODM) encoding site-wise distributions. A valid ODM must satisfy three coupled constraints: site-wise full occupancy, mass conservation matching $\mathbf{c}$, and non-negativity. We illustrate diverse types of disordered structures in Appendix D.1 and detail the mapping between macroscopic formulas and microscopic site counts in Appendix D.2.

Together, these requirements define a formula-dependent transportation polytope, as illustrated in Fig. 1. Unlike standard crystal generators that predict site labels independently or correct compositions after generation, ODM generation is globally coupled: changing one site occupancy changes the feasible allocations at other sites. Moreover, the feasible domain changes with both composition and site budget. Standard diffusion or flow models are therefore difficult to apply directly, because their trajectories must respect a heterogeneous family of constrained domains rather than a single Euclidean space.

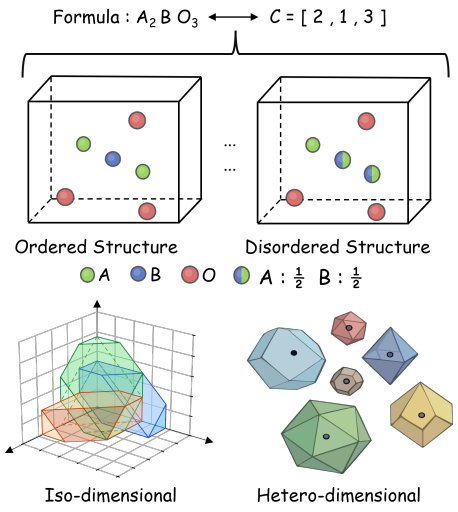


Figure 1: Formula transportation polytopes.

To address this challenge, we propose EP-Flow, a marginal-constrained flow matching framework for disordered CSP without site annotation. EP-Flow canonicalizes each formula-dependent transportation polytope into a shared double-centered linear subspace by applying log-domain double centering to the ODM. In this space, the model learns a marginal-preserving flow whose velocity field remains in the zero-marginal tangent space. A Sinkhorn-based inverse map then recovers an ODM on the target transportation polytope, enforcing row-sum, column-sum, and non-negativity constraints. Built on this occupancy formulation, EP-Flow jointly generates occupancies, fractional coordinates, and lattice parameters.

We evaluate EP-Flow on large-scale COD [7] and MPDS [8] benchmarks covering substitutional disorder, vacancy-inclusive disorder, and systems with up to N=50 crystallographic sites. Compared to the adapted baselines, EP-Flow achieves state-of-the-art performance under this no-site-annotation CSP setting, substantially improving structure match rate, occupancy accuracy, formula consistency, and validity. Case studies further show that EP-Flow recovers sparse and chemically meaningful disorder motifs rather than merely matching global composition statistics.

In summary, our contributions are threefold: **(1)** we formulate disordered CSP without site annotation as generation over formula-dependent transportation polytopes using an ODM representation that unifies ordered crystals, substitutional disorder, vacancy disorder, and interstitial occupancy; **(2)** we introduce EP-Flow, a marginal-constrained flow matching framework with double-centered canonicalization and Sinkhorn recovery; and **(3)** we construct COD and MPDS benchmarks with occupancy-sensitive metrics, demonstrating state-of-the-art performance for formula-conditioned disordered CSP without site-level disorder priors.

2 Related Work

Ordered Crystal Structure Prediction. Generative modeling for crystals has progressed from latent-variable models [1, 9] to diffusion models [5, 3, 10, 11, 12, 13] and, more recently, flow-based and Bayesian flow approaches [14, 4, 15, 16, 17, 18]. These methods have substantially improved the

modeling of periodic atomic coordinates, but they are fundamentally designed for *ordered* crystals, where each site is assigned a single discrete species. By contrast, EP-Flow targets disordered crystal generation, where occupancies are fractional and coupled by global stoichiometric constraints.

Disordered Crystal Structure Prediction. Recognizing the limitations of ordered models, recent work has started to address disordered structures explicitly. Dis-Gen [19] represents crystals in a tabular form and reconstructs them with a VAE-style architecture, while DMFlow [20] uses flow matching to model substitutional and positional disorder under simplex constraints. However, these methods still rely on site-level disorder annotations, e.g., which sites are doped and what occupancies they should carry. This assumption breaks down in the more challenging setting considered here, where only the global chemical formula is given. In contrast, EP-Flow models the full occupancy distribution jointly, so that local full-occupancy and global mass-conservation constraints are handled within a unified generative process without requiring explicit site-specific supervision.

Constrained Generative Modeling on Manifolds. More broadly, our work is related to constrained generative modeling beyond materials science. Prior studies have extended diffusion and flow-based models to Riemannian manifolds, simplices, and more general polytopes [21, 22, 23, 24, 25], using projection-based corrections [26] or barrier/mirror-map formulations [27, 28]. These strategies, however, can either break trajectory smoothness or become numerically fragile near sharp polytope boundaries. Our approach instead develops a marginal-preserving flow in a shared double-centered space, followed by a Sinkhorn inverse map, yielding exact transportation-polytope constraints.

3 Preliminaries

We briefly review the two ingredients underlying our method: transportation-polytope constraints and the double-centered zero-marginal space that provides a shared representation across formulas.

3.1 Transportation Polytopes and Sinkhorn Projection

Let $\boldsymbol{r} \in \mathbb{R}_+^N$ and $\boldsymbol{c} \in \mathbb{R}_+^M$ be prescribed marginal vectors with equal total mass, i.e., $\mathbf{1}_N^\top \boldsymbol{r} = \mathbf{1}_M^\top \boldsymbol{c}$. The associated transportation polytope is defined as the set of non-negative matrices satisfying these exact marginal constraints:

$$\mathcal{P}(\boldsymbol{r}, \boldsymbol{c}) \;=\; \Big\{ \boldsymbol{P} \in \mathbb{R}_+^{N \times M} \;\Big|\; \boldsymbol{P}\mathbf{1}_M = \boldsymbol{r},\; \mathbf{1}_N^\top \boldsymbol{P} = \boldsymbol{c}^\top \Big\}. \tag{1}$$

Given a non-negative matrix $\boldsymbol{X} \in \mathbb{R}_+^{N \times M}$, the Sinkhorn algorithm [29] alternates row and column rescaling to project $\boldsymbol{X}$ onto $\mathcal{P}(\boldsymbol{r}, \boldsymbol{c})$. We denote the resulting differentiable operator by $\mathcal{S}(\boldsymbol{X}; \boldsymbol{r}, \boldsymbol{c}) \in \mathcal{P}(\boldsymbol{r}, \boldsymbol{c})$. In practice, a finite number of iterations is sufficient to satisfy the marginal constraints up to a prescribed tolerance [30]. For strictly positive inputs, this projection is differentiable and coincides with the unique Kullback-Leibler (KL) projection onto the target polytope [31].

3.2 Double-Centering Operator and Zero-Marginal Space

To map complex constrained distributions into a tractable representation, we rely on the double-centering operator. For any arbitrary matrix $\boldsymbol{X} \in \mathbb{R}^{N \times M}$, the double-centering operator $\mathfrak{D}(\cdot)$ subtracts its row and column means, expressed as:

$$\mathfrak{D}(\boldsymbol{X}) = (\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top)\boldsymbol{X}(\mathbf{I}_M - \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top). \tag{2}$$

The image of this operator defines the zero-marginal subspace:

$$\mathcal{M}_0(N, M) \triangleq \Big\{ \boldsymbol{Z} \in \mathbb{R}^{N \times M} \;\Big|\; \boldsymbol{Z}\mathbf{1}_M = \mathbf{0}_N,\; \mathbf{1}_N^\top \boldsymbol{Z} = \mathbf{0}_M^\top \Big\}. \tag{3}$$

Because $\mathcal{M}_0(N, M)$ is a linear subspace, it serves as a well-behaved Euclidean manifold whose tangent space is isomorphic to itself. This makes $\mathcal{M}_0(N, M)$ a convenient Euclidean space for defining generative vector fields while preserving the zero-sum structure induced by the constraints.

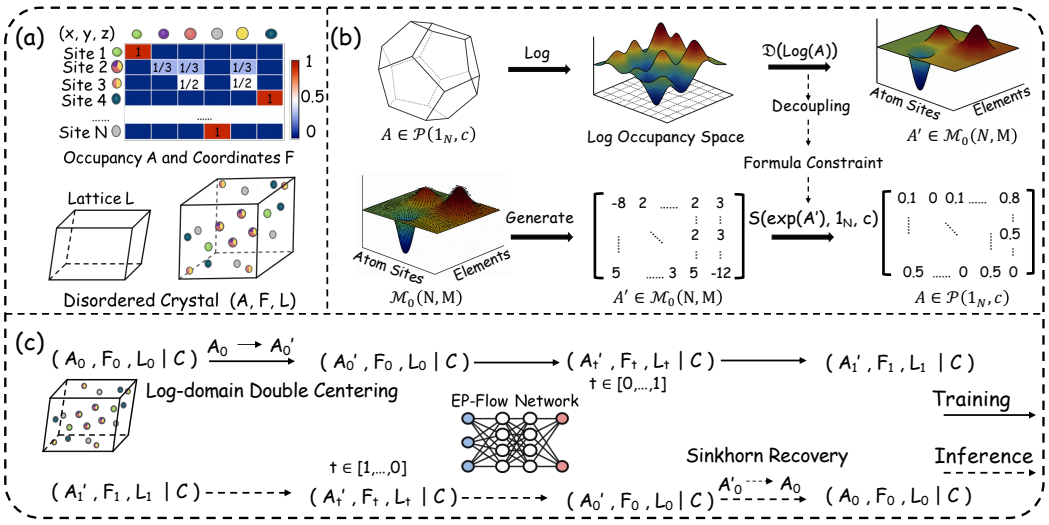


Figure 2: **Overview of EP-Flow for disordered crystal generation.** **(a)** Disordered crystal represented by $(\boldsymbol{A}, \boldsymbol{F}, \boldsymbol{l})$. **(b)** Occupancy matrix canonicalized from formula polytope $\mathcal{P}(\mathbf{1}_N, \boldsymbol{c})$ to double-centered space $\mathcal{M}_0(N, M)$ via log-domain double centering and mapped back to $\boldsymbol{A}$ via Sinkhorn projection. **(c)** Joint flow matching over $(\boldsymbol{A}', \boldsymbol{F}, \boldsymbol{l})$ conditioned on $\boldsymbol{c}$, with projection back to $\boldsymbol{A}$ for the final structure.

4 Proposed Approach: EP-Flow

4.1 Task Definition

We represent a disordered crystal as $\mathcal{D} = (\boldsymbol{A}, \boldsymbol{F}, \boldsymbol{l})$, where $\boldsymbol{A} \in [0,1]^{N\times M}$ is the ODM over N sites and M species, $\boldsymbol{F} \in [0,1)^{N\times 3}$ denotes fractional coordinates, and $\boldsymbol{l} = (a, b, c, \alpha, \beta, \gamma) \in \mathbb{R}^6_+$ denotes the lattice parameters, as shown in Fig. 2a. Given a chemical formula, we encode its stoichiometry as $\boldsymbol{c} \in \mathbb{R}^M$, where c_j is the total occupancy of species j over all sites. Physical validity requires $\boldsymbol{A}$ to satisfy row-sum constraints $\sum_{j=1}^{M} A_{ij} = 1$ for full site occupancy, column-sum constraints $\sum_{i=1}^{N} A_{ij} = c_j$ for global mass conservation, and non-negativity $A_{ij} \geq 0$. Hence, $\boldsymbol{A}$ lies in a formula-dependent transportation polytope $\mathcal{P}(\mathbf{1}, \boldsymbol{c})$. By introducing vacancy as a pseudo-species, this representation naturally covers ordered crystals, substitutional disorder, and vacancy disorder.

Disordered CSP is then formulated as learning the conditional distribution $p(\boldsymbol{A}, \boldsymbol{F}, \boldsymbol{l} \mid \boldsymbol{c})$ via flow matching; see Appendix A.5 for details. This differs fundamentally from conventional CSP formulations that assume fixed site-wise assignments and model only $p(\boldsymbol{F}, \boldsymbol{l} \mid \boldsymbol{A})$. Here the model receives only $\boldsymbol{c}$ and must generate a full crystal triplet whose occupancy matrix remains inside $\mathcal{P}(\mathbf{1}, \boldsymbol{c})$ while the occupancy, atomic positions, and lattice geometry remain mutually consistent. This shift from explicit site labels to a continuous, constraint-aware occupancy representation is the key to exploring a much richer space of disordered materials, without assuming prior knowledge of specific doping patterns or site preferences.

4.2 Occupancy Distribution Matrix Generation

The primary challenge in generating $\boldsymbol{A} \in [0,1]^{N\times M}$ is that its support, the transportation polytope $\mathcal{P}(\mathbf{1}, \boldsymbol{c})$, changes with the stoichiometry $\boldsymbol{c}$ and the number of sites N. ODMs therefore live on a heterogeneous family of constrained domains with no shared geometry, making direct generative modeling difficult. We resolve this by canonicalizing each ODM into the shared double-centered space $\mathcal{M}_0(N, M)$, where all matrices satisfy zero-sum marginals regardless of the underlying formula.

Definition 1 (Space Canonicalization)**.** *Given an ODM $\boldsymbol{A} \in \mathcal{P}(\mathbf{1}, \boldsymbol{c})$, we define a forward mapping $\Psi : \mathcal{P}(\mathbf{1}, \boldsymbol{c}) \to \mathcal{M}_0(N, M)$ by*

$$\boldsymbol{A}' = \Psi(\boldsymbol{A}) \triangleq \mathfrak{D}(\log \boldsymbol{A}) \tag{4}$$

where $\mathfrak{D}$ is the double-centering operator, as shown in Eq. (2). The image lies in $\mathcal{M}_0(N, M)$.

Intuitively, double-centering removes the additive row- and column-wise components of $\log \boldsymbol{A}$*, which can be interpreted as site- and species-specific chemical potentials, and retains only the residual interaction structure in a stoichiometry-invariant space [32]. In practice, we apply* ϵ*-smoothing to prevent numerical instability during logarithmic operations. Algorithm details are provided in Appendix D.3.*

To recover a valid ODM under a target stoichiometry $\boldsymbol{c}$, we map back from $\mathcal{M}_0(N, M)$ using a Sinkhorn projection [29]:

$$\boldsymbol{A} = \Psi^{-1}(\boldsymbol{A}') \triangleq \mathcal{S}(\exp(\boldsymbol{A}'),\, \mathbf{1},\, \boldsymbol{c})\,. \tag{5}$$

In this context, $\mathcal{S}$ acts as a projection that finds the matrix in $\mathcal{P}(\mathbf{1}, \boldsymbol{c})$ that minimizes the Kullback-Leibler (KL) divergence to the exponentiated centered matrix $\exp(\boldsymbol{A}')$, thereby restoring the physical stoichiometry. We illustrate the process in Fig. 2b, and provide more details in Appendix D.4.

As established in Eqs. (4) and (5), the function Ψ is a bijective mapping, as stated in the following proposition:

Proposition 1 (Bijectivity of Canonicalization)**.** *For any fixed stoichiometry vector* $\boldsymbol{c}$*, the mapping* Ψ *establishes a one-to-one correspondence between the interior of the transportation polytope* $\mathcal{P}(\mathbf{1}, \boldsymbol{c})$ *and the double-centered space* $\mathcal{M}_0(N, M)$*. Specifically, for any* $\boldsymbol{A}' \in \mathcal{M}_0(N, M)$*, there exists a unique matrix* $\boldsymbol{A} \in \mathit{int}(\mathcal{P}(\mathbf{1}, \boldsymbol{c}))$ *such that* $\Psi(\boldsymbol{A}) = \boldsymbol{A}'$*, which is uniquely recoverable via the Sinkhorn operator:* $\boldsymbol{A} = \mathcal{S}(\exp(\boldsymbol{A}'), \mathbf{1}, \boldsymbol{c})$*. This bijectivity ensures that the generative trajectory in* $\mathcal{M}_0(N, M)$ *maps to a unique and physically consistent path in the valid ODM space. A proof is provided in Appendix A.1.*

With this shared representation, we learn the centered ODM distribution by flow matching. We use the linear path $\boldsymbol{A}'_t = (1-t)\boldsymbol{A}'_0 + t\boldsymbol{A}'_1$ for $t \in [0, 1]$, where $\boldsymbol{A}'_1 = \Psi(\boldsymbol{A}_1)$ comes from data and $\boldsymbol{A}'_0$ is sampled from a prior supported on $\mathcal{M}_0(N, M)$. Since $\mathcal{M}_0(N, M)$ is linear, the entire path remains in $\mathcal{M}_0(N, M)$ as long as the predicted velocity is projected back by $\mathfrak{D}$, and we provide the proof of trajectory feasibility in Appendix A.2. We additionally apply a formula-derived mask to restrict training to site–species pairs that are feasible in the data. The resulting target velocity is the constant field $u_t^{\boldsymbol{A}'} = \boldsymbol{A}'_1 - \boldsymbol{A}'_0$, which supervises the projected network output in $\mathcal{M}_0(N, M)$.

4.3 Fractional Coordinates Generation

Fractional coordinates $\boldsymbol{F} \in [0, 1)^{N\times 3}$ are periodic along each lattice direction, so we define their flow on the N-torus $(\mathbb{T}^3)^N$. Given a noise sample $\boldsymbol{F}_0$ and a data sample $\boldsymbol{F}_1$, the target velocity is the shortest periodic displacement

$$u_t^{\boldsymbol{F}} = \Delta_{\mathbb{T}^3}(\boldsymbol{F}_1, \boldsymbol{F}_0) = \big((\boldsymbol{F}_1 - \boldsymbol{F}_0 + 0.5) \bmod 1\big) - 0.5. \tag{6}$$

This torus-aware displacement is the geodesic direction on the periodic coordinate space. The corresponding interpolation is

$$\boldsymbol{F}_t = \big(\boldsymbol{F}_0 + t u_t^{\boldsymbol{F}}\big) \bmod 1, \qquad t \in [0, 1]. \tag{7}$$

The model is trained to predict $u_t^{\boldsymbol{F}}$ from $\boldsymbol{F}_t$, ensuring periodicity is respected throughout the trajectory.

4.4 Lattice Parameters Generation

Following FlowMM [4], we model the lattice parameters $\boldsymbol{l} = (a, b, c, \alpha, \beta, \gamma)$ in an unconstrained latent space for stable learning. The lengths (a, b, c) are parameterized in log space to guarantee positivity, while the angles (α, β, γ) are parameterized by unconstrained logits and mapped to the Niggli-reduced range $[60°, 120°]$. This yields a canonical lattice representation (i.e., $\boldsymbol{l}^{\mathrm{unc}}$) while preserving the coupled statistics between cell geometry and atomic arrangement. Flow matching supervision is defined in the unconstrained space with target velocity $u_t^{\boldsymbol{l}} = \boldsymbol{l}_1^{\mathrm{unc}} - \boldsymbol{l}_0^{\mathrm{unc}}$.

4.5 Unified Training Objective

The model is trained by jointly regressing the target velocity fields of all structural components. At time $t \in [0, 1]$, the intermediate state is $\mathcal{D}_t = (\boldsymbol{A}_t, \boldsymbol{F}_t, \boldsymbol{l}_t)$, and the network $v_\theta(\mathcal{D}_t, t \mid \boldsymbol{c})$ outputs three velocity heads for occupancy, coordinates, and lattice variables, conditioned on the stoichiometry $\boldsymbol{c}$.

To account for the different scales of lattice lengths and angles, the lattice head is further split into length and angle components, denoted by v_θ^{len} and v_θ^{ang}. The overall objective is a weighted sum of mean-squared errors:

$$\begin{aligned}\mathcal{L} = \lambda_{\boldsymbol{A}'}\,\mathbb{E}_{t,\,\mathcal{D}_1,\,\mathcal{D}_t}\left[\left\|v_\theta^{\boldsymbol{A}'}(\mathcal{D}_t,t\mid \boldsymbol{c}) - u_t^{\boldsymbol{A}'}\right\|_2^2\right] + \lambda_{\boldsymbol{F}}\,\mathbb{E}_{t,\,\mathcal{D}_1,\,\mathcal{D}_t}\left[\left\|v_\theta^{\boldsymbol{F}}(\mathcal{D}_t,t\mid \boldsymbol{c}) - u_t^{\boldsymbol{F}}\right\|_2^2\right] \\ + \mathbb{E}_{t,\,\mathcal{D}_1,\,\mathcal{D}_t}\left[\lambda_{\text{len}}\left\|v_\theta^{\text{len}}(\mathcal{D}_t,t\mid \boldsymbol{c}) - u_t^{\text{len}}\right\|_2^2 + \lambda_{\text{ang}}\left\|v_\theta^{\text{ang}}(\mathcal{D}_t,t\mid \boldsymbol{c}) - u_t^{\text{ang}}\right\|_2^2\right],\end{aligned} \tag{8}$$

where $\lambda_{\boldsymbol{A}'}, \lambda_{\boldsymbol{F}}, \lambda_{\text{len}}, \lambda_{\text{ang}}$ balance occupancy, coordinate, and lattice learning. Optimizing this objective encourages a physically consistent *joint* distribution over these structural components under the stoichiometric constraint. The training and inference process is illustrated in Fig. 2c.

4.6 Stoichiometry-Conditioned Joint Velocity Network

To predict the coupled velocity fields for occupancy, coordinates, and lattice parameters, we use a lightweight stoichiometry-conditioned message-passing network. At time t, the input consists of the centered occupancy state $\boldsymbol{A}'_t$, fractional coordinates $\boldsymbol{F}_t$, lattice state $\boldsymbol{l}_t$, and stoichiometry vector $\boldsymbol{c}$. We provide physical intuition and conceptual inspirations in Appendix B. Specifically, we draw a heuristic analogy between statistical thermodynamics and our mathematical formulations, such as the logarithmic mapping, double-centering, and Sinkhorn recovery.

For each site i, we construct an initial node embedding by combining a local occupancy embedding, a global formula embedding, a global lattice embedding, and a time embedding:

$$\boldsymbol{h}_i^{(0)} = \phi_{\text{init}}\left(\phi_{\text{energy_diff}}(\boldsymbol{A}'_{t,i}),\, \phi_{\text{form}}(\boldsymbol{c}),\, \phi_{\text{lat}}(\boldsymbol{l}_t),\, \phi_{\text{time}}(t)\right), \tag{9}$$

where ϕ_* are lightweight MLPs. Within each crystal, we build a fully connected graph and perform several rounds of message passing. For an edge (i,j), the message depends on the current node states, the coordinate displacement $\Delta\boldsymbol{f}_{ij}$, the occupancy difference $\boldsymbol{A}'_{t,i} - \boldsymbol{A}'_{t,j}$, and the global formula/lattice context. Denoting the hidden state at layer ℓ by $\boldsymbol{h}_i^{(\ell)}$, the updates are

$$\boldsymbol{m}_{ij}^{(\ell)} = \phi_e\left(\boldsymbol{h}_i^{(\ell)}, \boldsymbol{h}_j^{(\ell)}, \text{SinEmb}(\Delta\boldsymbol{f}_{ij}), \phi_{\text{energy_diff}}(\boldsymbol{A}'_{t,i} - \boldsymbol{A}'_{t,j}), \boldsymbol{c}, \boldsymbol{l}_t\right), \tag{10}$$

$$\boldsymbol{h}_i^{(\ell+1)} = \boldsymbol{h}_i^{(\ell)} + \phi_v\left(\boldsymbol{h}_i^{(\ell)}, \text{mean}_j\ \boldsymbol{m}_{ij}^{(\ell)}\right), \tag{11}$$

where ϕ_e, $\phi_{\text{energy_diff}}$, and ϕ_v are MLPs, and SinEmb is the sinusoidal embedding used to encode periodic displacements [3]. The residual update preserves local information while enabling each site to aggregate composition- and geometry-aware messages from the full crystal.

After the final message-passing layer, the backbone outputs three heads. Two site-wise heads predict the coordinate velocity $v_\theta^{\boldsymbol{F}}$ and occupancy velocity $v_\theta^{\boldsymbol{A}'}$, with the latter projected by $\mathfrak{D}$ back into $\mathcal{M}_0(N,M)$. A graph-level pooled representation is then used to predict the lattice velocity $v_\theta^{\boldsymbol{l}}$. This network satisfies SO(3) invariance and translation invariance, as shown in Appendices A.3 and A.4.

5 Experiments

5.1 Experimental Setup

Currently, there are no publicly available benchmarks or models specifically designed for disordered occupancy distribution tasks. Therefore, we establish a comprehensive evaluation framework, encompassing self-constructed datasets, modified strong baselines, and a multi-tiered metric suite.

Datasets. Our benchmark is curated from two complementary databases, COD [7] and MPDS [8]. The benchmark contains 17,239 structures from COD and 110,966 structures from MPDS. We organize four subsets: Sub-20 and Sub-50 for **substitutional disorder**, and Disorder-20 and Disorder-50 for the more general setting that additionally includes **vacancies**. Here, 20 and 50 indicate that the crystal contains at most 20 or 50 atomic sites, respectively. This split allows us to separately evaluate the disorder type and scaling with the number of sites. For each subset, we use an approximate 8:1:1 train/validation/test split. Detailed statistics and processing details are deferred to Appendix C.1.

Baselines. Existing crystal generators are primarily designed for *ordered* structures and therefore cannot be directly applied to our more challenging disordered setting. To construct strong and fair

Table 1: Disordered CSP Performance on the COD dataset. When match rate is zero (i.e., when no predictions are successfully matched), RMSD and OccMAE are undefined (N/A). The best results are marked in bold. "Ada" denotes the adapted baselines.

Dataset	Model	Match Rate (↑)	RMSD (↓)	OccMAE (↓)	Validity (↑)
Sub-20	Ada-FlowMM	0.65%	**0.0146**	**0.0002**	87.40%
	Ada-OMatG	10.21%	0.0298	0.0147	90.40%
	Ada-DMFlow	21.61%	0.0580	0.0189	90.60%
	EP-Flow	**45.39%**	0.0393	0.0111	**95.10%**
Sub-50	Ada-FlowMM	0.67%	0.0728	**0.0092**	86.90%
	Ada-OMatG	3.28%	0.0459	0.0218	90.90%
	Ada-DMFlow	10.84%	**0.0229**	0.0503	80.00%
	EP-Flow	**26.77%**	0.0357	0.0107	**92.60%**
Disorder-20	Ada-FlowMM	0.49%	0.0365	0.0407	83.80%
	Ada-OMatG	9.78%	**0.0193**	**0.0154**	88.40%
	Ada-DMFlow	15.16%	0.0279	0.0298	87.20%
	EP-Flow	**31.80%**	0.036	0.0163	**89.20%**
Disorder-50	Ada-FlowMM	0.64%	**0.0146**	**0.0079**	75.70%
	Ada-OMatG	3.59%	0.0472	0.0154	**79.70%**
	Ada-DMFlow	9.74%	0.0348	0.0517	74.10%
	EP-Flow	**17.10%**	0.0334	0.0148	77.70%

baselines, we adapt representative SOTA methods, including FlowMM [4] and OMatG [33], to operate on continuous site-wise occupancies rather than discrete one-hot species representations. On the generation side, these baselines first predict an unconstrained ODM in Euclidean space, which is then mapped to the positive orthant through a sigmoid transformation, enabling the subsequent application of the Sinkhorn projection. Furthermore, we adapt DMFlow [20], a strong baseline specifically designed for disordered crystal generation. We adopt the spherical reparameterization mechanism from its DNG, enabling the occupancy to diffuse on the sphere, which naturally satisfies the site-wise row-sum constraint. For all baselines, we apply the same formula-derived mask based on the ground-truth chemical composition as used in EP-Flow during both training and generation. In the inference phase, we uniformly apply the same Sinkhorn projection to project the occupancies onto the transportation polytope space, enforcing both the global formula sum and site-wise sum constraints. Further implementation details of all baseline adaptations are provided in Appendix C.2.

Evaluation Metrics. Our models are evaluated by four metrics: **(1)** Match Rate, which measures structure-level matching under the extended matcher; **(2)** RMSD, which quantifies coordinate accuracy for matched structures; **(3)** OccMAE, which measures the error in site-wise occupancy distributions; and **(4)** Validity, which measures the fraction of physically plausible outputs. The exact metric definitions, matcher tolerances, and implementation details are provided in Appendix C.3.

5.2 Main Results on Disordered Crystal Generation

We summarize the quantitative performance in Tables 1 and 2. Three observations stand out: (1) **EP-Flow achieves the highest Match Rate across all evaluated subsets.** Across all eight subsets, EP-Flow achieves Match Rates ranging from 17.10% to 60.82%, while Ada-FlowMM remains below 1%, Ada-OMatG achieves Match Rates of 3.28%–11.78%, and Ada-DMFlow achieves Match Rates of 9.74%–45.17%. EP-Flow outperforms the strongest baseline, Ada-DMFlow, by 7.36–23.78 percentage points, suggesting the benefit of modeling dynamics on a formula-dependent transportation polytope rather than in a generic Euclidean space. This comparison should also be read together with RMSD and OccMAE: since both are averaged only over matched cases, a model with higher Match Rate can include harder successful matches and therefore exhibit slightly larger average RMSD, so these metrics should be interpreted jointly rather than in isolation. (2) **High validity does not imply correct structure recovery.** The adapted baselines can obtain high validity despite low Match Rate. For instance, on MPDS Disorder-20, Ada-FlowMM achieves 95.90% validity but only 0.24% Match Rate. This indicates that generating physically plausible structures alone does not ensure

Table 2: Disordered CSP Performance on the MPDS dataset. When match rate is zero (i.e., when no predictions are successfully matched), RMSD and OccMAE are undefined (N/A). The best results are marked in bold. "Ada" denotes the adapted baselines.

Dataset	Model	Match Rate (↑)	RMSD (↓)	OccMAE (↓)	Validity (↑)
Sub-20	Ada-FlowMM	0.49%	0.0335	0.0457	96.90%
	Ada-OMatG	11.78%	**0.0331**	0.0813	88.30%
	Ada-DMFlow	45.17%	0.0416	0.0096	92.10%
	EP-Flow	**60.82%**	0.0483	**0.0075**	**97.80%**
Sub-50	Ada-FlowMM	0.18%	0.0501	0.0391	**98.10%**
	Ada-OMatG	5.25%	**0.0242**	0.0594	93.50%
	Ada-DMFlow	25.00%	0.0354	0.0320	81.50%
	EP-Flow	**39.33%**	0.0538	**0.0076**	91.90%
Disorder-20	Ada-FlowMM	0.24%	**0.0136**	0.0474	**95.90%**
	Ada-OMatG	9.91%	0.0263	0.0860	92.40%
	Ada-DMFlow	42.59%	0.0329	0.0211	94.10%
	EP-Flow	**57.74%**	0.0413	**0.0092**	94.90%
Disorder-50	Ada-FlowMM	0.13%	0.0354	0.0401	**97.30%**
	Ada-OMatG	3.76%	**0.0178**	0.0460	87.10%
	Ada-DMFlow	23.84%	0.0485	0.0250	84.10%
	EP-Flow	**32.60%**	0.0550	**0.0069**	89.60%

Table 3: Ablation results on MPDS-Sub-20.

Variant	Match Rate (↑)	RMSD (↓)	OccMAE (↓)	Validity (↑)
w/o Formula	20.01%	0.0529	0.0106	96.7%
w/o Energy-diff	35.23%	0.0527	0.0159	93.50%
w/o Double Centering	46.06%	0.0537	0.0115	91.0%
w/o Projection	15.78%	**0.0434**	0.1277	90.5%
w/o DC & Sinkhorn	0.00%	N/A	N/A	0.00%
Full Model (Reference)	**60.82%**	0.0483	**0.0075**	**97.8%**

correct recovery of coupled site-wise occupancy and geometry. This further highlights the necessity of employing stringent structure-level matching metrics to evaluate the true predictive capacity of models in disordered CSP. (3) **Larger and more disordered systems are intrinsically harder.** From 20 to 50 sites, and from substitutional to vacancy-inclusive disorder subsets, both EP-Flow and the strong baseline Ada-DMFlow degrade consistently on both COD and MPDS, reflecting the greater combinatorial ambiguity of site assignment and the higher sparsity of occupancy patterns. Meanwhile, they perform consistently better on MPDS than on COD, suggesting that larger disordered datasets provide richer supervision for learning transferable site–species compatibility.

5.3 Ablation Studies

To isolate the contribution of each core component, we conduct five ablations on MPDS-Sub-20 in Table 3: (1) **w/o Formula**, which removes chemical-formula conditioning embedded in the network layers and sharply reduces Match Rate from 60.82% to 20.01%, showing that accurate occupancy prediction requires explicit stoichiometric guidance beyond local geometry; (2) **w/o Energy-diff**, which removes the relative energy-difference edge feature and lowers Match Rate to 35.23% while increasing OccMAE to 0.0159, indicating that pairwise relative affinities provide important local signals for resolving fine-grained disorder (see Appendix B.4 for a detailed physical interpretation); (3) **w/o Double Centering**, which removes the $\mathfrak{D}$ operator and learns in the log space, reducing Match Rate to 46.06% and confirming that mapping formula-dependent polytopes into a shared centered space improves structural alignment; (4) **w/o Projection**, which learns directly in the raw bounded polytope and causes OccMAE to surge from 0.0075 to 0.1277, showing that the unflattened constraint geometry is too irregular for occupancy learning, even if the matched-sample RMSD appears competitive; and (5) **w/o DC & Sinkhorn**, which removes both normalization and

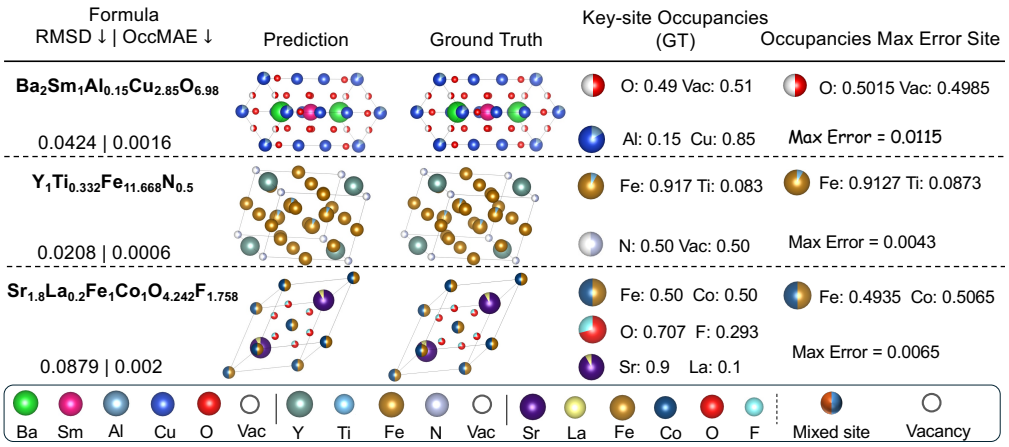


Figure 3: **Reconstruction results for disordered functional materials.** Each row compares predicted and ground-truth structures, reporting RMSD and OccMAE. Occupancies are annotated, and the rightmost column highlights the site with the largest occupancy error. Three disorder patterns are shown: oxygen vacancies, interstitial N/vacancies, and mixed-cation/mixed-anion substitution.

constraint projection and collapses completely to 0.00% Match Rate, demonstrating that geometric normalization and marginal preservation are critically important for staying on the feasible manifold.

5.4 More Analysis

5.4.1 Case Studies on Functional Disordered Materials

To assess whether EP-Flow captures realistic disorder beyond simple cation mixing, we examine three representative functional materials spanning substitutional, vacancy, and interstitial disorder. As shown in Fig. 3, these examples probe whether the model recovers not only average occupancies but also property-relevant local configurations.

Layered cuprate with oxygen-vacancy disorder. $Ba_2SmAl_{0.15}Cu_{2.85}O_{6.98}$ is an Al-substituted $SmBa_2Cu_3O_{7-\delta}$-type cuprate with dilute Cu/Al substitution and oxygen-vacancy disorder. In these high-T_c materials, the functional impact of vacancies depends critically on whether they appear in the charge-regulating Cu–O chains or in the superconducting CuO_2 planes. EP-Flow localizes the vacancies outside the CuO_2 planes rather than distributing them randomly, indicating that it captures chemically meaningful defect preferences rather than only average geometry.

$ThMn_{12}$-type magnet with interstitial disorder. $YTi_{0.332}Fe_{11.668}N_{0.5}$ is a $ThMn_{12}$-type Fe-rich compound with dilute Fe/Ti substitution and partially occupied interstitial nitrogen sites. While the Ti occupancy is only about 8.3% and the interstitial N site has an occupancy of 0.5, both disorder channels are directly tied to permanent-magnet behavior through their effects on structural stability, Fe–Fe distances, and magnetocrystalline anisotropy. This case shows that EP-Flow can capture the coupling between dilute substitutional disorder and interstitial N/vacancy disorder in a physically meaningful way.

Mixed-anion perovskite with multi-sublattice substitution. $Sr_{1.8}La_{0.2}FeCoO_{4.242}F_{1.758}$ can be approximately written as $(Sr_{0.9}La_{0.1})(Fe_{0.5}Co_{0.5})(O_{0.707}F_{0.293})_3$, corresponding to an oxyfluoride perovskite with simultaneous A-site Sr/La mixing, B-site Fe/Co mixing, and anion-site O/F mixing. This is a challenging multi-sublattice disorder case because both cation and anion sublattices are disordered, and the O/F arrangement directly affects local charge balance, covalency, and crystal-field splitting around the transition-metal centers. The example therefore tests whether EP-Flow can recover coupled mixed-cation and mixed-anion disorder within a single framework.

Taken together, these examples show that EP-Flow is not restricted to simple cation-site disorder: it recovers mixed-anion substitution in a perovskite, interstitial N/vacancy disorder in an Fe-rich magnetic framework, and oxygen-vacancy disorder in a layered superconducting oxide. Because these disorder patterns are directly linked to functional properties, successful reconstruction provides stronger evidence that the model captures realistic, property-relevant local chemistry. For reconstructions and analyses of more complex cases, please see Appendix D.5.

5.4.2 Disorder-Pattern Analysis

To probe what the model learns, Fig. 4 visualizes site-wise occupancy correlations. EP-Flow produces sparse, structured site–species assignments rather than diffuse occupancies: strong diagonal blocks indicate stable site preferences, while localized off-diagonal responses capture mixed-occupancy behavior. Moreover, some predicted mixing patterns absent from the smaller COD training split can still be verified in the larger MPDS dataset, suggesting that the model learns transferable site–species compatibility rather than merely memorizing observed occupancy pairs.

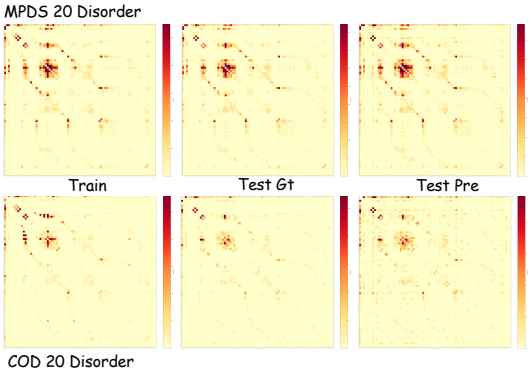


Figure 4: Site-wise occupancy-pattern analysis.

6 Conclusion

We presented EP-Flow, a geometric flow matching framework for disordered CSP conditioned on the formula and site number. A key contribution is formalizing this generation task to unify ordered structures with substitutional, vacancy, and interstitial disorder through the Occupancy Distribution Matrix (ODM). Methodologically, EP-Flow canonicalizes heterogeneous, formula- and site-dependent occupancy polytopes into a shared, unbounded double-centered subspace. It then learns a marginal-preserving flow and maps samples back via Sinkhorn projection, thereby ensuring physical validity. Together with a complete evaluation framework built from large-scale COD and MPDS data and metrics such as the Structure Match Rate tailored for disordered structures, this yields a unified pipeline for jointly generating occupancies, fractional coordinates, and lattice parameters directly from the given formula and site count. Experiments show that EP-Flow achieves state-of-the-art performance across disorder types and site-count scales, highlighting the advantage of jointly modeling occupancy, geometry, and lattice under exact compositional constraints.

Acknowledgments and Disclosure of Funding

Funding. This work was jointly supported by the National Natural Science Foundation of China (Grant Nos. 52573261 and 62376276). The AI-driven experiments, simulations, and model training were performed on the Robotic AI-Scientist Platform of the Chinese Academy of Sciences.

Acknowledgments. We gratefully acknowledge the financial and computational support provided by the Data Center for Condensed Matter Physics, Chinese Academy of Sciences.

Competing interests. The authors declare no competing interests.

References

[1] 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.

[2] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

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

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

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

[6] Alec Belsky, Mariette Hellenbrandt, Vicky Lynn Karen, and Peter Luksch. New developments in the inorganic crystal structure database (icsd): accessibility in support of materials research and design. *Structural Science*, 58(3):364–369, 2002.

[7] Saulius Gražulis, Adriana Daškevič, Andrius Merkys, Daniel Chateigner, Luca Lutterotti, Miguel Quiros, Nadezhda R Serebryanaya, Peter Moeck, Robert T Downs, and Armel Le Bail. Crystallography open database (cod): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic acids research*, 40(D1):D420–D427, 2012.

[8] Evgeny Blokhin and Pierre Villars. Materials platform for data science, 2022.

[9] Xiaoshan Luo, Zhenyu Wang, Pengyue Gao, Jian Lv, Yanchao Wang, Changfeng Chen, and Yanming Ma. Deep learning generative model for crystal structure prediction. *npj Computational Materials*, 10(1):254, 2024.

[10] Peijia Lin, Pin Chen, Rui Jiao, Qing Mo, Cen Jianhuan, Wenbing Huang, Yang Liu, Dan Huang, and Yutong Lu. Equivariant diffusion for crystal structure prediction. In *Forty-first International Conference on Machine Learning*, 2024.

[11] 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*, 2026.

[12] Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. *arXiv preprint arXiv:2503.03965*, 2025.

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

[14] Xiaoshan Luo, Zhenyu Wang, Qingchang Wang, Jian Lv, Lei Wang, Yanchao Wang, and Yanming Ma. Crystalflow: A flow-based generative model for crystalline materials. *arXiv preprint arXiv:2412.11693*, 2024.

[15] Qi Li, Rui Jiao, Liming Wu, Tiannian Zhu, Wenbing Huang, Shifeng Jin, Yang Liu, Hongming Weng, and Xiaolong Chen. Powder diffraction crystal structure determination using generative models. *Nature Communications*, 16(1):7428, 2025.

[16] Rui Jiao, Hanlin Wu, Wenbing Huang, Yuxuan Song, Yawen Ouyang, Yu Rong, Tingyang Xu, Pengju Wang, Hao Zhou, Wei-Ying Ma, et al. Mof-bfn: Metal-organic frameworks structure prediction via bayesian flow networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[17] Laura Ruple, Luca Torresi, Henrik Schopmans, and Pascal Friederich. Symmetry-aware bayesian flow networks for crystal generation. *arXiv preprint arXiv:2502.03146*, 2025.

[18] Hanlin Wu, Yuxuan Song, Jingjing Gong, Ziyao Cao, Yawen Ouyang, Jianbing Zhang, Hao Zhou, Wei-Ying Ma, and Jingjing Liu. A periodic bayesian flow for material generation. *arXiv preprint arXiv:2502.02016*, 2025.

[19] Martin Hoffmann Petersen, Ruiming Zhu, Haiwen Dai, Savyasanchi Aggarwal, Nong Wei, Andy Paul Chen, Arghya Bhowmik, Juan Maria Garcia Lastra, and Kedar Hippalgaonkar. Dis-gen: Disordered crystal structure generation. *arXiv preprint arXiv:2507.18275*, 2025.

[20] Liming Wu, Rui Jiao, Qi Li, Mingze Li, Songyou Li, Shifeng Jin, and Wenbing Huang. Dmflow: Disordered materials generation by flow matching. *arXiv preprint arXiv:2602.04734*, 2026.

[21] Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in neural information processing systems*, 35:2406–2422, 2022.

[22] Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to dna sequence design. In *International Conference on Machine Learning*, pages 46495–46513. PMLR, 2024.

[23] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385, 2024.

[24] Oscar Davis, Samuel Kessler, Mircea Petrache, İsmail İ Ceylan, Michael Bronstein, and Avishek J Bose. Fisher flow matching for generative modeling over discrete data. *Advances in Neural Information Processing Systems*, 37:139054–139084, 2024.

[25] Tomek Diederen and Nicola Zamboni. Flows on convex polytopes. *arXiv preprint arXiv:2503.10232*, 2025.

[26] Utkarsh Utkarsh, Pengfei Cai, Alan Edelman, Rafael Gomez-Bombarelli, and Christopher Vincent Rackauckas. Physics-constrained flow matching: Sampling generative models with hard constraints. *arXiv preprint arXiv:2506.04171*, 2025.

[27] Nic Fishman, Leo Klarner, Valentin De Bortoli, Emile Mathieu, and Michael Hutchinson. Diffusion models for constrained domains. *arXiv preprint arXiv:2304.05364*, 2023.

[28] Guan-Horng Liu, Tianrong Chen, Evangelos Theodorou, and Molei Tao. Mirror diffusion models for constrained and watermarked generation. *Advances in Neural Information Processing Systems*, 36:42898–42917, 2023.

[29] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

[30] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[31] Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

[32] Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

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

[34] Michael Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR 2023 Conference*, 2023.

[35] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[36] Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024.

[37] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[38] 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.

A Proofs and Foundational Concepts

A.1 Proof of the Bijectivity of Canonicalization

We restate the main claim using the notation of the paper. For a fixed stoichiometry vector $\boldsymbol{c} \in \mathbb{R}^M_{>0}$ with total mass $N = \mathbf{1}^\top \boldsymbol{c}$, define the transportation polytope

$$\mathcal{P}(\mathbf{1}, \boldsymbol{c}) = \left\{ \boldsymbol{P} \in \mathbb{R}^{N \times M}_{>0} \;\middle|\; \boldsymbol{P}\mathbf{1}_M = \mathbf{1}_N,\ \boldsymbol{P}^\top \mathbf{1}_N = \boldsymbol{c} \right\}, \tag{12}$$

and the centered space

$$\mathcal{M}_0(N, M) \triangleq \left\{ \boldsymbol{Z} \in \mathbb{R}^{N \times M} \;\middle|\; \boldsymbol{Z}\mathbf{1}_M = \mathbf{0}_N,\ \mathbf{1}_N^\top \boldsymbol{Z} = \mathbf{0}_M^\top \right\}. \tag{13}$$

Proof. We first prove surjectivity. Fix any $\boldsymbol{Z} \in \mathcal{M}_0(N, M)$ and define $\boldsymbol{B} = \exp(\boldsymbol{Z})$ elementwise. Since $\boldsymbol{B}$ is strictly positive, the Sinkhorn–Knopp theorem guarantees the existence of positive diagonal matrices $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ such that

$$\boldsymbol{A} = \boldsymbol{D}_1 \boldsymbol{B} \boldsymbol{D}_2 \in \mathcal{P}(\mathbf{1}, \boldsymbol{c}). \tag{14}$$

We can explicitly express these strictly positive diagonal matrices using real-valued vectors $\boldsymbol{a} \in \mathbb{R}^N$ and $\boldsymbol{b} \in \mathbb{R}^M$ as $\boldsymbol{D}_1 = \mathrm{diag}(\exp(\boldsymbol{a}))$ and $\boldsymbol{D}_2 = \mathrm{diag}(\exp(\boldsymbol{b}))$.

Substituting these into the scaling equation and taking the element-wise logarithm seamlessly translates the multiplicative Sinkhorn scaling into additive row and column shifts:

$$\log \boldsymbol{A} = \boldsymbol{Z} + \boldsymbol{a}\mathbf{1}^\top + \mathbf{1}\boldsymbol{b}^\top. \tag{15}$$

Applying the double-centering operator gives

$$\Psi(\boldsymbol{A}) \triangleq \mathfrak{D}(\log \boldsymbol{A}) = \mathfrak{D}(\boldsymbol{Z}) = \boldsymbol{Z}, \tag{16}$$

because $\mathfrak{D}$ removes all additive row-wise and column-wise components and $\boldsymbol{Z}$ is already centered. Hence every $\boldsymbol{Z} \in \mathcal{M}_0(N, M)$ has a preimage.

To prove injectivity, suppose $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathcal{P}(\mathbf{1}, \boldsymbol{c})$ satisfy $\Psi(\boldsymbol{A}_1) = \Psi(\boldsymbol{A}_2) = \boldsymbol{Z}$. Then both $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are strictly positive matrices whose logarithms differ from $\boldsymbol{Z}$ only by row-wise and column-wise additive terms. Exponentiating shows that both matrices are diagonal scalings of $\exp(\boldsymbol{Z})$ with the same target marginals $(\mathbf{1}, \boldsymbol{c})$. By uniqueness of Sinkhorn scaling for strictly positive matrices, these two scalings must coincide, so $\boldsymbol{A}_1 = \boldsymbol{A}_2$.

Therefore Ψ is bijective, and its inverse is exactly the Sinkhorn recovery

$$\Psi^{-1}(\boldsymbol{Z}) = \mathcal{S}(\exp(\boldsymbol{Z}), \mathbf{1}, \boldsymbol{c}). \tag{17}$$

□

A.2 Trajectory Feasibility in the Centered Space

We next justify why projecting the occupancy velocity onto $\mathcal{M}_0(N, M)$ preserves the centered constraints along the entire ODE trajectory.

Proof. Let $\boldsymbol{A}'_t \in \mathbb{R}^{N \times M}$ satisfy the ODE

$$\frac{d}{dt}\boldsymbol{A}'_t = v_\theta^{\boldsymbol{A}'}(\mathcal{D}_t, t \mid \boldsymbol{c}), \qquad \boldsymbol{A}'_0 \in \mathcal{M}_0(N, M), \tag{18}$$

and assume that $v_\theta^{\boldsymbol{A}'} \in \mathcal{M}_0(N, M)$ for all t. Consider the row-sum vector $\boldsymbol{r}'_t = \boldsymbol{A}'_t\mathbf{1}$ and the column-sum vector $\boldsymbol{c}'_t = (\boldsymbol{A}'_t)^\top \mathbf{1}$. Differentiating yields

$$\frac{d}{dt}\boldsymbol{r}'_t = \frac{d}{dt}\boldsymbol{A}'_t\mathbf{1} = v_\theta^{\boldsymbol{A}'}(\mathcal{D}_t, t \mid \boldsymbol{c})\mathbf{1} = \mathbf{0}, \tag{19}$$

and similarly

$$\frac{d}{dt}\boldsymbol{c}'_t = \left(v_\theta^{\boldsymbol{A}'}(\mathcal{D}_t, t \mid \boldsymbol{c})\right)^\top \mathbf{1} = \mathbf{0}. \tag{20}$$

Since $\boldsymbol{A}'_0 \in \mathcal{M}_0(N, M)$, we have $\boldsymbol{r}'_0 = \mathbf{0}$ and $\boldsymbol{c}'_0 = \mathbf{0}$. Therefore $\boldsymbol{r}'_t = \mathbf{0}$ and $\boldsymbol{c}'_t = \mathbf{0}$ for every $t \in [0, 1]$, which implies $\boldsymbol{A}'_t \in \mathcal{M}_0(N, M)$ for all t. □

Combining this fact with the inverse map Ψ^{-1} immediately shows that the recovered trajectory $\boldsymbol{A}_t = \mathcal{S}(\exp(\boldsymbol{A}'_t), \mathbf{1}, \boldsymbol{c})$ stays inside $\mathcal{P}(\mathbf{1}, \boldsymbol{c})$ at all times.

A.3 $SO(3)$ Invariance of the Joint Velocity Field

For our model, a global rigid rotation acts only on the orientation of the lattice basis in Cartesian space. However, the network does not use the oriented basis $\boldsymbol{L}_t$ directly. Instead, it takes the lattice parameters $\boldsymbol{l}_t = (a, b, c, \alpha, \beta, \gamma)$ as input, together with $\boldsymbol{A}'_t$, $\boldsymbol{F}_t$, and $\boldsymbol{c}$. These six lattice parameters are invariant under any global rotation $Q \in SO(3)$, because such a rotation changes the embedding of the unit cell in Euclidean space but does not change its lengths or inter-axis angles. Therefore,

$$\boldsymbol{l}_t(Q\boldsymbol{L}_t) = \boldsymbol{l}_t(\boldsymbol{L}_t). \tag{21}$$

At the same time, the fractional coordinates $\boldsymbol{F}_t$ are intrinsic coordinates with respect to the lattice basis, so they are unchanged by a simultaneous rotation of the whole crystal. Hence the full model input is rotation-invariant. Since the network is a deterministic function of these rotation-invariant inputs, its outputs are also invariant.

Therefore, a global $SO(3)$ rotation does not affect the predicted occupancy, coordinate, or lattice-parameter velocities. In this sense, the model is rotation-invariant at the level of its input representation and output predictions.

A.4 Translation Invariance of the Joint Velocity Field

We next consider periodic translations on the torus. It suffices to focus on the fractional coordinates $\boldsymbol{F}_t$, since the other inputs $\boldsymbol{A}'_t$, $\boldsymbol{l}_t$, and $\boldsymbol{c}$ are unaffected by a global shift of the crystal origin. Let $w(\cdot)$ denote the elementwise wrapping operator on the torus, and consider a periodic translation by $\boldsymbol{\tau} \in \mathbb{R}^3$:

$$\boldsymbol{F}_t \mapsto w(\boldsymbol{F}_t + \mathbf{1}\boldsymbol{\tau}^\top). \tag{22}$$

Because the network only uses relative fractional displacements, for any edge (i, j) we have

$$w(w(\boldsymbol{f}_j + \boldsymbol{\tau}) - w(\boldsymbol{f}_i + \boldsymbol{\tau})) = w(\boldsymbol{f}_j - \boldsymbol{f}_i). \tag{23}$$

Hence the periodic offset is unchanged under a common translation of all sites,

$$\Delta\boldsymbol{f}_{ij}\big(w(\boldsymbol{F}_t + \mathbf{1}\boldsymbol{\tau}^\top)\big) = \Delta\boldsymbol{f}_{ij}(\boldsymbol{F}_t), \tag{24}$$

and therefore its sinusoidal embedding is also unchanged. Since the remaining inputs are unchanged as well, the entire network input representation is translation-invariant.

A.5 Flow Matching

Flow Matching (FM) [34, 35] defines a continuous-time transformation between a base distribution p_0 and a target data distribution p_1. Let $\boldsymbol{x}_t \in \mathbb{R}^d$ denote a time-dependent data point along the probability path $p_t(\boldsymbol{x})$, where $t \in [0, 1]$. This path is governed by an Ordinary Differential Equation (ODE): $\frac{d}{dt}\boldsymbol{x}_t = v_t(\boldsymbol{x}_t)$, where v_t is a time-dependent vector field. Since the true score or vector field of the marginal distribution p_t is typically intractable, Conditional FM [35] introduces a conditional probability path $p_t(\boldsymbol{x}|\boldsymbol{x}_1)$ that interpolates between a noise sample $\boldsymbol{x}_0 \sim p_0$ and a data sample $\boldsymbol{x}_1 \sim p_1$. Under a linear interpolation scheme, the intermediate state is defined as $\boldsymbol{x}_t = t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0$. The model is trained to approximate the ideal vector field by minimizing the expected L_2 distance between the learned vector field $v_\theta(\boldsymbol{x}_t, t)$ and the tractable conditional vector field $v_t(\boldsymbol{x}_t|\boldsymbol{x}_1)$. Once trained, new samples are generated by numerically integrating the ODE from $t = 0$ to $t = 1$.

Notably, when data are naturally constrained to specific geometric structures, such as the torus or the simplex, standard Euclidean FM is extended to Riemannian FM [36]. This framework defines the probability path and vector field directly on the manifold's tangent bundle, ensuring that the generative trajectory respects the underlying geometric boundaries.

B Physical Intuition and Conceptual Inspirations

The algebraic operations of the canonicalization map Ψ and its inverse can be understood as a **conceptual analogy** to the principles of statistical thermodynamics governing continuous atomic distributions. We detail the physical mappings of the four main modeling components below.

B.1 Logarithmic Mapping as Energy Space Evolution

In our continuous formulation, the fractional occupancy $A_{ij} \in (0, 1)$ represents the probability of species j occupying site i. This suggests a **Boltzmann-type interpretation**, where the occupancy is conceptually related to a latent local formation energy E_{ij} via:

$$A_{ij} = \frac{1}{Q} \exp\left(-\frac{E_{ij}}{k_B T}\right), \tag{25}$$

where Q is the partition function and T serves as an effective temperature. Taking the element-wise logarithm yields:

$$\log A_{ij} = -\frac{E_{ij}}{k_B T} - \log Q. \tag{26}$$

By mapping the bounded probability simplex into an unconstrained energy space, the generative ODE avoids learning a velocity field over restricted probability boundaries. Consequently, the vector field learned by the model can be viewed as an analogy to the rate of change of an energy landscape with respect to generative time ($\propto -\partial E_{ij}/\partial t$). This enables the network to simulate a continuous thermodynamic-like evolution, driving the system from high-entropy noise toward a low-energy configuration. Note that this formulation characterizes the generative trajectory as an evolution toward potential stability, rather than assuming the system maintains global thermodynamic equilibrium at every intermediate step.

B.2 Double-Centering as Relative Potential Extraction

In thermodynamics, absolute energy scales are unobservable; system dynamics are driven exclusively by relative energy differences. We decompose the effective local formation energy E_{ij} into a centered core interaction and macroscopic background fields:

$$E_{ij} = E_{ij}^{\text{core}} + \mu_i^{\text{site}} + \mu_j^{\text{element}} + E_0. \tag{27}$$

Here, μ_i^{site} is the site-specific potential well, μ_j^{element} is the species-specific chemical potential determined by stoichiometry $\boldsymbol{c}$, and E_{ij}^{core} is the intrinsic relative energy. Applying the double-centering operator $\mathfrak{D}$ to $\log \boldsymbol{A}$ **algebraically eliminates** these additive background fields and constant terms, leaving:

$$Z_{ij} = \mathfrak{D}(\log \boldsymbol{A})_{ij} = -\frac{E_{ij}^{\text{core}}}{k_B T}. \tag{28}$$

Therefore, $\boldsymbol{Z}$ represents a dimensionless relative affinity, acting as an **intrinsic structural invariant** of the local chemical bonding. Operating in $\mathcal{M}_0(N, M)$ therefore preserves the information needed to recover the ODM under fixed marginals, while providing an energy-like representation for learning.

B.3 Sinkhorn Recovery as Thermodynamic Annealing

The inverse mapping reconstructs the physical occupancy $\boldsymbol{A}$ by applying the Sinkhorn algorithm to $\exp(\boldsymbol{Z})$, converging to:

$$A_{ij} = \exp(\boldsymbol{a}_i) \exp\left(-\frac{E_{ij}^{\text{core}}}{k_B T}\right) \exp(\boldsymbol{b}_j). \tag{29}$$

In the context of a grand canonical ensemble, the scaling variables $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$ function as dynamic chemical potentials. The iterative Sinkhorn scaling provides a mathematical **analogy** to an isothermal annealing process where atomic distributions are adjusted to satisfy global stoichiometric constraints. Structural formation is finalized when the system reaches a state that **effectively satisfies** the physical boundary constraints imposed by the target formula.

B.4 Physical Interpretation of Energy-Difference Edge Features

In the message-passing network, we include the difference between connected sites in the double-centered log-domain representation:

$$\Delta \boldsymbol{A}'_{ij} = \boldsymbol{A}'_{i,:} - \boldsymbol{A}'_{j,:}. \tag{30}$$

This feature is computed from the centered representation $\boldsymbol{A}' = \mathfrak{D}(\log \boldsymbol{A})$, which removes background terms and can be interpreted as a relative affinity field. This log-domain difference has a direct probabilistic interpretation: for a given species e, a difference in log occupancy corresponds to a probability ratio $\log(A_{i,e}/A_{j,e})$. Under the Boltzmann-type interpretation, this ratio further corresponds to a relative energy difference $-(E_{i,e} - E_{j,e})/k_B T$. Subtraction in the log-domain space thus encodes relative site preferences. By providing the network with explicit information about how the relative energy landscape changes across the crystal graph, $\Delta \boldsymbol{A}'_{ij}$ helps the model distinguish locally compatible site–species configurations. The ablation results in Table 3 confirm that this feature is critical for recovering accurate local disorder patterns.

Concluding Remarks on Physical Analogy. This conceptual bridge between manifold-constrained flow and statistical mechanics represents an **exploratory attempt** to provide physical intuition for disordered crystal generation. While these analogies are intended as a heuristic framework rather than a formal derivation, they offer a new vantage point for designing constraint-aware generative models in materials science. We hope this perspective stimulates further cross-disciplinary exploration between the machine learning and physical science communities.

C Data Preprocessing and Metrics

C.1 Datasets

We evaluate the model on four subsets derived from the COD and MPDS databases: substitutional-disorder and general-disorder settings, each at two scales (up to 20 and 50 sites). Following the protocol in the main text, each subset is split into training, validation, and test sets with an approximate ratio of 8:1:1. All models are trained conditionally on the stoichiometry vector $\boldsymbol{c}$. Crucially, to provide a unified thermodynamic representation, we treat crystallographic vacancies as an explicit pseudo-atom. Consequently, the stoichiometry vector $\boldsymbol{c}$ includes this vacancy species, ensuring that the total sum of the formula strictly equals the number of crystallographic sites ($\mathbf{1}^\top \boldsymbol{c} = N$).

All raw CIF files are standardized by primitive-cell reduction followed by Niggli reduction.

We then apply a set of filters to ensure physical validity, computational tractability, and sufficient statistical support. Structures with interatomic distances below 0.5 Å, cell volume below 0.1 Å^3, or any lattice parameter larger than 40 Å are removed.

To evaluate different structural scales, we construct subsets with the number of crystallographic sites satisfying $N \leq 20$ or $N \leq 50$.

To control the compositional complexity, we limit the number of distinct species in each extended formula to $M \leq 6$, including the vacancy pseudo-atom when present.

To focus on inorganic crystals and avoid molecular orientational disorder, we exclude structures containing both C and H.

To restrict local disorder complexity, we allow at most four co-occupying species on each mixed site and require each retained structure to contain at least one disordered site.

To ensure statistically reliable elemental embeddings, we remove structures containing rare elements, defined as elements appearing fewer than 5 times in raw COD or fewer than 50 times in raw MPDS.

Figure 5 illustrates the element-to-position index map used in this work and the resulting element-occurrence distribution in the MPDS Disorder-50 subset.

To prevent data leakage during evaluation, we apply a strict structural deduplication protocol before the train/validation/test split. Structures with matching formulas and site counts are compared using structure matchers with

$$(\text{ltol}, \text{stol}, \text{angle_tol}, \text{occtol}) = (0.2, 0.3, 5, 0.1). \tag{31}$$

Table 4: Dataset statistics for COD and MPDS databases.

Database	Setting	Train	Val	Test	Total	Ratio (Train:Val:Test)
COD	Substitution 20	7,100	879	921	8,900	7.98 : 0.99 : 1.03
	Substitution 50	9,483	1,141	1,190	11,814	8.03 : 0.97 : 1.01
	Disorder 20	9,408	1,185	1,217	11,810	7.97 : 1.00 : 1.03
	Disorder 50	13,825	1,689	1,725	17,239	8.02 : 0.98 : 1.00
MPDS	Substitution 20	53,492	6,648	6,786	66,926	7.99 : 0.99 : 1.01
	Substitution 50	65,187	8,412	8,499	82,098	7.94 : 1.02 : 1.03
	Disorder 20	69,032	8,715	8,835	86,582	7.97 : 1.01 : 1.02
	Disorder 50	87,882	11,531	11,553	110,966	7.92 : 1.04 : 1.04

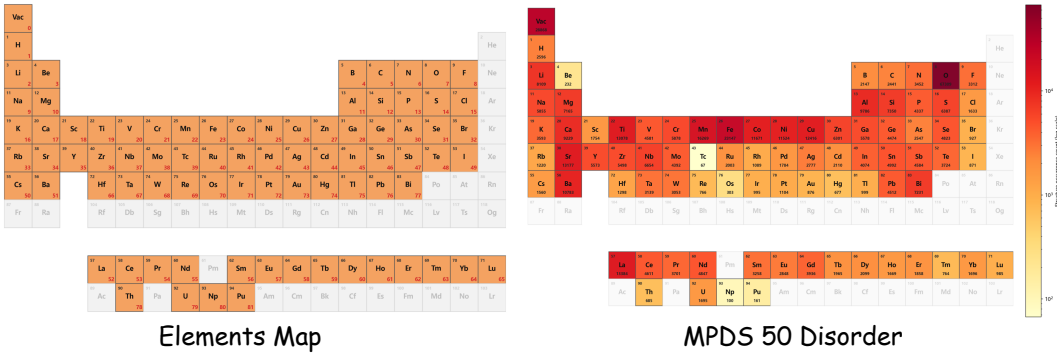


Figure 5: **Element index map and MPDS Disorder-50 distribution.** Left: the element-to-position index map used in this work. Right: the element-occurrence heatmap for the MPDS Disorder-50 subset.

Duplicates are rigorously removed from the validation and test sets.

Table 4 reports the size of each de-duplicated subset. The Sub-20/50 subsets contain substitutional disorder only, whereas Disorder-20/50 additionally include vacancy disorder and therefore represent a more challenging generation regime.

C.2 Baseline Adaptations

Ordered Baselines. Since FlowMM [4] and OMatG [33] are originally designed for ordered crystal structure prediction, we adapt them to support continuous site-wise occupancies while preserving their original backbone architectures. Specifically, we replace the discrete atomic-species representation with continuous occupancy distributions and introduce an additional prediction head to jointly generate site occupancies with fractional coordinates and lattice parameters. The input chemical formula is used to mask species absent from the target composition, and a Sinkhorn projection is applied to the predicted occupancy matrix at inference time to enforce the target composition and site-wise normalization constraints. For OMatG, we additionally incorporate the disordered formula embedding used in EP-Flow, which provides explicit composition and site information to the network. Other components of the original FlowMM and OMatG architectures are kept unchanged.

Disordered Baselines. The original CSP setting of DMFlow [20] assumes that site-wise occupancy distributions are provided as input and remain fixed during generation, which is incompatible with our formula-conditioned setting where the occupancies are unknown. We therefore adapt the occupancy-generation mechanism from DMFlow's De Novo Generation (DNG) setting to predict site-wise occupancies from the chemical formula. In particular, we employ its Riemannian spherical reparameterization to constrain the generated occupancy distributions, while jointly generating the structural variables using the original DMFlow framework. At inference time, we further apply a Sinkhorn projection to the predicted occupancies to enforce consistency with the target chemical formula. Thus, Adapted-DMFlow does not use ground-truth site-level occupancy information as input and receives the same task-level conditioning information as EP-Flow.

Summary of constraint handling. These adaptations also provide a progressive comparison of how occupancy constraints are incorporated into the generative process. Adapted-FlowMM and Adapted-OMatG generate occupancies without explicitly enforcing either the row or column constraints during generation, and rely on a post-hoc Sinkhorn projection to satisfy both constraints. Adapted-DMFlow introduces the site-wise row constraint through its spherical reparameterization, such that each site represents a normalized occupancy distribution, but does not enforce the column constraint corresponding to the target composition during generation; the latter is again imposed through the Sinkhorn post-processing. In contrast, EP-Flow explicitly preserves both the row and column constraints throughout the generative process. Therefore, these baselines form a progressive comparison from unconstrained generation, to row-constrained generation, and finally to the jointly row- and column-constrained generation of EP-Flow, allowing us to directly assess the benefit of incorporating the full transportation-polytope constraints into the generative dynamics.

C.3 Evaluation Metrics

To rigorously evaluate disordered crystals, we propose an extended structure matching metric suite. Building upon the standard evaluation protocol for ordered crystals, we additionally incorporate continuous site-wise occupancies into the structural matching process. Based on this, our models are evaluated using four key metrics: Match Rate, RMSD, OccMAE, and structural validity.

Let $\{(\hat{\mathcal{D}}^{(n)}, \mathcal{D}^{(n)})\}_{n=1}^{N_{\text{test}}}$ denote the generated/reference pairs. The structure-matching stage uses

$$(\text{ltol}, \text{stol}, \text{angle_tol}, \text{occtol}) = (0.3, 0.5, 10, 0.3), \tag{32}$$

where ltol is the relative lattice-length tolerance, stol is the normalized site-distance tolerance, angle_tol is the lattice-angle tolerance in degrees, and occtol is the occupancy tolerance.

The choice of $\text{occtol} = 0.3$ is designed to align seamlessly with the relative scale of the lattice and spatial tolerances ($\text{ltol} = 0.3, \text{stol} = 0.5$). This threshold permits reasonable generative variance in continuous fractional occupancies while rigorously rejecting fundamentally mismatched atomic assignments.

To demonstrate that this threshold does not artificially inflate performance, we comprehensively evaluated EP-Flow on the representative MPDS Sub-20 dataset under increasingly stringent occupancy tolerances, ranging from 0.30 down to a highly strict 0.01.

As summarized in Table 5, tightening occtol from 0.30 to 0.01 only reduces the Match Rate from 60.82% to 50.09%, while the OccMAE naturally decreases to 0.0004. The robust retention of the Match Rate under such extreme constraints confirms that the model genuinely captures highly precise continuous site-wise occupancies, rather than merely passing a loose matching threshold.

Table 5: Performance of EP-Flow on the MPDS Sub-20 dataset under varying occupancy tolerances.

occtol	Successful Matches	Match Rate	Normalized RMSD	OccMAE
0.30	4,127	60.82%	0.0483	0.0075
0.25	4,067	59.93%	0.0483	0.0058
0.20	4,012	59.12%	0.0480	0.0048
0.15	3,916	57.71%	0.0480	0.0033
0.10	3,804	56.06%	0.0479	0.0020
0.05	3,653	53.83%	0.0475	0.0010
0.01	3,399	50.09%	0.0444	0.0004
0.005	3,186	46.95%	0.0438	0.0002

For each pair, we first compute a vacancy-aware site-wise occupancy discrepancy matrix. Let $p_{i,e}^{(n)}$ and $g_{j,e}^{(n)}$ denote the occupancies of species e at predicted site i and reference site j, respectively, with vacancy included explicitly as an additional species. Then

$$d_{\text{occ}}^{(n)}(i,j) = \frac{1}{2} \sum_{e} \left| p_{i,e}^{(n)} - g_{j,e}^{(n)} \right|, \tag{33}$$

which is exactly the total variation distance. A site pair is chemically feasible if $d_{\text{occ}}^{(n)}(i,j) \leq \text{occtol}$.

Next, for every lattice-aligned candidate and periodic shift, the pairwise spatial distance is computed on the ground-truth lattice and normalized by the average free length

$$\ell_{\text{free}}^{(n)} = \left(\frac{\text{Vol}(\mathcal{D}^{(n)})}{N_n}\right)^{1/3}, \qquad d_{\text{spa}}^{(n)}(i,j) = \frac{\text{dist}^{(n)}(i,j)}{\ell_{\text{free}}^{(n)}}, \tag{34}$$

where $\text{Vol}(\cdot)$ denotes the volume and $\text{dist}(\cdot,\cdot)$ is the distance between two atoms. A site pair is spatially feasible if $d_{\text{spa}}^{(n)}(i,j) \le \text{stol}$.

The final valid edge set is the intersection of the occupancy-feasible and spatially feasible masks. In this bipartite graph, we solve a Hungarian assignment [37] by minimizing the normalized spatial cost. A sample is counted as a successful match only when a perfect N_n-to-N_n assignment exists.

Let $\mathcal{V}$ denote the set of samples for which both the predicted and reference structures are valid:

$$\mathcal{V} = \left\{ n \mid \hat{\mathcal{D}}^{(n)} \text{ and } \mathcal{D}^{(n)} \text{ are both valid 3D structures} \right\}. \tag{35}$$

We define the set of successful matches as

$$\mathcal{S} = \{ n \in \{1, \dots, N_{\text{test}}\} \mid \text{sample } n \text{ admits a valid perfect assignment} \}. \tag{36}$$

For each $n \in \mathcal{S}$, let π_n denote the optimal assignment. In our implementation, matching is performed only when both the prediction and the ground truth pass the validity checks. Therefore, $\mathcal{S} \subseteq \mathcal{V}$. The reported **Match Rate** is

$$\text{MR} = \frac{|\mathcal{S}|}{N_{\text{test}}}. \tag{37}$$

For each successful sample $n \in \mathcal{S}$, the per-structure normalized RMSD is

$$\text{RMSD}^{(n)} = \sqrt{\frac{1}{N_n} \sum_{i=1}^{N_n} \left(d_{\text{spa}}^{(n)}\big(i, \pi_n(i)\big) \right)^2}, \tag{38}$$

and the reported dataset-level **RMSD** is the average over successful samples,

$$\text{RMSD} = \frac{1}{|\mathcal{S}|} \sum_{n \in \mathcal{S}} \text{RMSD}^{(n)}. \tag{39}$$

Similarly, the per-structure occupancy error is

$$\text{OccMAE}^{(n)} = \frac{1}{N_n} \sum_{i=1}^{N_n} d_{\text{occ}}^{(n)}\big(i, \pi_n(i)\big), \tag{40}$$

and the reported **OccMAE** is

$$\text{OccMAE} = \frac{1}{|\mathcal{S}|} \sum_{n \in \mathcal{S}} \text{OccMAE}^{(n)}. \tag{41}$$

The **Validity** score follows the crystal-construction and sanity checks: a prediction is valid only if it can be converted to a pymatgen structure [38] and passes the structural validity test. Therefore,

$$\text{Validity} = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \mathbb{I}\left[\hat{\mathcal{D}}^{(n)} \text{ has no structure overlap} \right]. \tag{42}$$

D Implementation Details

D.1 Disordered Crystal Structure

Crucially, the Occupancy Distribution Matrix (ODM) should not be interpreted as a single, static snapshot of a disordered crystal. Instead, it serves as a statistical summary of all possible local atomic arrangements. As illustrated in Fig. 6, while an ordered structure assigns exactly one specific element to each crystallographic site, a disordered structure allows equivalent sites to host different elements—or remain completely empty—across different regions of the material. The vacancy-extended ODM captures this variability by assigning each site a probability distribution over all candidate species, explicitly including the vacancy pseudo-atom. Therefore, each entry A_{ij} simply represents the probability, or the average fraction, that species j occupies site i. This formulation enables EP-Flow to model disordered crystals at an average, macroscopic level. It elegantly translates the physical complexity of mixed atomic arrangements into clear geometric marginal constraints, while still preserving physically meaningful, site-specific disorder patterns.

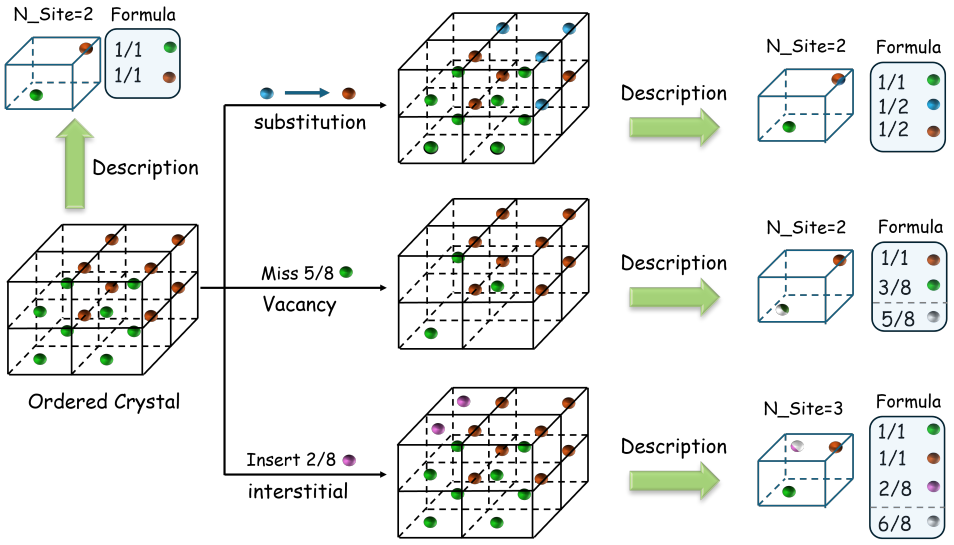


Figure 6: **Unified representation of diverse disorder types via the vacancy-extended Occupancy Distribution Matrix (ODM).** In an ordered crystal (left), the site count N is well-defined by the formula with discrete occupancies. When transitioning to disordered systems (right), this assumption breaks down. For **substitutional disorder** (top right), occupancies become continuously fractional while N remains unchanged. For **vacancy** (middle right) and **interstitial disorder** (bottom right), the site count cannot be determined from the raw formula alone. To resolve this ambiguity, a vacancy pseudo-atom (Vac, depicted as grey/white spheres) is introduced to ensure that the occupancy at each site sums to one. This yields an extended formula $\mathbf{c}$ that incorporates the total vacancy count, satisfying $\sum_{j=1}^{M} c_j = N$. This formulation explicitly conditions the structural representation on both the formula and the site number, allowing all disorder types to be modeled within a unified ODM framework.

D.2 Formula and Number of Sites

In ordered crystal generation, the number of atomic sites is usually well defined from the reduced chemical formula. Each crystallographic site is occupied by a single atom, and therefore the total number of atoms in the formula naturally equals the number of nodes in the crystal graph. This convention is adopted by many graph-based crystal generation models, such as DiffCSP and FlowMM, where the model typically needs to know or sample the number of nodes before generating fractional coordinates and lattice parameters.

However, this assumption does not directly hold for disordered crystals. In a disordered structure, a crystallographic site may be fractionally occupied by multiple species, or may be partially vacant.

As a result, the raw chemical formula may contain non-integer stoichiometric coefficients and its total sum does not necessarily reveal the number of crystallographic sites. For example, a partially occupied site with occupancy 0.5 contributes only half an atom to the chemical formula, while it still corresponds to one crystallographic site in the structure representation. Therefore, the site count cannot be inferred by simply summing the non-vacancy species in the formula.

To obtain a consistent representation, we explicitly introduce a vacancy species, denoted as Vac, to complete each crystallographic site to full occupancy. For each site i, the occupancy vector satisfies

$$\sum_{j=1}^{M} A_{ij} = 1, \tag{43}$$

where M includes both real chemical species and the vacancy species. Accordingly, the extended stoichiometric vector $\boldsymbol{c}$ also includes the total vacancy count. This gives

$$\sum_{j=1}^{M} c_j = N, \tag{44}$$

where N is the number of crystallographic sites. In other words, after vacancy completion, the extended formula is fully occupied at the site level, and the sum of the extended formula equals the number of sites.

To better illustrate how the vacancy-augmented formula bridges macroscopic chemical formulas with microscopic site occupancies, let us consider a common macroscopic formula, $A_{1.5}B_1$.

Case 1 (A structure with 3 sites): Suppose we read the structural information of a crystal with this formula and find it has exactly $N = 3$ crystallographic sites. The actual site occupancies might be: Site 1 (A, occupancy = 1.0), Site 2 (B, occupancy = 1.0), and Site 3 (A, occupancy = 0.5). To address the partial occupancy on Site 3, we introduce a "vacancy" (Vac) pseudo-atom to pad the remaining probability. Thus, Site 3 becomes a mixed site: (A: 0.5, Vac: 0.5). This transforms the vacancy defect into a purely substitution-disorder format. The corresponding augmented formula becomes $A_{1.5}B_1\mathrm{Vac}_{0.5}$. The sum of the base formula subscripts ($1.5 + 1 + 0.5 = 3$) perfectly matches the actual number of sites ($N = 3$).

Case 2 (A structure with 4 sites): Now, suppose another structure with the exact same conventional formula $A_{1.5}B_1$ has $N = 4$ crystallographic sites. Its actual site occupancies might be: Site 1 (A, occupancy = 1.0), Site 2 (B, occupancy = 1.0), Site 3 (A, occupancy = 0.25), and Site 4 (A, occupancy = 0.25). We pad the empty probabilities on Sites 3 and 4 with the vacancy pseudo-atom, resulting in Site 3 (A: 0.25, Vac: 0.75) and Site 4 (A: 0.25, Vac: 0.75). The augmented formula for this specific structure becomes $A_{1.5}B_1\mathrm{Vac}_{1.5}$. Again, the sum of the base formula subscripts ($1.5 + 1 + 1.5 = 4$) perfectly matches the site count ($N = 4$).

In practice, this simple yet effective padding mechanism guarantees that our transportation-polytope framework can unify substitution disorder and vacancy defects under a single mathematical formulation. During training, we utilize these fully padded, vacancy-augmented formulas alongside their corresponding ODMs. During inference, the augmented formula serves as the deterministic conditional input. It is directly derived from the conventional chemical formula and the target number of sites N by calculating the vacancy amount as:

$$\mathrm{Amount}_{\mathrm{Vac}} = N - \sum(\text{base formula subscripts}). \tag{45}$$

For our examples, this is $3 - (1.5 + 1) = 0.5$ for Case 1, and $4 - (1.5 + 1) = 1.5$ for Case 2.

This treatment is important for disordered crystal generation. It allows substitutional disorder and vacancy disorder to be represented in a unified occupancy distribution matrix (ODM), while preserving the same node-level structure as graph-based crystal models. The model therefore generates a fixed number of crystallographic sites N, each associated with a fractional coordinate and a probability distribution over real species and vacancies. The final physical composition is recovered by marginalizing out the vacancy species, whereas the extended formula ensures that every site remains fully occupied in the ODM representation.

D.3 Log-domain Double Centering

ϵ-smoothing and log-domain canonicalization. Real ODMs are highly sparse, and many entries of $\boldsymbol{A}$ are exactly zero. Directly applying $\log \boldsymbol{A}$ is therefore ill-defined. We first apply a small ϵ-smoothing to move the ODM from the boundary of the transportation polytope to its relative interior:

$$\boldsymbol{A}^{\epsilon} = (1-\epsilon)\boldsymbol{A} + \epsilon \boldsymbol{U}_{\boldsymbol{B}}, \qquad U_{ij} = \frac{B_{ij}}{\sum_{k=1}^{M} B_{ik}}. \quad (46)$$

We then perform masked log-domain double centering:

$$\boldsymbol{A}' = \mathfrak{D}_{\boldsymbol{B}} \left(\log \boldsymbol{A}^{\epsilon} \right). \quad (47)$$

This removes site-wise and species-wise additive effects and provides a shared zero-marginal representation for flow matching.

Log-domain Double Centering

Ensure: Centered representation $\boldsymbol{A}'$

$\boldsymbol{U}_{\boldsymbol{B}} \leftarrow \mathrm{MaskedUniform}(\boldsymbol{B})$
$\boldsymbol{A}^{\epsilon} \leftarrow (1-\epsilon)\boldsymbol{A} + \epsilon \boldsymbol{U}_{\boldsymbol{B}}$
$\boldsymbol{X} \leftarrow \log \boldsymbol{A}^{\epsilon}$
$\widetilde{\boldsymbol{X}} \leftarrow \boldsymbol{X} \odot \boldsymbol{B}$
$r \leftarrow \mathrm{MaskedRowMean}(\widetilde{\boldsymbol{X}}, \boldsymbol{B})$
$q \leftarrow \mathrm{MaskedColumnMean}(\widetilde{\boldsymbol{X}}, \boldsymbol{B})$
$g \leftarrow \mathrm{MaskedGlobalMean}(\widetilde{\boldsymbol{X}}, \boldsymbol{B})$
$\boldsymbol{A}' \leftarrow \boldsymbol{B} \odot (\widetilde{\boldsymbol{X}} - r - q + g)$
return $\boldsymbol{A}'$

D.4 Sinkhorn Recovery

Formula-conditioned recovery with truncation. The generated centered matrix $\boldsymbol{A}'$ is not itself a valid occupancy matrix. We first mask species absent from the target formula and exponentiate the stabilized logits to obtain a positive kernel:

$$\boldsymbol{K} = \exp(\boldsymbol{Z}) \odot \boldsymbol{B} + \epsilon \boldsymbol{B}, \qquad \boldsymbol{B} = \mathbf{1}[\boldsymbol{c} > 0]. \quad (48)$$

The kernel is then projected onto the formula-conditioned transportation polytope by Sinkhorn normalization:

$$\boldsymbol{A} = \mathcal{S}\left(\boldsymbol{K}, \mathbf{1}_N, \boldsymbol{c}\right), \quad (49)$$

Specifically, the Sinkhorn algorithm seeks two positive scaling vectors $\boldsymbol{u} \in \mathbb{R}^{N}_{>0}$ and $\boldsymbol{v} \in \mathbb{R}^{M}_{>0}$ such that the scaled matrix $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{u})\boldsymbol{K}\mathrm{diag}(\boldsymbol{v})$ meets the prescribed marginals. This is efficiently achieved through alternating multiplicative updates:

$$\boldsymbol{u}^{(k+1)} = \frac{\mathbf{1}_N}{\boldsymbol{K}\boldsymbol{v}^{(k)}}, \qquad \boldsymbol{v}^{(k+1)} = \frac{\boldsymbol{c}}{\boldsymbol{K}^{\top}\boldsymbol{u}^{(k+1)}}, \quad (50)$$

where the division is performed element-wise, and $\boldsymbol{v}^{(0)}$ is typically initialized to $\mathbf{1}_M$. After a finite number of iterations K_{SK}, the recovered matrix $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{u}^{(K)})\boldsymbol{K}\mathrm{diag}(\boldsymbol{v}^{(K)})$ satisfies the marginal constraints up to numerical precision:

$$\boldsymbol{A}\mathbf{1}_M = \mathbf{1}_N, \qquad \boldsymbol{A}^{\top}\mathbf{1}_N = \boldsymbol{c}, \qquad A_{ij} \geq 0. \quad (51)$$

During CIF reconstruction, we observe that the recovered ODM may contain very small nonzero occupancies on some site–species pairs with negligible contribution, sometimes around 10^{-5}. If directly written into CIF files, these trace values may be interpreted as additional mixed species, leading to unnecessarily dense occupancy annotations.

This behavior is mainly caused by the numerical recovery procedure. First, Sinkhorn recovery operates on strictly positive kernels induced by ϵ-smoothing, so valid entries may retain small positive masses after normalization. Second, Sinkhorn normalization is implemented with a finite number of iterations, which can leave tiny residual values on numerically feasible site–species pairs.

To obtain cleaner reconstructed structures, we truncate entries with $0 < A_{ij} < \tau$ to zero after the first Sinkhorn projection. In our experiments, we set $\tau = 0.005$, corresponding to occupancies below 0.5% of a crystallographic site. This threshold is intentionally conservative: it removes only trace values below the occupancy scale considered meaningful in our reconstruction protocol, while preserving genuine fractional occupancies. After truncation, we apply a second Sinkhorn projection to restore the prescribed row and column marginals up to numerical tolerance.

This post-processing produces a sparse and formula-consistent ODM, helps avoid interpreting negligible numerical residuals as meaningful mixed-site disorder during CIF reconstruction, and yields reconstructed crystals with more physically interpretable occupancies.

Sinkhorn Recovery

Ensure: Recovered ODM $\boldsymbol{A}$

$\boldsymbol{B} \leftarrow \mathbf{1}[\boldsymbol{c} > 0]$
$\boldsymbol{Z} \leftarrow \mathrm{MaskInvalid}(\boldsymbol{A}', \boldsymbol{B})$
$\boldsymbol{Z} \leftarrow \boldsymbol{Z} - \mathrm{RowMax}(\boldsymbol{Z})$
$\boldsymbol{K} \leftarrow \exp(\boldsymbol{Z}) \odot \boldsymbol{B} + \epsilon \boldsymbol{B}$
$\boldsymbol{K} \leftarrow \mathrm{Sinkhorn}(\boldsymbol{K}; \mathbf{1}_N, \boldsymbol{c}, K)$
$K_{ij} \leftarrow 0, \quad \text{if } 0 < K_{ij} < \tau$
$\boldsymbol{A} \leftarrow \mathrm{Sinkhorn}(\boldsymbol{K}; \mathbf{1}_N, \boldsymbol{c}, K)$
return $\boldsymbol{A}$

D.5 Structural Reconstruction

Figure 7 visualizes representative reconstruction trajectories for three disordered oxide systems. At $T = 1$, the structures are highly noisy and mainly preserve coarse compositional information. As the trajectory evolves toward $T = 0$, EP-Flow progressively recovers ordered atomic positions and site-wise occupancy patterns. The final predictions are valid, formula-matched, and successfully matched to the corresponding ground-truth structures.

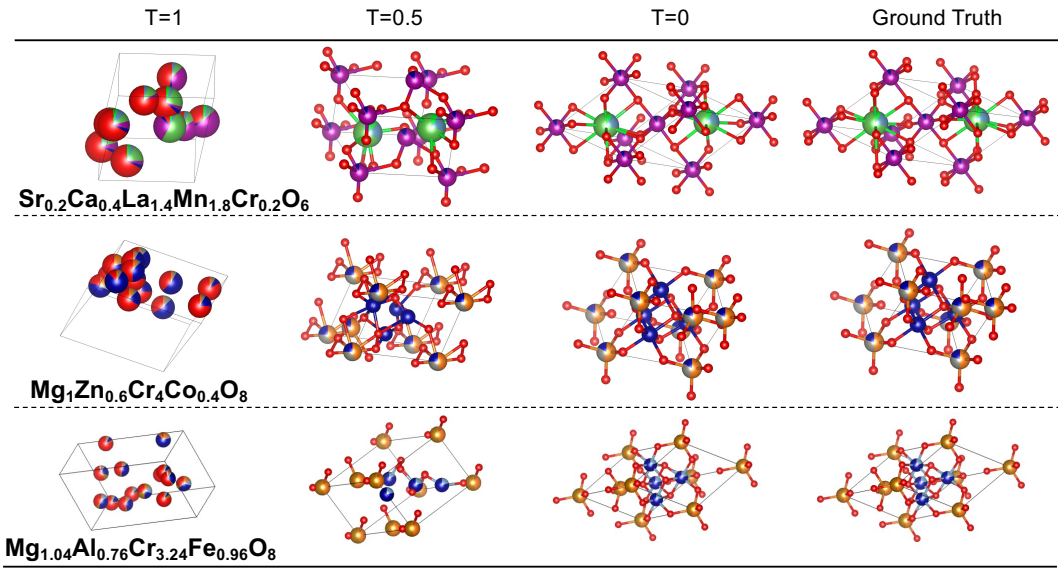


Figure 7: **Structural reconstruction trajectories of three representative disordered oxides.** Each row shows the evolution from the noisy state at $T = 1$ to the reconstructed structure at $T = 0$, compared with the ground truth.

Table 6: Reconstruction metrics for the representative disordered oxide cases in Fig. 7.

Formula	**RMSD (↓)**	**OccMAE (↓)**	**Max Occ Error (↓)**
$Sr_{0.2}Ca_{0.4}La_{1.4}Mn_{1.8}Cr_{0.2}O_6$	0.0188	0.0006	0.0019
$Mg_1Zn_{0.6}Cr_4Co_{0.4}O_8$	0.0191	0.0053	0.0371
$Mg_{1.04}Al_{0.76}Cr_{3.24}Fe_{0.96}O_8$	0.0151	0.0052	0.0335

Table 6 summarizes the quantitative reconstruction accuracy. All three examples achieve RMSD below 0.02, indicating accurate reconstruction of fractional coordinates and lattice geometry. The OccMAE values are also small, showing that EP-Flow recovers site-wise fractional occupancies with high fidelity. The largest occupancy errors mainly occur on mixed cation sites, where multiple chemically similar species compete for the same crystallographic positions.

Table 7: Site-wise occupancy comparison for representative structural reconstruction cases. Only mixed-occupancy sites are reported; fully occupied sites are omitted for clarity.

Formula	**Mixed site**	**Reference occupancy**	**Predicted occupancy**	**ΔOcc**
$Sr_{0.2}Ca_{0.4}La_{1.4}Mn_{1.8}Cr_{0.2}O_6$	B-site	$Mn_{0.900}Cr_{0.100}$	$Mn_{0.8987}Cr_{0.1013}$, $Mn_{0.9013}Cr_{0.0987}$	0.0013
	A-site	$Sr_{0.100}Ca_{0.200}La_{0.700}$	$Sr_{0.1019}Ca_{0.1988}La_{0.6993}$, $Sr_{0.0981}Ca_{0.2012}La_{0.7007}$	0.0019
$Mg_1Zn_{0.6}Cr_4Co_{0.4}O_8$	Cation site	$Mg_{0.500}Zn_{0.300}Co_{0.200}$	$Mg_{0.5371}Zn_{0.2726}Co_{0.1903}$, $Mg_{0.4629}Zn_{0.3274}Co_{0.2097}$	0.0371
$Mg_{1.04}Al_{0.76}Cr_{3.24}Fe_{0.96}O_8$	Mg/Fe site	$Mg_{0.520}Fe_{0.480}$	$Mg_{0.5535}Fe_{0.4465}$, $Mg_{0.4865}Fe_{0.5135}$	0.0335
	Al/Cr site	$Al_{0.190}Cr_{0.810}$	$Al_{0.1884–0.1914}Cr_{0.8086–0.8116}$	0.0016

Table 7 further compares the reference and predicted occupancies on mixed sites. For $Sr_{0.2}Ca_{0.4}La_{1.4}Mn_{1.8}Cr_{0.2}O_6$, EP-Flow almost exactly recovers both A-site Sr/Ca/La disorder and

B-site Mn/Cr disorder, with maximum occupancy errors below 0.002. This case demonstrates that EP-Flow can accurately reconstruct coupled A/B-site disorder in a perovskite-like oxide framework.

For $\mathrm{Mg_1Zn_{0.6}Cr_4Co_{0.4}O_8}$, the Mg/Zn/Co mixed site is more challenging because three species compete on the same cation sublattice. The prediction preserves the global formula but slightly redistributes occupancy between the two mixed sites: one site becomes slightly Mg-rich, while the other becomes relatively Zn/Co-rich. This suggests that the main residual error arises from local redistribution among multiple mixed species rather than failure to satisfy the global composition.

For $\mathrm{Mg_{1.04}Al_{0.76}Cr_{3.24}Fe_{0.96}O_8}$, the Al/Cr mixed sites are reconstructed accurately, with maximum occupancy error around 0.0016. The main residual error comes from the Mg/Fe mixed sites, where EP-Flow predicts one site as slightly Mg-rich and the other as slightly Fe-rich. Importantly, the two predicted Mg/Fe sites still sum to the correct total composition, $\mathrm{Mg_{1.04}Fe_{0.96}}$. Thus, the error mainly reflects local redistribution among chemically similar mixed sites rather than violation of the global stoichiometry. Such site-level differentiation is chemically plausible in disordered crystals, although the crystallographic average structure remains the reference for quantitative evaluation.

D.6 Hyperparameters

Table 8 summarizes the concrete architecture hyperparameters and optimization-related quantities exposed by the current implementation. All experiments were conducted on NVIDIA GeForce RTX 4090 GPUs with 24 GB memory. Models on COD were trained on a single GPU, while MPDS models were trained with two-GPU distributed data parallelism due to the larger dataset scale. We used 16-bit mixed precision for all training runs.

Table 8: Architecture and training-related hyperparameters.

Hyperparameter	Symbol / module	Value
Hidden dimension	d_h	512
Message-passing layers	L	6
Activation function	σ	SiLU
Learning rate	lr	5×10^{-4}
LR decay factor	γ_{lr}	0.85
LR scheduler patience (epochs)	P_{lr}	50
Training epochs	T_{epoch}	1000
Sinkhorn iterations	K_{SK}	100
ODE steps	I_{step}	200

Table 9: Dataset-specific training hyperparameters and generative loss weights.

Database	**Setting**	**Batch Size**	**Sampling** `step_lr`	$\boldsymbol{\lambda_A}$	$\boldsymbol{\lambda_F}$	$\boldsymbol{\lambda_{\mathrm{len}}}$	$\boldsymbol{\lambda_{\mathrm{ang}}}$
COD	Substitution	256	3	20.0	100.0	1.0	100.0
	Disorder			20.0	100.0	1.0	100.0
MPDS	Substitution	512	5	20.0	100.0	5.0	500.0
	Disorder			20.0	100.0	5.0	500.0

E Limitations

EP-Flow currently focuses on formula-conditioned generation for disordered crystals. Although the global stoichiometry provides the main compositional constraint, the generation process also implicitly depends on the number of crystallographic sites. In practice, the site count is either specified or sampled from the empirical distribution of the benchmark setting. Therefore, fully determining the site number directly from the formula alone remains an open problem, especially for disordered structures where the summed non-vacancy stoichiometry does not necessarily equal the number of crystallographic sites.

Second, EP-Flow does not currently support property-conditioned generation. The present framework is designed to generate structurally valid disordered crystals under stoichiometric constraints, but it does not directly optimize target properties such as ionic conductivity, magnetism, superconductivity, or thermodynamic stability. Extending EP-Flow to property-guided or inverse-design settings is an important direction for future work.

Finally, EP-Flow does not currently condition on experimental measurements such as PXRD, PDF, or spectroscopy data. Incorporating such information could further constrain the generation process and make the framework more useful for experimental structure determination and refinement.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope as formula-to-disordered-crystal generation, the EP-Flow method, and the empirical gains on curated COD/MPDS benchmarks, which are exactly the contributions supported by the theory and experiments (Abstract; Sections 1, 4 and 5).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main paper states the key theoretical claims, and the appendix provides formal proofs for the bijectivity of canonicalization, trajectory feasibility, and the invariance properties used by the method (Section 4 and Appendix A).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies the task formulation, baseline adaptations, evaluation protocol, dataset construction, metric definitions, and training hyperparameters needed to reproduce the reported results (Section 5 and Appendices B and C).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The datasets and code are currently being organized and will be made publicly available in due course.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main text gives the benchmark setup, baselines, and evaluation protocol, while the appendix reports split statistics, preprocessing rules, metric implementation details, and architecture/training hyperparameters (Section 5.1; Appendices C–D).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The current experiments do not involve repeated runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided the computer resources in Appendix D.6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work is computational materials-science research using existing scientific structure databases and does not involve human subjects, sensitive personal data, or unsafe intervention studies.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer:[No]

    Justification: The paper motivates the materials-discovery value of disordered crystal generation, but it does not currently include a dedicated discussion of broader positive and negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work studies scientific crystal-structure generation and does not release a high-risk generative model or scraped human-centered dataset that would require special misuse safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper credits and cites the original databases and baseline methods used in this work. Detailed license and terms-of-use information can be found at their original sources.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The datasets and code are currently being organized and will be made publicly available in due course.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, human participants, or human-annotated data collection.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The work does not involve human subjects or participant studies, so IRB review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the proposed method, experiments, or scientific claims, so no declaration is required under the checklist policy.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.