GRAND CANONICAL GENERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce Grand Canonical Generators (GCG), a generative framework that extends Boltzmann generators to the grand canonical ensemble. We present two designs. The first conditions a variable-size generative model on the chemical potential, sampling particle number and configuration jointly. The second factorizes the grand canonical distribution into a particle-number distribution and the corresponding canonical Boltzmann density. This factorized formulation can use any existing Boltzmann generator for the canonical component, encodes the known linear chemical-potential dependence analytically, and yields a tractable likelihood that supports self-normalized importance sampling (SNIS). Empirically, GCG accurately reproduces grand canonical observables on a Lennard–Jones fluid and methane adsorption in a zeolite, demonstrating generalization across chemical potentials and correction via SNIS and grand canonical Monte Carlo.

1 INTRODUCTION

Sampling molecular systems that exchange particles with an external reservoir is of central interest in problems ranging from gas adsorption in porous materials for carbon capture (Millward & Yaghi, 2005) to solvation free-energy estimation in drug design (Jorgensen & Ravimohan, 1985). Such open systems are described by the grand canonical ensemble, whose equilibrium distribution spans both a variable particle number and a continuous molecular configuration at each particle number. Accurately computing thermodynamic observables, like adsorption isotherms, solvation free energies, and fluid-phase coexistence, requires sufficient sampling across all relevant particle-numbers.

The dominant approach to sampling open systems is grand canonical Monte Carlo (GCMC) (Adams, 1975), which augments conventional Monte Carlo with particle insertion and deletion moves. However, trial insertions frequently overlap existing particles, which leads to low acceptance rates and long correlation times (Adams, 1975; Melling et al., 2023). Because Monte Carlo methods move through configuration space in small sequential steps, escaping metastable states and exploring the full space is slow. The problem worsens when the system spans a wide range of particle numbers: each particle number must be visited often enough to estimate its statistical weight reliably, which further extends simulation times.

Boltzmann generators (BGs) provide an amortized alternative to molecular simulation, by learning to generate independent, approximately equilibrium configurations in the canonical ensemble (Noé et al., 2019). When the generator provides a tractable likelihood, importance-weighted reweighting corrects residual model error without retraining. However, existing BGs fix the particle number and therefore cannot be used for the grand canonical ensemble.

Going from the canonical to the grand canonical ensemble introduces two challenges. First, the state space becomes a union of configuration spaces of different dimension. Second, the probability of each particle number depends on the canonical partition functions, which cancel in canonical reweighting and are therefore never estimated by a BG. Because these probabilities shift exponentially with the chemical potential, a grand canonical sampler must implicitly learn free-energy differences across particle numbers.

Extending the BG paradigm to open systems thus requires (1) a variable-size generative model, (2) a particle-number distribution that respects the chemical-potential dependence, (3) a tractable joint likelihood for inference-time correction via importance-sampling. No existing method addresses all three requirements.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

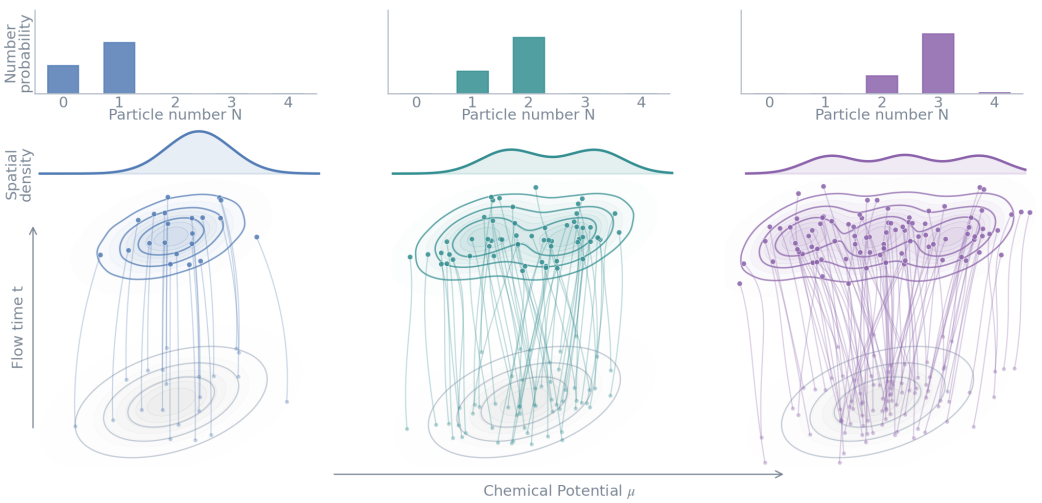

Figure 1: Grand Canonical Generators learn to sample from the μVT ensemble, which defines a distribution over both the particle number and geometries, conditioned on the chemical potential μ.

Present work. In this work, we introduce **Grand Canonical Generators (GCG)**, to our knowledge, the first Boltzmann generator for the grand canonical ensemble that operates without lattices and provides a tractable joint likelihood. We propose two instantiations of that framework that exploit the ensemble structure to different degrees.
A **Joint GCG** models the complete variable-size state in a single generative process. It meets requirement (1), but it must learn the chemical-potential dependence from data, and the trans-dimensional Markov process we use provides no tractable likelihood.
A **Factorized GCG** exploits that at fixed particle number, the grand canonical distribution reduces to the canonical Boltzmann distribution, independent of the chemical potential. It first samples from a learned particle-number model and then from a conditional BG. The particle-number model applies known factors analytically, as in grand canonical histogram reweighting (Ferrenberg & Swendsen, 1989), learning only ratios of canonical partition functions, which improves transfer to unseen chemical potentials. This satisfies requirement (2). The product of both likelihoods yields a tractable joint likelihood that supports grand canonical self-normalized importance sampling (SNIS), including a stratified variant that maintains coverage across all relevant particle numbers. The factorized design thus meets all three requirements.

We validate GCG on three systems that probe distinct aspects of the framework. On an **ideal gas**, the model recovers the exact particle-number distribution and spatial uniformity, confirming correct variable-size generation without interaction effects. On a periodic **Lennard–Jones fluid**, the factorized construction reproduces reference thermodynamic observables after importance-sampling correction and transfers to nearby chemical potentials outside the training range. On **methane adsorption in MFI zeolite**, GCG generates adsorbed configurations inside the crystalline host and recovers the reference adsorption isotherm.

2 Background

We seek to sample from the grand canonical density , also known as the μVT-ensemble

$$\gamma_\mu(N, X; c) = \frac{\exp[-\beta U_N(X; c) + \beta\mu N]}{N!\,\Lambda^{3N}}, \tag{1}$$

$$\pi(N, X \mid \beta, \mu, c) = \frac{\gamma_\mu(N, X; c)}{\Xi(\beta, \mu, c)}, \quad \Xi(\beta, \mu, c) = \sum_{N=0}^{\infty} \int_{\mathcal{V}^N} \gamma_\mu(N, X; c)\, \mathrm{d}X. \tag{2}$$

It defines a distribution over both particle numbers $N \in \mathbb{Z}_{\geq 0}$ and coordinates $X \in \mathbb{R}^{3N}$, given the inverse temperature $\beta = 1/(k_\mathrm{B}T)$, volume V, and chemical potential μ. The denominator Ξ is the (intractable) grand canonical partition function. $\mathcal{V}^N$ is the configuration space for N particles in the volume V, and Λ is the thermal de Broglie wavelength, which accounts for momentum

degrees of freedom. Unlike the canonical ensemble, which fixes N, the grand canonical ensemble allows the system to exchange particles with a reservoir. The chemical potential μ controls this particle exchange and therefore the probability of each particle number. The context c describes the surrounding system, e.g. a periodic cell.

Boltzmann generators instead approximate the canonical density at a fixed particle number:

$$\pi(X \mid N, \beta, c) = \frac{\exp[-\beta U_N(X; c)]}{Z_N(\beta, c)}, \qquad Z_N(\beta, c) = \int_{\mathcal{V}^N} \exp[-\beta U_N(X; c)]\, \mathrm{d}X. \tag{3}$$

treating N, V, T as given and fixed. The canonical ensemble is also known as the Boltzmann distribution or the NVT-ensemble. Typically, a BG uses a generative model $q_\theta(X \mid N, \beta, c) \approx \pi(X \mid N, \beta, c)$ (Noé et al., 2019) to transform a simple reference distribution into the target distribution through a learned transport map, like a normalizing flow. Importantly, a BG never requires Z_N, since it cancels in self-normalized reweighting. In contrast, marginalizing Eq. (2) over X yields

$$\pi(N \mid \beta, \mu, c) = \frac{e^{\beta\mu N}\, Z_N(\beta, c)}{N!\, \Lambda^{3N}\, \Xi(\beta, \mu, c)}, \tag{4}$$

so the particle-number distribution depends on the intractable Z_N and shifts exponentially with μ. This makes the grand canonical distribution significantly more challenging.

Flow matching (Lipman et al., 2023) trains a continuous normalizing flow (CNF) without backpropagating through its sampling dynamics. Given source and target samples $X_0 \sim \rho_N$, $X_1 \sim \pi(\cdot \mid N, \beta, c)$, a conditional probability path yields an interpolant X_t and a target velocity $u_t(X_t \mid X_0, X_1)$. A neural velocity field v_θ is trained by regressing onto the conditional flow matching (CFM) target:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t, X_0, X_1, X_t} \left\| v_\theta(X_t, t; N, \beta, c) - u_t(X_t \mid X_0, X_1) \right\|^2. \tag{5}$$

At inference, integrating the ordinary differential equation (ODE) $\dot{X}_t = v_\theta(X_t, t; N, \beta, c)$ from $t{=}0$ to $t{=}1$ maps source samples to generated configurations. Under standard regularity conditions, the change-of-variables formula gives the endpoint log-density

$$\log q_\theta(X_1 \mid N, \beta, c) = \log \rho_N(X_0) - \int_0^1 \nabla_X \cdot v_\theta(X_t, t; N, \beta, c)\, \mathrm{d}t. \tag{6}$$

The divergence is the trace of the $3N \times 3N$ velocity-field Jacobian, which dominates the cost of likelihood evaluation and can also be predicted directly (OuYang et al., 2026; Burger et al., 2025).

Self-normalized importance sampling corrects the generative model when the log-density in Eq. (6) is available. Let $(N_i, X_i) \sim q$ for a proposal $q(N, X \mid \beta, \mu, c) > 0$ whenever $\gamma_\mu(N, X; c) > 0$. For an observable $\mathcal{O}$, the self-normalized estimator and the associated importance weights are

$$\widehat{\mathbb{E}}_\pi[\mathcal{O}] = \frac{\sum_{i=1}^K w_i\, \mathcal{O}(N_i, X_i)}{\sum_{i=1}^K w_i}, \qquad w_i = \frac{\gamma_\mu(N_i, X_i; c)}{q(N_i, X_i \mid \beta, \mu, c)} = \frac{\exp[-\beta U_{N_i}(X_i; c) + \beta\mu N_i]}{N_i!\, \Lambda^{3N_i}\, q(N_i, X_i \mid \beta, \mu, c)}. \tag{7}$$

SNIS is asymptotically unbiased but has finite-sample bias. The effective sample size (ESS), $\mathrm{ESS} = (\sum_i w_i)^2 / \sum_i w_i^2$, measures proposal–target overlap.

Generator Matching and Morph. Flow matching assumes a fixed state dimension. Generator Matching extends flow matching to time-dependent Markov processes with both continuous and discrete components (Holderrieth et al., 2025). For a variable-size state $z = (N, X)$, the infinitesimal generator combines coordinate drift with discrete jumps:

$$\mathcal{A}_t f(z) = v_t(z) \cdot \nabla_X f(z) + \int \big[f(z') - f(z)\big]\, K_t(z, \mathrm{d}z'), \tag{8}$$

where f is an arbitrary test function, v_t is the continuous drift, and K_t defines transition rates between states with different particle numbers. Generator Matching trains the drift and jump components by matching the model to the generator of a conditional probability path. This loss jointly matches the continuous drift and discrete transition rates, so the learned process follows the desired probability path while allowing changes in particle number. Morph applies this framework to variable-size molecular generation and uses unbalanced optimal transport (OT) to couple source and target structures of different sizes (Franke et al., 2026). Unlike a CNF, the likelihood of a sample using a jump process is generally not tractable, because computing the density requires knowing and integrating over all possible paths that lead to that sample.

3 Related Work

Neural equilibrium samplers learn transformations from simple distributions to independent equilibrium configurations (Noé et al., 2019; Wirnsberger et al., 2022; Dibak et al., 2022), with extensions to equivariant and periodic particle models (Klein et al., 2023; Hoffmann et al., 2026; Grenioux et al., 2025). When likelihoods are tractable, importance reweighting yields asymptotically correct estimates (Nicoli et al., 2020; Rehman et al., 2026). However, these existing methods assume a fixed particle number. We extend equilibrium generators to the grand canonical ensemble by learning a tractable particle-number distribution alongside the spatial generator.

Learned grand canonical sampling on lattices has been explored through autoregressive models for semi-grand canonical distributions, where chemical-potential differences change the number of each species at fixed site number (Damewood et al., 2022; Du et al., 2026), and through Exponentially Tilted Thermodynamic Maps for lattice-gas states across temperature and chemical potential (Lee et al., 2026). Neural Thermodynamic Integration estimates canonical partition functions across particle numbers but does not provide samples (Máté et al., 2024). JANUS samples site occupations, vacancies, and displacements on crystalline lattice sites, correcting estimates via forward–backward path weights, and learning from energy evaluations instead of precollected samples (Blessing et al., 2026). Grand canonical diffusion changes the number of atoms through continuous voxel fields but targets structure generation (Lei et al., 2024). (Christiansen & Hammer, 2025) relax discrete atom insertion and deletion into continuous existence variables for gradient-based structure optimization at fixed chemical potential, but do not sample the equilibrium distribution. Contrary to these works we target continuous particle positions without lattice constraints, and learn a normalized proposal that enables grand canonical importance sampling.

Variable-size generation with permutation-invariant normalizing flows provide tractable densities for unordered point sets if the particle number is supplied (Rasul et al., 2019; Biloš & Günnemann, 2021). Autoregressive models generate molecules atom by atom (Cheng et al., 2025; Rehman et al., 2026), and jump processes insert and delete components during generation (Campbell et al., 2023; Holderrieth et al., 2025; Franke et al., 2026). These methods target data generation or molecular design, but have not yet been adapted to the grand canonical ensemble. As such, they can be used inside the Grand Canonical Generator framework presented in this paper.

4 Grand Canonical Generators

We introduce GCG, a framework for learned sampling in the grand canonical ensemble. We present two constructions, joint and factorized, that offer different trade-offs between flexibility and likelihood evaluation. We then describe a structured particle-number model that improves generalization across chemical potentials, and a stratified importance-sampling correction that removes model bias at inference time. We conclude with the model parameterization used in our experiments.

Joint GCG. The joint construction directly models the complete variable-size state $q_\psi^{\text{joint}}(N, X \mid \beta, \mu, c)$. This can be realized using jump diffusion processes, autoregressive models, or, in our case, with the Generator Matching framework from Morph (Holderrieth et al., 2025; Franke et al., 2026). Its generator changes particle coordinates and particle number within one process, which we condition on (β, μ, c). Joint GCG places few structural restrictions on the learned distribution. However, it must learn the chemical-potential dependence from data, and its jump process lacks a tractable endpoint likelihood.

Factorized GCG. Instead we can factorize the grand canonical distribution into particle-number and conditional configuration distributions:

$$\pi(N, X \mid \beta, \mu, c) = \pi(N \mid \beta, \mu, c) \underbrace{\frac{\exp[-\beta U_N(X; c)]}{Z_N(\beta, c)}}_{\pi(X \mid N, \beta, c)}. \tag{9}$$

using the canonical partition function $Z_N(\beta, c)$ from Eq. (3). The conditional configuration distribution $\pi(X \mid N, \beta, c)$ is now the canonical Boltzmann ensemble, which no longer depends on μ. This allows us to learn a *Factorized GCG*:

$$q_{\phi,\theta}(N, X \mid \beta, \mu, c) = q_\phi(N \mid \beta, \mu, c)\, q_\theta(X \mid N, \beta, c). \tag{10}$$

Sampling first draws $N \sim q_\phi$ and then $X \sim q_\theta(\cdot \mid N, \beta, c)$. Conveniently, any BG architecture can be conditioned on the particle number and serve as q_θ. If $q_\phi(N \mid \beta, \mu, c)$ and $q_\theta(X \mid N, \beta, c)$ support likelihood evaluation, the resulting joint likelihood can be used for SNIS.

Particle-number model. Sampling from $q_{\phi,\theta}(N, X \mid \beta, \mu, c)$ requires a tractable distribution over particle numbers whose likelihood we can evaluate exactly. A natural choice is a categorical distribution $q_\phi(N \mid \beta, \mu, c) = \mathrm{softmax}_N[\ell_\phi(N; \beta, \mu, c)]$ over $\{0, \ldots, N_{\max}\}$. While unrestricted logits suffice in principle, we can exploit the known structure of grand canonical weights to improve transfer across chemical potentials.

At fixed (β, c), and with the thermal-wavelength factor Λ^3 absorbed into the activity $e^{\beta\mu}$, the grand canonical ensemble assigns each particle number the unnormalized weight $\widetilde{p}(N) = e^{\beta\mu N} Z_N/N!$ (Adams, 1975). Taking the logarithm separates two analytically known terms from the unknown canonical partition function $\log \widetilde{p}(N) = \beta\mu N - \log(N!) + \log Z_N$. We retain the known terms and represent only the residual $\log Z_N$ with a learned function r_ϕ: $\ell_\phi(N; \beta, \mu, c) = \beta\mu N - \log(N!) + r_\phi(N, \beta, c)$. Crucially, r_ϕ does not depend on μ. Changing the chemical potential applies the exponential tilt $\beta\mu N$ without retraining. Transfer to chemical potentials outside the training range thus reduces to extrapolating $\log Z_N$ to particle numbers outside the training data. This parameterization mirrors grand canonical histogram reweighting (Ferrenberg & Swendsen, 1989; Wilding, 1995). With a free residual per particle number, the maximum-likelihood estimate of r_ϕ on data pooled from several chemical potentials coincides with the multiple-histogram (WHAM/MBAR) estimate of $\log Z_N$ (Kumar et al., 1992; Shirts & Chodera, 2008). The neural residual additionally shares information across β and c, but, as in histogram reweighting, transfer is reliable only where the pooled data cover the relevant particle numbers.

Since the softmax is invariant to additive shifts, the model only needs to learn the ratios $\log(Z_N/Z_0)$. We fix this gauge freedom by setting $r_\phi(0, \beta, c) = 0$. We train the particle-number model jointly with q_θ by adding the maximum-likelihood loss to the flow-matching objective.

$$\mathcal{L}_{\mathrm{N}}(\phi) = -\mathbb{E}_{(N,\beta,\mu,c)\sim p_{\mathrm{data}}}\big[\log q_\phi(N \mid \beta, \mu, c)\big] \tag{11}$$

Grand canonical self-normalized importance sampling (GC-SNIS). The factorized model provides the joint log-likelihood $\log q_{\phi,\theta}(N, X) = \log q_\phi(N) + \log q_\theta(X \mid N)$, where we omit the conditioning on (β, μ, c). Given K samples $(N_i, X_i) \sim q_{\phi,\theta}$, we can estimate an observable $\mathcal{O}$ via

$$w_i = \frac{\gamma_\mu(N_i, X_i; c)}{q_{\phi,\theta}(N_i, X_i \mid \beta, \mu, c)}, \quad \widehat{\mathbb{E}}_\pi[\mathcal{O}] = \frac{\sum_{i=1}^K w_i\, \mathcal{O}(N_i, X_i)}{\sum_{i=1}^K w_i}. \tag{12}$$

The grand partition function cancels in the self-normalized ratio. Direct sampling from q_ϕ can miss particle numbers with small proposal probability but significant target weight. To address this, we use stratified sampling: we allocate n_N samples to each particle number N, set $K = \sum_N n_N$ and $\alpha_N = n_N/K$, and draw $X_{N,i} \sim q_\theta(\cdot \mid N, \beta, c)$. The stratified weight and estimator are

$$w^{\mathrm{strat}}_{N,i} = \frac{\gamma_\mu(N, X_{N,i}; c)}{\alpha_N\, q_\theta(X_{N,i} \mid N, \beta, c)}, \quad \widehat{\mathbb{E}}^{\mathrm{strat}}_\pi[\mathcal{O}] = \frac{\sum_N \sum_{i=1}^{n_N} w^{\mathrm{strat}}_{N,i}\, \mathcal{O}(N, X_{N,i})}{\sum_N \sum_{i=1}^{n_N} w^{\mathrm{strat}}_{N,i}}. \tag{13}$$

GCG samples can also initialize conventional GCMC chains. These chains preserve the target distribution while starting near equilibrium and do not require likelihood evaluations.

Architecture. We implement the joint and factorized GCG using Morph (Franke et al., 2026) with a Diffusion Transformer (DiT) backbone.

Morph baseline. We first make minimal changes to adapt Morph to the sampling task and use it as a baseline (Franke et al., 2026). We adapt Morph from molecular design to single-species particle systems by removing bond, atom-type, and charge predictions, as well as bond inputs. We disable center-of-mass removal and condition on (β, μ, c) through the existing adaptive layer normalization (AdaLN) blocks used for conditioning on the flow time.

Joint GCG model. To obtain an improved Joint GCG, we make the following changes. First, we replace the parent-based insertion mechanism with a graph-level insertion intensity and a small CFM head for insertion coordinates, which allows the model to handle empty states and spatially homogeneous systems. The CFM head is a small vector-field network attached to the shared DiT backbone. It transports each insertion position from a uniform periodic source using a short ODE

integration. Second, we represent coordinates on a periodic torus with a uniform source distribution and wrap coordinates into the simulation cell after each integration step. The coordinate readout aggregates relative offsets from neighbouring particles within a cutoff using the minimum-image convention. We choose a cutoff that is smaller than L/2, covered by a smooth envelope, which avoids the discontinuities at the box boundary and at the cutoff, which can hurt model likelihood evaluations. Third, we construct flow-matching couplings with periodic OT. The assignment cost uses minimum-image distances with Hungarian matching but without Kabsch alignment, which is invalid for a fixed periodic cell. For the periodic Lennard–Jones experiment, we replace the existing symmetry augmentations with random periodic translations and the 24 cubic rotations.

Factorized GCG model. For Factorized GCG, the conditional model $q_\theta(X \mid N, \beta, c)$ is a CNF that shares the periodic backbone of Joint GCG. Its velocity network receives (N, β, c, t) but not μ. The particle-number model q_ϕ uses a small multilayer perceptron (MLP) for r_ϕ (Section 4).

For the zeolite adsorption experiment, we continue using the Factorized GCG architecture. Since the zeolite lattice breaks the translational symmetry, we add absolute position features encoded with Fourier features on the torus. We remove translation and rotation augmentations, replace optimal-transport coupling with independent coupling, and substitute global attention biases with local message passing to improve scalability.

5 Experiments

Our experiments test whether GCG meets the three requirements identified in Section 1: (1) Can a variable-size generator sample the joint distribution over N and X of an interacting fluid? (2) Does the structured particle-number model improve transfer to chemical potentials outside the training range? (3) Can the tractable likelihood correct residual model error? Finally, (4) we test whether these properties carry over to a realistic application, methane adsorption in a zeolite.

We compare our **Joint GCG** and **Factorized GCG** against **Morph** (Franke et al., 2026) as a baseline. The models are trained and validated on GCMC reference simulations (Hatch et al., 2024; Dubbeldam et al., 2016). Datasets and hyperparameters are listed in Section A.2 and Table 1.

Lennard–Jones fluid. We first ask whether variable-size generators can sample an interacting grand canonical ensemble, and whether an existing variable-size model for molecular design suffices. We train on a periodic Lennard–Jones fluid at four chemical potentials $\beta\mu \in \{-3.75, -3.25, -3.00, -2.75\}$ and evaluate at the held-out interpolation $\beta\mu = -3.5$. The units are in terms of the particle diameter σ and well depth ϵ (Section A.2).

All three models reproduce the particle-number distribution (Fig. 2a). However, Morph places the first peak of $g(r)$ at $r \approx 1.5\sigma$ instead of 1.1σ (Fig. 2b) and consequently overestimates the potential energy (mean -1.9ϵ vs. -2.7ϵ, Fig. 2c). Both GCG variants recover $g(r)$ and the energy density $p(U)$, with Factorized GCG matching the reference most closely and Joint GCG slightly overpopulating large N ($\langle N \rangle = 10.4$ vs. 9.9). On an ideal gas without any interactions, both GCG variants recover the Poisson particle-number distribution and uniform spatial marginals, whereas Morph

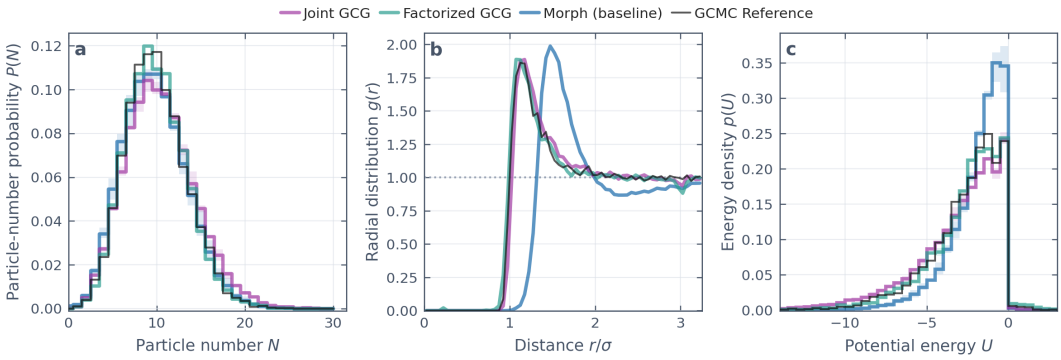

Figure 2: Lennard–Jones. Curves and bands show the mean and range over three training seeds.

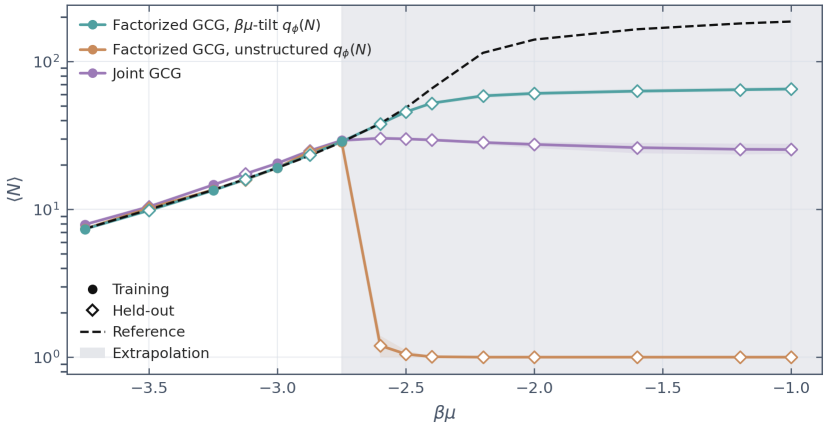

Figure 3: Chemical-potential extrapolation. Mean and ± 1 std over three training seeds.

overestimates the central density by $\sim 13\times$ (Section A.1.1). We attribute Morph's failure to its insertion relative to existing particles and non-periodic coordinates, which bias particle placement.

Chemical-potential transfer. In applications, a sampler must cover chemical potentials beyond those in its training data, e.g., to extend an adsorption isotherm to higher pressures. We ask whether building the known μ-dependence into $q_\phi(N)$ enables such transfer to unseen chemical potentials. To separate the effect of the factorization from the effect of the tilt, we compare Joint GCG with two Factorized GCG variants: one with an unstructured $q_\phi(N)$ that receives μ as input, and one with the $\beta\mu$-tilted $q_\phi(N)$ from Section 4. We evaluate the learned particle-number distributions on held-out states between the training conditions $\beta\mu \in \{-3.75, -3.25, -3.00, -2.75\}$ and on extrapolation states $\beta\mu \in \{-2.6, -2.5, -2.4\}$, where the reference $\langle N \rangle$ increases from 28.4 to 65.4.

All models reproduce the training states, but differ outside the training range (Fig. 3). On the held-out states inside the training range, the $\beta\mu$-tilted $q_\phi(N)$ has a relative error of at most 1.1%, compared to 8% for Joint GCG and 6% for the unstructured $q_\phi(N)$. The tilted model stays within 0.6% at $\beta\mu = -2.6$ and 6% at $\beta\mu = -2.5$, while Joint GCG errs by 21% and 38%. The unstructured $q_\phi(N)$ collapses to $\langle N \rangle \approx 1$ at all extrapolation states. At $\beta\mu = -2.4$, all models fail. The tilted model underestimates $\langle N \rangle$ by 20% and Joint GCG by 55%. Thus, the tilt extends the accurate range, but nevertheless accumulates error from extrapolating $\log Z_N$ to unseen particle numbers. Section A.1.2 shows that adding a single-particle reference to $q_\phi(N)$ further improves extrapolation.

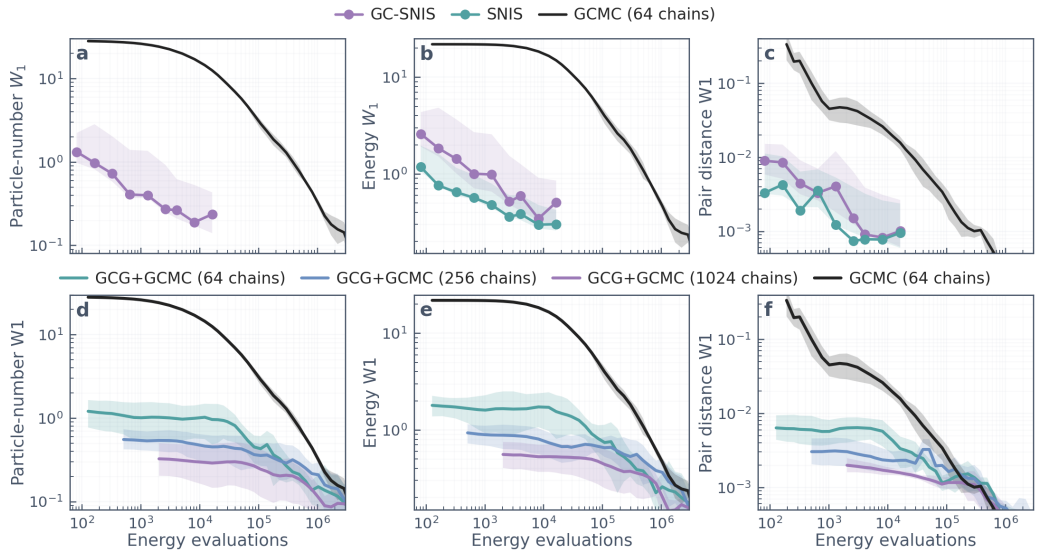

Figure 4: Inference-time correction of Factorized GCG on LJ with SNIS (top) and GCMC (bottom). Bands show 95% bootstrap intervals (200 resamples for SNIS; 2,000 for GCMC over five seeds).

Inference-time correction. Next, we ask whether the tractable likelihood of a trained Factorized GCG can remove residual model bias. We treat training as a one-time cost that is amortized over chemical potentials and systems in the future, and therefore compare only inference-time estimators. We draw 400 configurations for each particle number from $N = 10$ to 50 (16,400 total), covering 99.7% of the model's particle-number mass. Conditional SNIS reweights configurations at each fixed N, whereas GC-SNIS additionally corrects the particle-number probabilities (Eq. (13)).

We measure the error of each scalar observable by its 1-Wasserstein distance W_1 to a long GCMC reference (Section A.2.4). With 1.6×10^4 target-energy evaluations, conditional SNIS halves the energy W_1 (0.6 to 0.3) and lowers the pair-distance W_1 from 1.6×10^{-3} to 9×10^{-4}. GCMC needs $92\times$ and $25\times$ more evaluations to match these errors (Fig. 4, top). However, GC-SNIS increases the particle-number W_1 from 0.08 to 0.2. The learned $q_\phi(N)$ is already accurate on this system and the finite-sample error of GC-SNIS exceeds the model's prediction error. Both estimators retain 12–13% of the configurations as effective samples (Section A.1.3). In Section A.1.4 we show that estimating the ratios Z_{N+1}/Z_N with the Bennett acceptance ratio (BAR) helps alleviate this error.

In order to avoid likelihood evaluations, including Joint GCG, generated configurations can instead be used to initialize GCMC chains. We test GCMC initialization with the same samples from the Factorized GCG model as for SNIS. Seeded chains start with a $23\times$ lower particle-number W_1 than cold-start GCMC after two trial moves per chain (1.2 vs. 28 for 64 chains). Their particle-number and energy errors remain below the cold-start baseline until $\sim 4 \times 10^6$ evaluations, while pair-distance errors meet at $\sim 2 \times 10^5$.

Methane adsorption in a zeolite. Finally, we test whether GCG carries over to a realistic adsorption problem. We apply Factorized GCG to methane adsorption in a rigid all-silica MFI zeolite, a benchmark system for gas separation. The crystalline host confines methane to a network of intersecting channels, with up to ~130 adsorbed molecules per cell. We train on seven pressures from 0.1 to 70 bar and evaluate at held-out pressures up to 150 bar, twice the largest training pressure. We show the mean and ± 1 std for three training seeds.

The generated adsorption isotherm agrees with GCMC across all held-out pressures, including the extrapolation regime, with the mean error in particle number across three training seeds being at most 0.5 molecules ($< 0.4\%$) (Fig. 5d). The particle-number distributions also match GCMC at 50 and 100 bar (Fig. 5e). At 100 bar, generated configurations reproduce the methane distribution across the channels (Fig. 5b,c) and the methane pair-distance distribution.

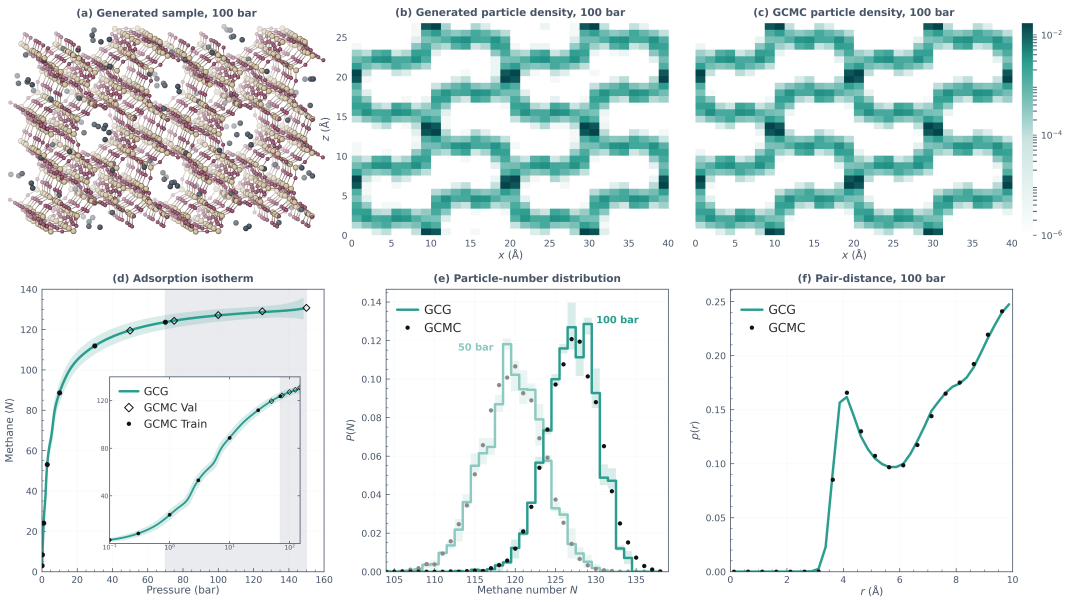

Figure 5: Methane adsorption in MFI zeolite for a Factorized GCG model, compared to GCMC.

6 CONCLUSION

We introduced Grand Canonical Generators, which extend BGs from the canonical to the grand canonical ensemble. Joint GCG samples particle number and configuration within a single variable-size process. Factorized GCG imposes additional structure, which leads to favourable results in our experiments. The Factorized GCG combines a particle-number model with any conditional BG, enabling a tractable joint likelihood. By encoding the known chemical-potential dependence analytically, the particle-number model only learns canonical partition function ratios and improves transfer to chemical potentials beyond the training range. Given a trained model, the tractable likelihood enables grand canonical importance sampling that corrects residual model error, and generated samples accelerate GCMC even for models without likelihoods. On methane adsorption in a zeolite, a Factorized GCG accurately reproduces the adsorption isotherm.

Looking forward, the factorized GCG is agnostic to the conditional BG, so advances in canonical generators carry over. Autoregressive architectures are a natural fit for the grand canonical ensemble, since they generate variable-size systems particle by particle while providing exact likelihoods, and could unify the joint and factorized variants. Further directions include simultaneous training on multiple hosts and particles, training on systems with flexible hosts that follow the canonical ensemble, and reducing the reliance on GCMC training data through energy-based training objectives.

Limitations. The GCG models presented here require equilibrium samples for training, which have to be collected from GCMC simulations. Amortizing model training across systems remains an open challenge. We treat methane as a point-like particle and the zeolite crystal as rigid. While these are standard modelling assumptions in the literature, extending the approach to polar adsorbates such as water, or to flexible frameworks such as proteins, would require sampling over additional degrees of freedom. Finally, SNIS incurs the cost of evaluating the target log-likelihood for every generated configuration. Likelihood flow maps could reduce the cost of likelihood evaluation and are a promising direction for future work. Correcting particle-number probabilities with GC-SNIS remains limited by weight variance.

AI USE STATEMENT

In this work, we used generative artificial intelligence (AI) tools to run the computational protocols for generating the datasets, create figures, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, and interpret results. Assisting with translation, formulating mathematical claims, providing critical ingredients for proving mathematical claims, or assisting in the writing of proofs, are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

David J. Adams. Grand canonical ensemble monte carlo for a Lennard-Jones fluid. *Molecular Physics*, 29(1):307–311, 1975. doi: 10.1080/00268977500100221.

Marin Biloš and Stephan Günnemann. Scalable normalizing flows for permutation invariant densities. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 957–967. PMLR, 2021. URL https://proceedings.mlr.press/v139/bilos21a.html.

Denis Blessing, Mouyang Cheng, Maximilian Schebek, Jutta Rogal, Mingda Li, Carles Domingo-Enrich, and Yuanqi Du. JANUS: A multi-modal foundation neural sampler for disordered materials. *arXiv preprint arXiv:2608.19116*, 2026. doi: 10.48550/arXiv.2608.19116.

Andreas Burger, Luca Thiede, Nikolaj Rønne, Varinia Bernales, Nandita Vijaykumar, Tejs Vegge, Arghya Bhowmik, and Alán Aspuru-Guzik. HIP: Hessian interatomic potentials without derivatives. *arXiv preprint arXiv:2509.21624*, 2025. doi: 10.48550/arXiv.2509.21624.

Andreas Burger, Luca A. Thiede, Abdulrahman Aldossary, Jorge A. Campos-Gonzalez-Angulo, Alex Zook, Jérôme F. Gonthier, and Alán Aspuru-Guzik. MōLe-Λ: Learning the coupled-cluster response state for energies, gradients, and properties. In *AI4Physics: An ICML 2026 Workshop on AI for Physics*, Seoul, South Korea, 2026. URL https://icml.cc/virtual/2026/72615. arXiv:2605.29622.

Andrew Campbell, William Harvey, Christian Weilbach, Valentin De Bortoli, Tom Rainforth, and Arnaud Doucet. Trans-dimensional generative modeling via jump diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42217–42257, 2023. doi: 10.52202/075280-1830.

Austin H. Cheng, Chong Sun, and Alán Aspuru-Guzik. Scalable autoregressive 3D molecule generation. *arXiv preprint arXiv:2505.13791*, 2025. doi: 10.48550/arXiv.2505.13791.

Mads-Peter Verner Christiansen and Bjørk Hammer. Gradient-based grand canonical optimization enabled by graph neural networks with fractional atomic existence. *Machine Learning: Science and Technology*, 6(4):045049, 2025. doi: 10.1088/2632-2153/ae1bf6.

James Damewood, Daniel Schwalbe-Koda, and Rafael Gómez-Bombarelli. Sampling lattices in semi-grand canonical ensemble with autoregressive machine learning. *npj Computational Materials*, 8(1):61, 2022. doi: 10.1038/s41524-022-00736-4.

Manuel Dibak, Leon Klein, Andreas Krämer, and Frank Noé. Temperature steerable flows and Boltzmann generators. *Physical Review Research*, 4(4):L042005, 2022. doi: 10.1103/PhysRevResearch.4.L042005.

Xiaochen Du, Juno Nam, Sulin Liu, and Rafael Gómez-Bombarelli. Scaling autoregressive models for lattice thermodynamics. *arXiv preprint arXiv:2603.14695*, 2026. doi: 10.48550/arXiv.2603.14695.

David Dubbeldam, Sofía Calero, Donald E. Ellis, and Randall Q. Snurr. RASPA: Molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Molecular Simulation*, 42(2):81–101, 2016. doi: 10.1080/08927022.2015.1010082.

Eike S. Eberhard, Luca A. Thiede, Abdulrahman Aldossary, Andreas Burger, Nicholas Gao, Vignesh Bhethanabotla, Alán Aspuru-Guzik, and Stephan Günnemann. Derivative informed learning of exchange-correlation functionals. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2606.04279.

Alan M. Ferrenberg and Robert H. Swendsen. Optimized Monte Carlo data analysis. *Physical Review Letters*, 63(12):1195–1198, 1989. doi: 10.1103/PhysRevLett.63.1195.

Malte Franke, Stefan P. Schmid, Žarko Ivković, Kjell Jorner, and Andreas Krause. Generative molecular morphing for flexible-size design via unbalanced optimal transport. *arXiv preprint arXiv:2606.07239*, 2026. doi: 10.48550/arXiv.2606.07239.

Louis Grenioux, Leonardo Galliano, Ludovic Berthier, Giulio Biroli, and Marylou Gabrié. Boltzmann generators for amorphous particle systems. *arXiv preprint arXiv:2512.16607*, 2025. doi: 10.48550/arXiv.2512.16607.

Harold W. Hatch, Daniel W. Siderius, and Vincent K. Shen. Monte carlo molecular simulations with FEASST version 0.25.1. *The Journal of Chemical Physics*, 161(9):092501, 2024. doi: 10.1063/5.0224283.

Emil Hoffmann, Maximilian Schebek, Leon Klein, Frank Noé, and Jutta Rogal. Boltzmann generators for condensed matter via Riemannian flow matching. In *AI4Mat Workshop at the International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=FTIxxfPLSG. arXiv:2602.18482.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary markov processes. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/819aaee144cb40e887a4aa9e781b1547-Abstract-Conference.html.

William L. Jorgensen and Ch. Ravimohan. Monte carlo simulation of differences in free energies of hydration. *The Journal of Chemical Physics*, 83(6):3050–3054, 1985. doi: 10.1063/1.449208.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/bc827452450356f9f558f4e4568d553b-Abstract.html.

Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992. doi: 10.1002/jcc.540130812.

Suemin Lee, Ruiyu Wang, Lukas Herron, and Pratyush Tiwary. Predicting phase transitions across temperature, pressure, and chemical potential using exponentially tilted thermodynamic maps. *Nature Communications*, 2026. doi: 10.1038/s41467-026-77612-y.

Bo Lei, Enze Chen, Hyuna Kwon, Tim Hsu, Babak Sadigh, Vincenzo Lordi, Timofey Frolov, and Fei Zhou. Grand canonical generative diffusion model for crystalline phases and grain boundaries. *arXiv preprint arXiv:2408.15601*, 2024. doi: 10.48550/arXiv.2408.15601.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Bálint Máté, François Fleuret, and Tristan Bereau. Neural thermodynamic integration: Free energies from energy-based diffusion models. *The Journal of Physical Chemistry Letters*, 15(45):11395–11404, 2024. doi: 10.1021/acs.jpclett.4c01958.

Oliver J. Melling, Marley L. Samways, Yunhui Ge, David L. Mobley, and Jonathan W. Essex. Enhanced grand canonical sampling of occluded water sites using nonequilibrium candidate monte carlo. *Journal of Chemical Theory and Computation*, 19(3):1050–1062, 2023. doi: 10.1021/acs.jctc.2c00823.

Andrew R. Millward and Omar M. Yaghi. Metal–organic frameworks with exceptionally high capacity for storage of carbon dioxide at room temperature. *Journal of the American Chemical Society*, 127(51):17998–17999, 2005. doi: 10.1021/ja0570032.

Kim A. Nicoli, Shinichi Nakajima, Nils Strodthoff, Wojciech Samek, Klaus-Robert Müller, and Pan Kessel. Asymptotically unbiased estimation of physical observables with neural samplers. *Physical Review E*, 101(2):023304, 2020. doi: 10.1103/PhysRevE.101.023304.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019. doi: 10.1126/science.aaw1147.

RuiKang OuYang, Hanlin Yu, Xinyue Ai, Yutong He, Nicholas M. Boffi, Pradeep Ravikumar, José Miguel Hernández-Lobato, Max Simchowitz, Benjamin Kurt Miller, and Omar Chehab. Few-step Boltzmann generators via scalable likelihood flow maps. *arXiv preprint arXiv:2606.29110*, 2026. doi: 10.48550/arXiv.2606.29110.

Kashif Rasul, Ingmar Schuster, Roland Vollgraf, and Urs Bergmann. Set flow: A permutation invariant normalizing flow. *arXiv preprint arXiv:1909.02775*, 2019. doi: 10.48550/arXiv.1909.02775.

Danyal Rehman, Charlie B. Tan, Yoshua Bengio, Avishek Joey Bose, and Alexander Tong. Autoregressive Boltzmann generators. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2606.27361. Spotlight.

Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, 2008. doi: 10.1063/1.2978177.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Nigel B. Wilding. Critical-point and coexistence-curve properties of the Lennard-Jones fluid: A finite-size scaling study. *Physical Review E*, 52(1):602–611, 1995. doi: 10.1103/PhysRevE.52.602.

Peter Wirnsberger, George Papamakarios, Borja Ibarz, Sébastien Racanière, Andrew J. Ballard, Alexander Pritzel, and Charles Blundell. Normalizing flows for atomic solids. *Machine Learning: Science and Technology*, 3(2):025009, 2022. doi: 10.1088/2632-2153/ac6b16.

A Appendix

A.1 Additional results

A.1.1 Ideal gas

Previous variable-size generative models were built for molecular design (Franke et al., 2026), and may incorporate architectural choices that are less suitable for fluids or gases. To study any spatial bias separate from interaction effects, we consider the ideal gas, for which the solution is known analytically.

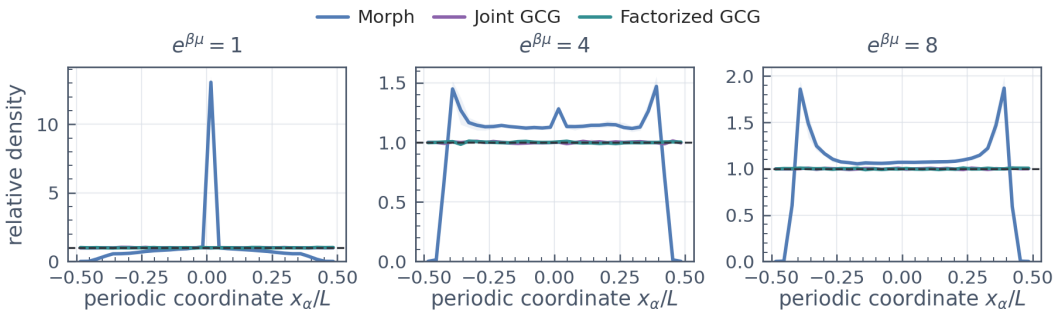

Figure 6: One-coordinate marginals for the periodic ideal gas. Lines are the mean density over the three coordinate axes and bands are their range. Dashed is the analytic uniform density. Morph develops a sharp central peak at $e^{\beta\mu} = 1$. Joint GCG and Factorized GCG remain close to the true marginal at all activities.

Setup. For a single-species ideal gas in a periodic box of unit volume ($V = 1$, $U_N = 0$, $\Lambda = 1$), the grand canonical distribution is

$$N \sim \mathrm{Poisson}\big(e^{\beta\mu}\big), \qquad \mathbf{x}_i \mid N \overset{\mathrm{iid}}{\sim} \mathrm{Uniform}(\mathcal{V}), \tag{14}$$

where $\mathbf{x}_i$ are the particle positions in $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and $\mathcal{V}$ is the periodic box. Thus every one-coordinate marginal is uniform and $g(r) = 1$. We train over the continuous range $\beta\mu \in [0, \log 16]$ and evaluate at activities $e^{\beta\mu} \in \{1, 4, 8\}$ with 65,536 configurations each.

Coordinate total variation. We divide each periodic coordinate interval $[-L/2, L/2]$ into $B_x = 32$ equal-width bins. Denoting by $\widehat{p}_{\alpha b}$ the normalized particle histogram for axis $\alpha \in \{1, 2, 3\}$ and bin b, we report the mean per-axis total variation (TV) distance from the analytic uniform marginal:

$$\mathrm{TV}_{\mathrm{coord}} = \frac{1}{3}\sum_{\alpha=1}^{3}\frac{1}{2}\sum_{b=1}^{B_x}\left|\widehat{p}_{\alpha b} - \frac{1}{B_x}\right|. \tag{15}$$

Radial-distribution error. We use $B_r = 25$ equal-width radial bins on $0 < r \le L/2$ with minimum-image pair distances. For radial edges r_b, the shell volume is $\Delta V_b = \frac{4\pi}{3}(r_{b+1}^3 - r_b^3)$. If C_b is the observed number of pairs in shell b, the ideal-gas normalization, the radial distribution function estimator, and its mean absolute error (MAE) are

$$E_b = \sum_i \frac{N_i(N_i-1)}{2}\frac{\Delta V_b}{V}, \qquad \widehat{g}_b = \frac{C_b}{E_b}, \qquad \mathrm{MAE}_g = \frac{1}{B_r}\sum_{b=1}^{B_r}|\widehat{g}_b - 1|\,. \tag{16}$$

The uniform-sample baseline uses the same 65,536 configurations to measure finite-sample error, which is largest for the low-occupancy inner radial shells.

Spatial marginals. Figure 6 compares one-coordinate marginals across methods. Morph develops strong spatial inhomogeneity. At activity one, its central density exceeds the target by a factor of $\sim 13\times$. The bias persists at activities four and eight, where Morph places no particles in the outer bins near the box faces and overpopulates the interior. This is consistent with its non-periodic coordinates and Gaussian source. Joint GCG and Factorized GCG both recover nearly uniform marginals at all three activities.

Quantitative comparison. Figure 7 reports the one-dimensional Wasserstein distance from the Poisson distribution, mean per-axis coordinate-histogram TV, and MAE of $g(r)$ from one. At activity four, the particle-number W_1 errors are 1.15, 0.034, and 0.005, and the coordinate TV errors are 0.143, 0.0036, and 0.0041 for Morph, Joint GCG, and Factorized GCG, respectively. Joint GCG thus reduces all three errors of Morph, including the $g(r)$ MAE, from 0.509 to 0.216. However, Joint GCG still introduces substantial pair correlations, compared with 0.052 for Factorized GCG and a finite-sample noise of 0.021. Factorized GCG gives the best overall agreement, while Joint GCG's pair-structure error shows that accurate one-particle marginals alone are not sufficient.

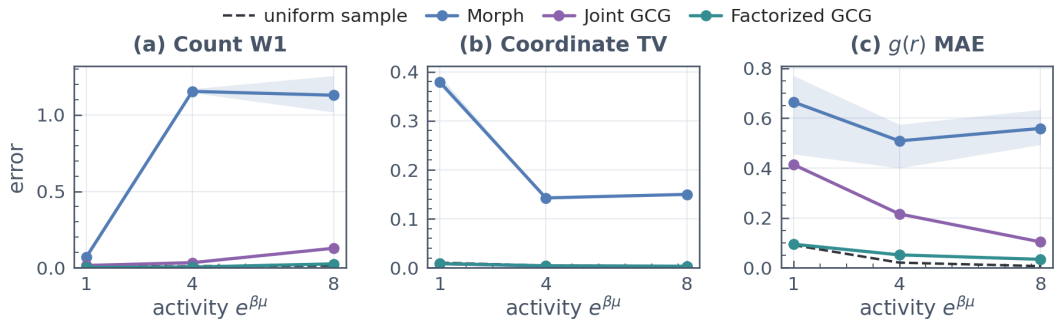

Figure 7: Ideal-gas errors across activity. Dashed: finite-sample error from independent uniform configurations.

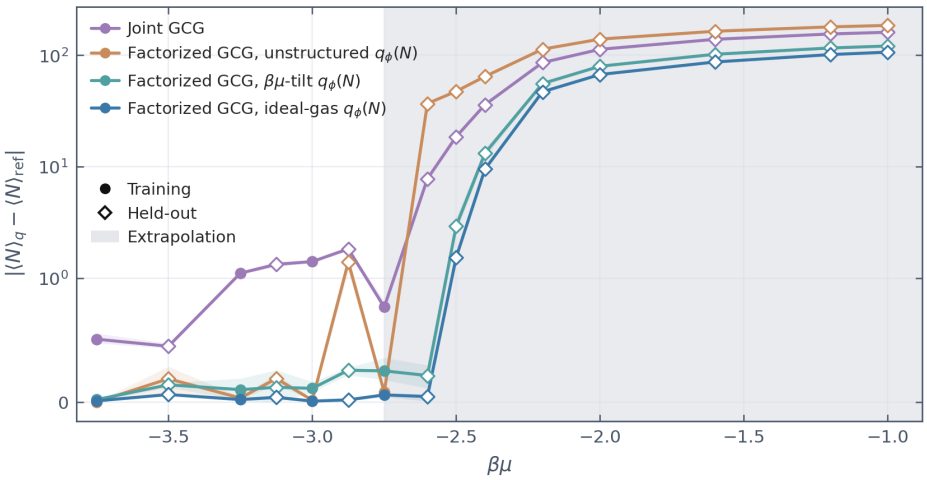

Figure 8: Particle-number transfer to unseen chemical potentials.

A.1.2 Particle-number model parameterization

Structured particle-number distribution. We seek a particle-number model that transfers to unseen chemical potentials and, ideally, to unseen systems. We start from the $\beta\mu$-tilted model in Section 4, where r_ϕ represents the full $\log Z_N$. While this applies the exact μ-dependence, r_ϕ must also learn the ideal-gas contribution $N \log V$, which grows linearly in N. At initialization, $r_\phi = 0$, so the model starts far from any physical distribution and must extrapolate this term to unseen particle numbers. Fortunately, the ideal gas has the analytic solution $Z_N = q_1^N$, which gives a Poisson particle-number distribution (Section A.1.1). Writing $\log Z_N = N \log q_1 + \log(Z_N/q_1^N)$ separates this one-particle term from the interaction contribution. Thus, we generalize the $\beta\mu$-tilted model with an ideal-gas reference:

$$q_\phi(N \mid \beta, \mu, c) = \frac{\exp\big[\beta\mu N + N \log q_1(\beta, c) - \log(N!) + r_\phi(N, \beta, c)\big]}{\sum_{M=0}^{N_{\max}} \exp\big[\beta\mu M + M \log q_1(\beta, c) - \log(M!) + r_\phi(M, \beta, c)\big]} = \mathrm{softmax}_N\big[\beta\mu N + N \log q_1(\beta, c) - \log(N!) + r_\phi(N, \beta, c)\big], \tag{17}$$

which differs from Section 4 only by the reference term $N \log q_1(\beta, c)$, and reduces to it for $q_1 = 1$. Here, $q_1(\beta, c)$ is a one-particle reference partition function, typically in analytic form, that we are free to choose. For heterogeneous systems, $q_1(\beta, c)$ represents an accessible volume or a host-specific partition function. For the homogeneous Lennard–Jones system, we set $q_1 = V$ in reduced coordinates.

Learned residual. The residual r_ϕ represents interactions and other deviations from the chosen one-particle reference. We parameterize it with an MLP that receives the particle number and returns one scalar. To transfer across systems, an encoder may additionally condition r_ϕ on e.g. host and adsorbate features. As in Section 4, we absorb Λ^3 into the activity $e^{\beta\mu}$. We fix the additive gauge by enforcing $r_\phi(0, \beta, c) = 0$ and initialize the final layer to zero, so the initial distribution coincides with the analytic reference. Interactions are then learned over the course of training.

Results. Figure 8 extends the main-text figure with a particle-number model that uses the analytic ideal gas as a reference. Up to $\beta\mu = -2.4$, the more structured particle-number models have smaller errors than the less structured models. The ideal-gas reference reduces the relative error from 6% to 3% at $\beta\mu = -2.5$ and from 20% to 15% at $\beta\mu = -2.4$, compared to the $\beta\mu$-tilted model. For $\beta\mu \geq -2.2$, no model is accurate.

A.1.3 Effective sample size

The ESS of SNIS depends on the overlap between the proposal and the target. If the proposal matches the target, all weights are equal and every configuration contributes fully. If the overlap is small, a few configurations carry most of the weight. We use the Kish effective sample size (Kish,

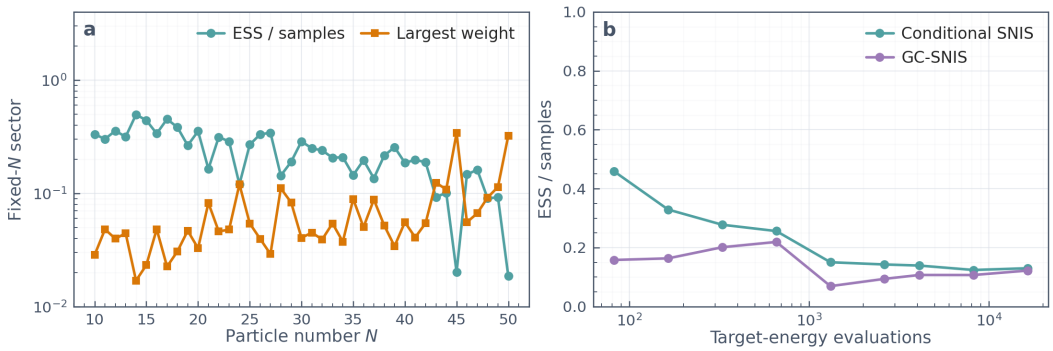

Figure 9: Normalized effective sample size ESS/M of the stratified pool in Fig. 4. (a) Per fixed-N sector at 400 configurations, with the largest normalized weight (dashed). (b) Pooled over all sectors for conditional SNIS and GC-SNIS as a function of target-energy evaluations.

1965) of the normalized weights of each estimator,

$$\mathrm{ESS} = \frac{\left(\sum_{i=1}^{M} w_i\right)^2}{\sum_{i=1}^{M} w_i^2}, \qquad w_i = \frac{\gamma_\mu(N_i, X_i; c)}{q_{\mathrm{prop}}(N_i, X_i \mid \beta, \mu, c)}, \tag{18}$$

which counts how many of the M configurations effectively contribute to the estimate. Here, q_{prop} is the proposal density, including the stratification factor when applicable. We report the normalized quantity ESS/M, which is 1 for a perfect proposal and $1/M$ if a single configuration carries all the weight. For each sector, w_i are the fixed-N weights of conditional SNIS; for the pooled values, w_i are the weights of conditional SNIS or GC-SNIS (Eq. (13)).

Figure 9 reports this quantity for the stratified pool of Fig. 4. Within a sector, ESS/M ranges from ~ 0.3 to ~ 0.5 for $N \leq 20$ and from ~ 0.1 to ~ 0.2 for $N \geq 35$. Only two sectors, $N = 45$ and $N = 50$, fall to ~ 0.02, where one configuration carries about one third of the sector weight. These sectors have model probabilities of only 0.5% and 0.1%, so their effect on pooled estimates is small. Pooled over all sectors, ESS/M is 0.13 for conditional SNIS and 0.12 for GC-SNIS at 16,400 configurations. The decrease with N is consistent with the growing configurational dimension $3N$. We leave allocation schemes that adapt to the per-sector ESS for future work.

A.1.4 Bennett acceptance ratio

Conditional and grand canonical correction. Conditional SNIS reweights configurations at each fixed N, but leaves the particle-number probabilities unchanged. GC-SNIS uses one global weight and corrects these probabilities, but its weight variance can be large.

BAR estimates the ratio Z_{N+1}/Z_N of canonical partition functions from Boltzmann samples at fixed particle numbers N and $N+1$. These ratios determine the relative probabilities of observing N and $N+1$ particles. BAR combines insertion and deletion estimates, which makes the estimate more robust when the two distributions have limited overlap compared to only using insertion or deletion estimates. Like for GC-SNIS, we stratify the configuration pool by particle number and allocate a fixed number of configurations to each value of N.

With the canonical partition function Z_N from Eq. (3), we define

$$R_N = \frac{Z_{N+1}}{V Z_N}, \qquad \Delta F_N = -\log R_N.$$

We estimate ΔF_N from the weighted BAR equation

$$\sum_i a_i \, s\big(\Delta u_i^+ - \Delta F_N\big) = \sum_j b_j \, s\big(-\Delta u_j^- + \Delta F_N\big), \qquad s(x) = \frac{1}{1+\exp(x)}, \tag{19}$$

where $\Delta u_i^+ = \beta[U_{N+1}(X_i \cup \{\mathbf{x}\}) - U_N(X_i)]$ is the reduced energy change for inserting a uniform trial particle $\mathbf{x}$ into a sample X_i with N particles, and $\Delta u_j^- = \beta[U_{N+1}(X_j) - U_N(X_j \setminus \{\mathbf{x}_k\})]$ is

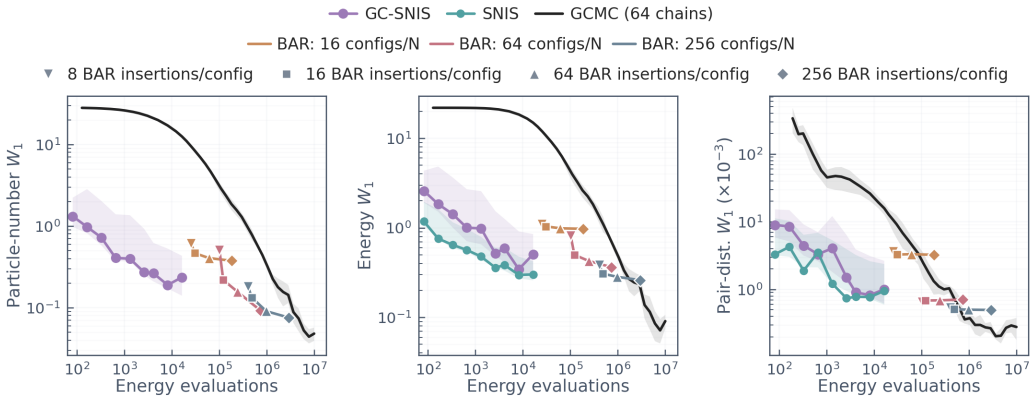

Figure 10: Inference-time correction with conditional SNIS, GC-SNIS, and BAR between N and $N+1$ particles. BAR combines insertion and deletion energy changes to reconstruct particle-number probabilities.

the change for removing particle $\mathbf{x}_k$ from a sample X_j with $N+1$ particles. The weights a_i and b_j are the SNIS weights of these samples, capped at the 99th percentile for their particle number (Eq. (21)). The ratio R_N then gives the relative probability of $N+1$ and N particles:

$$\frac{\pi(N+1 \mid \beta, \mu, c)}{\pi(N \mid \beta, \mu, c)} = \frac{\exp(\beta\mu)V}{N+1} R_N. \tag{20}$$

Thus, conditional SNIS corrects the configurations at each fixed N, whereas BAR corrects the relative probabilities of different N.

Weight capping. For the BAR curves, we cap the conditional SNIS weights within each particle number at their 99th percentile and renormalize them:

$$\widetilde{w}_i^{(N)} = \frac{\min\left(w_i^{(N)}, Q_{0.99}(w^{(N)})\right)}{\sum_{j:N_j=N} \min\left(w_j^{(N)}, Q_{0.99}(w^{(N)})\right)}. \tag{21}$$

We apply the cap before estimating the adjacent particle-number ratios.

Results. Figure 10 compares particle-number, energy, and pair-distance W_1 errors against target-energy evaluations. Lower values indicate better agreement. Conditional SNIS and GC-SNIS reach low errors through direct reweighting, whereas GCMC improves with the chain budget. BAR also improves with more configurations per particle number and insertion trials, but its performance depends on the proposal budget.

A.2 Datasets

A.2.1 2D repelling particles in a trap

For Fig. 1, we consider N charged particles at positions $\mathbf{x}_i = (x_i, y_i)$ in an anisotropic harmonic trap, that repel each other via Coulomb interaction

$$U_N = \frac{1}{2}\sum_{i=1}^{N}\left(x_i^2 + 16\, y_i^2\right) + \sum_{i<j} \frac{1}{\sqrt{\|\mathbf{x}_i - \mathbf{x}_j\|^2 + \delta^2}}, \tag{22}$$

where the softening length $\delta = 0.15$ regularizes the interaction at short range. The anisotropic confinement is four times stronger along y than along x, favoring elongated arrangements. For a single particle, the trap pulls it toward the center. For two or more particles, the repulsion separates the particles along the x axis. The plotted flow trajectories represent the generator's transport from its prior.

A.2.2 Periodic Lennard–Jones fluid

We consider a single-species Lennard–Jones fluid in a cubic periodic box of side $L = 6.5\sigma$ with particle diameter $\sigma = 1$, well depth $\epsilon = 1$, reduced temperature $T^* = 1.5$ ($\beta^* = 2/3$), cutoff $r_c = 3\sigma$ (no energy shift, with long-range correction), and thermal wavelength $\Lambda = 1$.

We sample configurations using GCMC in FEASST v0.25.19 (Hatch et al., 2024), with Metropolis translation, insertion, and deletion moves. Each trial selects a particle translation or an insertion/deletion move with weights 1 and 2, respectively. After 2,000,000 equilibration trials, we record configurations every 5,000 trials at four training chemical potentials $\beta\mu \in \{-3.75, -3.25, -3.00, -2.75\}$.

At each chemical potential we generate five independent chains of 20,000 configurations for training and one chain of 4,096 configurations for validation, all with independent pseudorandom seeds. We generate four independent GCMC chains at the held-out chemical potential $\beta\mu = -3.5$, each containing 8,192 configurations after the same equilibration and recording procedure, for a total of 32,768 reference configurations. Coordinates are wrapped into the periodic cell and scaled by $L/\sqrt{12}$. We retain particle number, coordinates, total energy, and Lennard–Jones pair energy for every configuration.

A.2.3 Methane adsorption in MFI zeolite

We study methane adsorption in a rigid all-silica MFI zeolite. The framework is the $2 \times 2 \times 2$ supercell of the bundled `MFI_SI` structure in RASPA2 (Dubbeldam et al., 2016), with box lengths $(40.044, 39.798, 26.766)$ Å. Methane uses the point-particle treatment from RASPA2 with the bundled `ExampleZeolitesForceField`. We use this classical force field for comparability with prior GCMC studies. Learned models that provide energies and gradients at higher levels of theory could be used in the future (Eberhard et al., 2026; Burger et al., 2026). All simulations run GCMC at $T = 298$ K with translation, reinsertion, and swap move probabilities of 0.5, 0.5, and 1.0, respectively.

At each pressure, we run 5,000 initialization cycles followed by 250,000 production cycles, writing a methane configuration and the particle-number and energy histograms every 500 cycles. We wrap all coordinates into the simulation cell before they enter the models. We train on $\{0.1, 0.3, 1, 3, 10, 30, 70\}$ bar, and test at $\{50, 75, 100, 125, 150\}$ bar. The highest pressure 150 bar contains up to 140 methane molecules.

A.2.4 Evaluation metrics

For a scalar observable x, such as the particle number N, the potential energy U, or a pair distance r_{ij}, we use the 1-Wasserstein distance (Villani, 2009) between the model distribution q and the GCMC reference p:

$$W_1(p, q) = \int_{-\infty}^{\infty} \left| F_p(x) - F_q(x) \right| \mathrm{d}x, \tag{23}$$

where F_p and F_q are the empirical cumulative distribution functions; for SNIS, F_q uses the normalized importance weights. W_1 measures the average distance that probability mass must move to turn q into p. We report W_1 because it has the units of the observable, bounds the error in its mean, $|\langle x \rangle_p - \langle x \rangle_q| \leq W_1(p, q)$, and is less prone to outliers than W_2.

A.3 Model and Hyperparameters

We train each model for 10 to 20 graphics processing unit (GPU) hours on an NVIDIA A100.

Table 1: Model hyperparameters for the Lennard–Jones (LJ) fluid and the MFI zeolite.

	Morph	Joint GCG	Factorized GCG	Factorized GCG
System	LJ	LJ	LJ	Zeolite
Temperature	1.5	1.5	1.5	298 K
Box	6.5^3	6.5^3	6.5^3	$40.04\times39.80\times26.77$ Å
Batch	1024	1024	1024	128
Steps	$60k$	$60k$	$60k$	$200k$
Backbone	8×512	8×512	8×512	8×768
Coord. features	none	none	none	Torus Fourier
Attention bias	yes	yes	yes	no
N support	n.a.	n.a.	$0-160$	$0-160$
N model	node head	graph intensity	MLP $+\log(N!)$	MLP $+\log(N!)$
Optimizer	Muon/AdamW	Muon/AdamW	Muon/AdamW	Muon/AdamW
LR	$2e{-3}/1e{-2}$	$2.5e{-4}/1.2e{-3}$	$1e{-3}/5e{-3}$	$7e{-4}/3.5e{-3}$
EMA	$0.999/5k$	$0.999/5k$	$0.999/5k$	disabled
Sample steps	200	200	200	200