Alignment Matters Inside and Out in Equivariant Graph Flow Matching

Anonymous authors
Paper under double-blind review

Abstract

Graphs are invariant under node permutations, motivating permutation-equivariant architectures in generative models. In flow matching, however, symmetry also affects the source–target coupling: one must choose both *which graphs to pair* and *which node representatives to align*. We study these two forms of alignment, termed outer and inner alignment, and their effect on equivariant graph flow matching. We connect inner alignment to transport on the graph quotient space, whose Euclidean quotient metric coincides with Gromov–Monge distance, and show that quotient couplings admit aligned representative lifts while symmetrization yields equivariant flow-matching minimizers, including for categorical endpoints. In practice, we compare random augmentation, permutation-blind minibatch optimal transport, approximate Gromov–Wasserstein alignment, and combinations of inner and outer alignment. Across continuous graph and molecular generation, alignment can substantially simplify trajectories and improve few-step generation, while its benefits depend on the alignment and computational budget. Our results highlight that effective graph flow matching benefits from alignment both inside and across graph pairs.

1 Introduction

Flow matching learns a velocity field that transports a tractable source law to data using simulation-free regression and ordinary differential equation (ODE) sampling (Albergo and Vanden-Eijnden, 2023; Lipman et al., 2023; Liu et al., 2023). Its conditional paths are determined by a source–target coupling. Although all such couplings have the same marginals at start and end time, transport-informed choices can produce shorter, straighter trajectories and simplify numerical integration (Chemseddine et al., 2025; Tong et al., 2023; Pooladian et al., 2023). The correct notion of transport is less obvious when data are defined only up to symmetry (Klein et al., 2023; Köhler et al., 2020).

For graphs the symmetry is node relabelling. Here, we encode edge channels and diagonal node features of an N-node graph by $E \in \mathbb{R}^{N\times N\times C}$. Because a graph has no distinguished vertex order, a permutation $\sigma \in \mathfrak{S}_N$ acts simultaneously on both node indices according to

$$\big(\rho_N^{\mathrm{graph}}(\sigma)E\big)_{ijc} \coloneqq E_{\sigma^{-1}(i)\sigma^{-1}(j)c}, \qquad 1 \le i, j \le N.$$

The unordered graph is the orbit

$$[E] \in Q \coloneqq \mathbb{R}^{N^2C}/G_N^{\mathrm{graph}}, \qquad G_N^{\mathrm{graph}} = \rho_N^{\mathrm{graph}}(\mathfrak{S}_N) \subset O(N^2C).$$

Equivariant architectures employed in graph generative models respect this symmetry (Eijkelboom et al., 2024; Jo et al., 2022; Vignac et al., 2023), but many models do not by themselves choose which labelled representatives to connect during flow matching. Each conditional bridge from E to an unordered target $[F]$ may select some $\sigma \cdot F$. Minimizing its length is a *Gromov–Monge* node-alignment problem, the hard-assignment counterpart of Gromov–Wasserstein transport (Mémoli, 2011; Mémoli and Needham, 2024).

We connect this representative choice to equivariant flow matching on general quotients $\mathbb{R}^D/G$ (Klein et al., 2023; Köhler et al., 2020; Song et al., 2023). Our theory shows that optimal quotient couplings lift without extra cost and that symmetrizing a coupling between chosen representatives yields an equivariant minimizer of the flow-matching objective, including for

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.

categorical endpoint prediction. For graphs, this yields an inner node-alignment problem corresponding to Gromov–Monge distance. We approximate it using Gromov–Wasserstein solvers and cheaper alignment heuristics (Mémoli, 2011; Bauer et al., 2025; Piening and Beinert, 2025), and separately study outer minibatch assignment between graphs. This changes only the source–target coupling, leaving the flow-matching setup unchanged.

Relation to prior work Graph generators commonly enforce node-relabeling symmetry through equivariant architectures, including molecular and discrete diffusion (Hoogeboom et al., 2022; Jo et al., 2022; Vignac et al., 2023) and variational flow matching (Eijkelboom et al., 2024). Recent graph flow methods additionally optimize the source–target graph pairing using minibatch optimal transport (Hou et al., 2026; Wijesinghe et al., 2026). We complement these approaches by explicitly separating alignment inside and across graph pairs: inner alignment selects representatives within each graph orbit, while outer alignment assigns graphs within a minibatch. We refer to Appendix C for a detailed comparison.
Optimal-transport couplings can straighten Euclidean flow-matching paths (Chemseddine et al., 2025; Pooladian et al., 2023; Tong et al., 2023) and can be extended to geometries with invariances (Klein et al., 2023; Köhler et al., 2020). In particular, if samples themselves carry an optimal-transport geometry, this leads to nested transport formulations, recently used for point-cloud generation (Haviv et al., 2025; Piening et al., 2026; Piening and Beinert, 2026). For graphs, node correspondences induce a Gromov–Monge ground cost, naturally relaxed by Gromov–Wasserstein methods for relational alignment (Beier et al., 2025; Chowdhury et al., 2021; Peyré et al., 2016; Vayer et al., 2020). Beyond particle permutations by linear assignment (Klein et al., 2023; Haviv et al., 2025; Hui et al., 2025; Piening et al., 2026), the proposed graph matching relabels graph and node indices before interpolation. Overall, this results in defining equivariant flow matching (Klein et al., 2023) over a Gromov-type graph geometry (Mémoli and Needham, 2024). Particularly, our contributions are:

- We show that optimal quotient couplings lift to Euclidean couplings with the same quadratic cost, whose interpolations project to constant-speed Wasserstein geodesics.
- We show that diagonal symmetrization preserves the flow-matching objective on equivariant fields and yields an equivariant minimizer, including for categorical endpoint prediction.
- For graphs, we separate alignment inside and across graph pairs: inner alignment selects node representatives, while outer alignment assigns graphs within a minibatch. We study Gromov–Wasserstein and finite-permutation approximations for the former and minibatch optimal transport for the latter.
- Across continuous and categorical graphs, we study how inner and outer alignment affect trajectory straightness and few-step generation, including their scaling with alignment compute. Better inner alignment yields substantial few-step gains, while outer alignment provides smaller, dataset-dependent improvements.

2 Background on transport and flow matching

In flow matching generative models, curves of probability measures are constructed that connect an easy-to-sample source measure with a target measure which is available only through samples (Lipman et al., 2023; Wald and Steidl, 2025). This construction is not limited to probability measures on Euclidean spaces (Chen and Lipman, 2024). After a short recap on Wasserstein geometry (Ambrosio et al., 2005; Santambrogio, 2015; Wald and Steidl, 2025), we will see how the same construction can be formulated on quotient spaces induced by orthogonal group actions.

2.1 Wasserstein geometry and dynamic transport

Let (X, d_X) be a Polish metric space and $\mathcal{P}_2(X)$ the set of Borel probability measures on X with finite second moments, i.e. $\int_X d_X(x, x_0)^2 \, \mathrm{d}\mu(x) < \infty$ for some $x_0 \in X$. The set $\mathcal{P}_2(X)$ becomes a complete metric space with the *Wasserstein distance* $\mathrm{W}_{2,X}$ (Santambrogio, 2015; Villani, 2009) which is given for any $\mu, \nu \in \mathcal{P}_2(X)$ by

$$\mathrm{W}_{2,X}^2(\mu, \nu) \coloneqq \min_{\pi \in \mathrm{c}_X(\mu,\nu)} \int_{X \times X} d_X(x, x')^2 \, \mathrm{d}\pi(x, x'). \tag{1}$$

The *couplings* or *transport plans* are given by

$$\mathrm{c}_X(\mu,\nu) = \{\pi \in \mathcal{P}_2(X \times X) : \mathrm{proj}^0_\sharp \pi = \mu, \mathrm{proj}^1_\sharp \pi = \nu\},$$

with $\mathrm{proj}^i : X \times X \to X$, $(x_0, x_1) \mapsto x_i$. By $\mathrm{c}^{\mathrm{opt}}_X(\mu,\nu)$ we denote *optimal* couplings, where the minimum in (1) is attained. Throughout, let $I = [0,1]$.
For $X = \mathbb{R}^D$, let $(\mu_t)_{t\in I}$ be narrowly continuous and let $v : I \times \mathbb{R}^D \to \mathbb{R}^D$ be Borel. The pair (μ_t, v_t) satisfies the continuity equation if

$$\partial_t \mu_t + \mathrm{div}(v_t \mu_t) = 0 \tag{2}$$

in the sense of distributions. If, in addition,

$$\int_I \int_{\mathbb{R}^D} \|v_t(x)\|^2 \, \mathrm{d}\mu_t(x) \, \mathrm{d}t < \infty,$$

then $(\mu_t)_{t\in I}$ is absolutely continuous with respect to $\mathrm{W}_{2,\mathbb{R}^D}$. Conversely, curves that are absolutely continuous with respect to $\mathrm{W}_{2,\mathbb{R}^D}$ and have square-integrable speed admit a Borel field v satisfying (2) and the displayed integrability condition (Ambrosio et al., 2005, Chapter 8). The definitions and bounds used below are recalled in Appendix A. If v is sufficiently regular, the characteristic ODE

$$\dot\gamma(t,x) = v(t,\gamma(t,x)), \qquad \gamma(0,x) = x, \tag{3}$$

transports μ_0 to μ_t, i.e., $\mu_t = \gamma_{t,\sharp}\mu_0$ (Ambrosio et al., 2005, Proposition 8.1.8).

2.2 Euclidean flow matching

Flow matching turns the dynamic description above into mean-squared-error (MSE) vector-field regression (see, e.g., Lipman et al., 2023; Wald and Steidl, 2025). For $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^D)$ and $\pi \in \mathrm{c}_{\mathbb{R}^D}(\mu,\nu)$, define $\mu_t := \mathrm{proj}^t_\sharp \pi$ with $\mathrm{proj}^t(x,x') := (1-t)x + tx'$. Then (μ_t, v^π_t) solves (2), where v^π minimizes

$$J_\pi(v) := \int_I \int_{\mathbb{R}^D\times\mathbb{R}^D} \left\| v_t(\mathrm{proj}^t(x,x')) - (x'-x) \right\|^2 \mathrm{d}\pi(x,x') \, \mathrm{d}t \tag{4}$$

over jointly Borel vector fields v. Equivalently, for $(X,X') \sim \pi$, we have that

$$v^\pi_t(z) = \mathbb{E}_\pi\big[X' - X \mid \mathrm{proj}^t(X,X') = z\big] \tag{5}$$

for $\mathrm{d}t\,\mathrm{d}\mu_t(z)$-a.e. (t,z). If $\pi \in \mathrm{c}^{\mathrm{opt}}_{\mathbb{R}^D}(\mu,\nu)$, we have that $\int_I \|v^\pi_t(z)\|^2_{L^2(\mu_t)} \, \mathrm{d}t = \mathrm{W}^2_{2,\mathbb{R}^D}(\mu,\nu)$.

Categorical endpoint prediction Let $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^D)$ and $\pi \in \mathrm{c}_{\mathbb{R}^D}(\mu,\nu)$. Suppose that the target is coordinatewise categorical. Write $[m] := \{1,\dots,m\}$ for a positive integer m, and fix a positive integer M. For every $d \in [D]$, fix distinct values $a_{d,1},\dots,a_{d,M} \in \mathbb{R}$ such that $\nu\Big(\prod_{d=1}^D \{a_{d,1},\dots,a_{d,M}\}\Big) = 1$. For $(X,X') \sim \pi$, set $Z_t := (1-t)X + tX'$ and $\mu_t := \mathrm{Law}(Z_t)$. Let $\kappa_d(x') \in [M]$ denote the unique index such that $x'_d = a_{d,\kappa_d(x')}$. Choosing jointly Borel versions, define the conditional coordinate probabilities

$$q^{*,d}_t(n \mid z) := \mathbb{P}_\pi(X'_d = a_{d,n} \mid Z_t = z), \qquad d \in [D], \quad n \in [M].$$

For a.e. $t < 1$ and μ_t-a.e. z, (5) yields

$$\big(v^\pi_t(z)\big)_d = \frac{1}{1-t}\left(\sum_{n=1}^M a_{d,n} q^{*,d}_t(n \mid z) - z_d\right). \tag{6}$$

Moreover, $(q^{*,d})_{d=1}^D$ minimizes the coordinatewise cross-entropy objective

$$J^{\mathrm{coord}}_\pi\big((q^d)_{d=1}^D\big) := -\int_I \int_{\mathbb{R}^D\times\mathbb{R}^D} \sum_{d=1}^D \log q^d_t(\kappa_d(x') \mid (1-t)x + tx') \, \mathrm{d}\pi(x,x') \, \mathrm{d}t \tag{7}$$

over jointly Borel kernels $q^d : I \times \mathbb{R}^D \to \Delta_M$, $d \in [D]$, where $\Delta_M := \{p \in [0,1]^M : \sum_{n=1}^M p_n = 1\}$ and $-\log 0 := +\infty$. Thus, although the joint endpoint takes up to M^D values, it suffices to train D distributions over M classes using (7) and recover the velocity via (6) without any conditional independence assumption (Chemseddine et al., 2026; Eijkelboom et al., 2024).

Few-step distillation Optimal-transport couplings can improve few-step inference by simplifying the teacher trajectories. Complementarily, recent flow-map methods learn finite-time updates of a given flow directly (Geng et al., 2025; Lee et al., 2026; Roos et al., 2026), where the distilled teacher flow still depends on the coupling π used for its training. Write q^{θ_π} for the corresponding frozen categorical teacher. For the categorical parametrization above, let $\delta^{\phi,d}_{s,t}(\cdot \mid z) \in \Delta_M$ and $D^{\phi}_{s,t}(z)_d \coloneqq \sum_{n=1}^{M} a_{d,n}\delta^{\phi,d}_{s,t}(n \mid z)$. Following (Lee et al., 2026), define $\Phi^{\phi}_{s,t}(z) \coloneqq \frac{1-t}{1-s}z + \frac{t-s}{1-s}D^{\phi}_{s,t}(z)$ and train it to match the teacher on the diagonal and satisfy $\Phi_{s,t} = \Phi_{u,t} \circ \Phi_{s,u}$. For $s < u < t$, set $\lambda \coloneqq \frac{(1-t)(u-s)}{(1-u)(t-s)}$ and $\bar{\delta}^{\phi,d}_{s,u,t} \coloneqq \mathrm{sg}\Big[\lambda\delta^{\phi,d}_{s,u}(\cdot \mid Z_s) + (1-\lambda)\delta^{\phi,d}_{u,t}(\cdot \mid \Phi^{\phi}_{s,u}(Z_s))\Big]$. With $Z_r = (1-r)x + rx'$ and independently sampled $(x, x') \sim \mu \otimes \nu$, let $\ell_{\mathrm{comp}} \coloneqq \sum_d \mathrm{KL}(\bar{\delta}^{\phi,d}_{s,u,t} \| \delta^{\phi,d}_{s,t}(\cdot \mid Z_s))$ and $\ell_{\mathrm{diag}} \coloneqq \sum_d \mathrm{KL}(q^{\theta_\pi,d}_{r}(\cdot \mid Z_r) \| \delta^{\phi,d}_{r,r}(\cdot \mid Z_r))$. We minimize

$$\int_{\mathcal{T}} \int \ell_{\mathrm{comp}} \, \mathrm{d}(\mu \otimes \nu) \, \mathrm{d}\rho + \int_{I} \int \ell_{\mathrm{diag}} \, \mathrm{d}(\mu \otimes \nu) \, \mathrm{d}\rho_0, \tag{8}$$

where ρ is supported on $\mathcal{T} \coloneqq \{(s,u,t) : 0 \le s < u < t \le 1\}$, ρ_0 on I, and sg denotes the stop-gradient operation. At zero loss, the diagonal term matches the teacher predictor and the first term enforces flow-map composition. Inference iterates $Z_{t_{k+1}} = \Phi^{\phi}_{t_k,t_{k+1}}(Z_{t_k})$.

3 Flow matching on quotient spaces

Let $G \subset O(D)$ be compact, let $Q \coloneqq \mathbb{R}^D/G$, and write $\mathfrak{q}(x) = [x]$ for the quotient map. The quotient metric is

$$d_Q([x],[y]) \coloneqq \min_{g \in G} \|x - gy\|,$$

and $\mathrm{W}_{2,Q}$ denotes the induced Wasserstein distance. For $\boldsymbol{\mu} \in \mathcal{P}_2(Q)$, write

$$[\boldsymbol{\mu}] \coloneqq \{\mu \in \mathcal{P}_2(\mathbb{R}^D) : \mathfrak{q}_\sharp \mu = \boldsymbol{\mu}\}$$

for its representative laws. In the graph setting, simultaneous node relabeling gives the following construction. Let $Z \coloneqq \mathbb{R}^C$ with $d_Z(u,v) \coloneqq \|u - v\|_2$. For $E, F \in Z^{N \times N}$,

$$\mathrm{GM}^2_{2,Z}([E],[F]) \coloneqq \min_{\sigma \in \mathfrak{S}_N} \sum_{i,j=1}^{N} d^2_Z\big(E_{ij}, F_{\sigma(i)\sigma(j)}\big). \tag{9}$$

Thus, choosing representatives of graph orbits is the (unnormalized) hard Gromov–Monge node-alignment problem, studied in optimal transport (Mémoli and Needham, 2024).

Theorem 3.1 (Representative lifts and quotient geodesics)**.** *Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(Q)$, fix $\mu \in [\boldsymbol{\mu}]$, and let $\Gamma \in \mathrm{c}_Q(\boldsymbol{\mu}, \boldsymbol{\nu})$. There exist $\nu_\Gamma \in [\boldsymbol{\nu}]$ and $\gamma_\Gamma \in \mathrm{c}_{\mathbb{R}^D}(\mu, \nu_\Gamma)$ such that $(\mathfrak{q}, \mathfrak{q})_\sharp \gamma_\Gamma = \Gamma$ and*

$$\int_{\mathbb{R}^D \times \mathbb{R}^D} \|x - y\|^2 \, \mathrm{d}\gamma_\Gamma(x,y) = \int_{Q \times Q} d_Q(\boldsymbol{x}, \boldsymbol{y})^2 \, \mathrm{d}\Gamma(\boldsymbol{x}, \boldsymbol{y}).$$

If Γ is optimal, then

$$\mathrm{W}_{2,Q}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathrm{W}_{2,\mathbb{R}^D}(\mu, \nu_\Gamma) = \mathrm{W}_{2,\mathbb{R}^D}(\mu^G, \nu^G) = \min_{\tilde{\nu} \in [\boldsymbol{\nu}]} \mathrm{W}_{2,\mathbb{R}^D}(\mu, \tilde{\nu}), \tag{10}$$

where $\nu \in [\boldsymbol{\nu}]$ is arbitrary and $\mu^G \coloneqq \mathbb{E}_{g \sim \lambda_G}[g_\sharp \mu]$ for λ_G the normalized Haar measure on G. Moreover, in this case, with $\mathrm{proj}^t(x,y) = (1-t)x + ty$ and $\mu_t \coloneqq \mathrm{proj}^t_\sharp \gamma_\Gamma$, the projected curve $\boldsymbol{\mu}_t \coloneqq \mathfrak{q}_\sharp \mu_t$ is a constant-speed $\mathrm{W}_{2,Q}$-geodesic, and the flow-matching field associated with γ_Γ satisfies

$$\int_I \int_{\mathbb{R}^D} \|v^{\gamma_\Gamma}_t(z)\|^2 \, \mathrm{d}\mu_t(z) \, \mathrm{d}t = \mathrm{W}^2_{2,Q}(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

The proof is given in Appendix A. The theorem shows that any quotient coupling can be lifted through aligned representatives without increasing its cost. In particular, an optimal quotient coupling admits a representative-space lift whose quadratic cost equals the quotient Wasserstein distance. The resulting lift γ_Γ, however, need not be invariant under applying

the same group action to both endpoints, because selecting representatives can break the symmetry. This matters because the velocity or endpoint network is constrained to be G-equivariant. For any representative coupling π—in particular, for γ_Γ—we therefore build a symmetrized coupling by applying the same group element to both endpoints and averaging over G. For a law η and a coupling π, set

$$\eta^G := \mathbb{E}_{g\sim\lambda_G}[g_\sharp\eta], \qquad \pi^G := \mathbb{E}_{g\sim\lambda_G}[(g,g)_\sharp\pi],$$

where the expectation denotes integration against the normalized Haar measure on G. For a finite group such as $\mathfrak{S}_N$, this is the average over all group elements. The resulting coupling is diagonally G-invariant, meaning that $(h,h)_\sharp\pi^G = \pi^G$ for every $h \in G$. It couples μ^G and ν^G and represents the same quotient coupling as π: $(\mathfrak{q},\mathfrak{q})_\sharp\pi^G = (\mathfrak{q},\mathfrak{q})_\sharp\pi$. We next show that, on the G-equivariant model class, training with π is equivalent to training with π^G.

Theorem 3.2 (Equivariant flow matching via symmetrization)**.** *Let $\pi \in \mathrm{c}_{\mathbb{R}^D}(\mu,\nu)$. Then:*

1. *For every G-equivariant field v, $J_\pi(v) = J_{\pi^G}(v)$. The field v^{π^G} can be chosen G-equivariant and minimizes J_π over all G-equivariant fields.*
2. *If the ODE of a G-equivariant field has a unique flow γ_t, then it induces a well-defined quotient flow $\bar\gamma_t([x]) := [\gamma_t(x)]$. For every $\eta \in \mathcal{P}_2(\mathbb{R}^D)$, $\mathfrak{q}_\sharp\gamma_{t,\sharp}\eta = \bar\gamma_{t,\sharp}\mathfrak{q}_\sharp\eta$. Hence the projected curve depends only on the quotient law $\mathfrak{q}_\sharp\eta$, not on the representative lift chosen at time 0.*
3. *If G consists of coordinate-permutation matrices and ν is supported on $\mathcal{A}^D$, $\mathcal{A} = \{a_1,\dots,a_M\}$, let*
$$q_t^{*,\pi^G,d}(n \mid z) := \mathbb{P}_{\pi^G}(X'_d = a_n \mid Z_t = z).$$
For $g \in G$, let $g\cdot d$ be the coordinate determined by $(gz)_{g\cdot d} = z_d$. These coordinate posteriors can be chosen equivariantly for $d \in [D]$ and $n \in [M]$:
$$q_t^{*,\pi^G,g\cdot d}(n \mid gz) = q_t^{*,\pi^G,d}(n \mid z).$$
They minimize J_π^{coord} over equivariant coordinate kernels and recover v^{π^G} via (6).

The full proof is given in Appendix B. Together, Theorems 3.1 and 3.2 connect the practical recipe used below to quotient transport: align representatives and train an equivariant Euclidean model whose flow, when well posed, descends to unordered graphs.

4 Inner and Outer Graph Alignment

For $C = C_\mathrm{e} + C_\mathrm{v}$, encode a graph by $E \in (\mathbb{R}^C)^{N\times N}$ with entries $E_{ik} = (\sqrt{\lambda_\mathrm{edge}}\, e^E_{ik}, \mathbf{1}_{\{i=k\}}\sqrt{\lambda_\mathrm{node}/C_\mathrm{v}}\, f^E_i)$, where $e^E_{ik} \in \mathbb{R}^{C_\mathrm{e}}$ contains edge features and $f^E_i \in \mathbb{R}^{C_\mathrm{v}}$ contains node features on the diagonal. Hence edge and node-feature discrepancies enter the squared Euclidean cost with weights λ_edge and λ_node, respectively, as in fused Gromov–Wasserstein transport (Vayer et al., 2020). We use $\lambda_\mathrm{edge} = \lambda_\mathrm{node} = \frac{1}{2}$ in all experiments. Because the same permutation acts on both node indices, the quotient cost (9) compares all pairwise graph relations after consistent node alignment. Its exact evaluation is an NP-hard quadratic assignment, so we approximate this inner alignment either by Gromov–Wasserstein relaxation (Bauer et al., 2025) or by minimizing over a finite random subset of node permutations, optionally followed by an outer minibatch assignment.

Inner alignment via GW For $E, F \in (\mathbb{R}^C)^{N\times N}$, we use a generalized version (Bauer et al., 2025) of the discrete Gromov–Wasserstein relaxation (Mémoli, 2011; Peyré et al., 2016), allowing for vector-valued edge features:

$$\mathrm{GW}^2_{2,Z}([E],[F]) := \min_{T\in\mathcal{U}_N} \sum_{i,j,k,l=1}^{N} \|E_{ik} - F_{jl}\|^2 T_{ij}T_{kl},$$

where $\mathcal{U}_N = \{T \geq 0 : T\mathbf{1} = T^\top\mathbf{1} = N^{-1}\mathbf{1}\}$. Restricting T to scaled permutation matrices recovers the Gromov–Monge objective, so $\mathrm{GW}^2_{2,Z} \leq \mathrm{GM}^2_{2,Z}/N^2$. We approximately solve

the non-convex problem by Frank–Wolfe iterations (Vayer et al., 2020): each iteration solves an exact linearized transport problem and updates the soft plan by an exact line search. From the final plan T, we recover the node permutation

$$\widehat{\sigma}_{\mathrm{GW}} \in \arg\max_{\sigma \in \mathfrak{S}_N} \sum_{i=1}^{N} T_{i,\sigma(i)},$$

using the Hungarian algorithm (Kuhn, 1955). Using the recovered permutation, we define the aligned representative by $F^{\mathrm{aligned}} \coloneqq \widehat{\sigma}_{\mathrm{GW}}^{-1} \cdot F, F_{ij}^{\mathrm{aligned}} = F_{\widehat{\sigma}_{\mathrm{GW}}(i)\widehat{\sigma}_{\mathrm{GW}}(j)}$. Since GW is a relaxation, this may not be the globally optimal hard alignment.

Inner alignment via Best-of-Augmentation (BoA) As a simple alternative, we sample $\mathcal{A}$ (distinct) permutations $\sigma_1, \ldots, \sigma_{\mathcal{A}} \in \mathfrak{S}_N$ for data augmentation and choose

$$\widehat{\sigma}_{\mathcal{A}} \in \arg\min_{\sigma \in \{\sigma_1, \ldots, \sigma_{\mathcal{A}}\}} \sum_{i,j=1}^{N} d_Z^2\big(E_{ij}, F_{\sigma(i)\sigma(j)}\big).$$

For $\Sigma_{\mathcal{A}} = (\sigma_1, \ldots, \sigma_{\mathcal{A}})$, define $\widehat{c}_{\mathcal{A}}(E, F; \Sigma_{\mathcal{A}}) \coloneqq \sum_{i,j} d_Z^2(E_{ij}, F_{\widehat{\sigma}_{\mathcal{A}}(i)\widehat{\sigma}_{\mathcal{A}}(j)})$. This is an upper bound on $\mathrm{GM}_{2,Z}^2([E],[F])$. For $\mathcal{A} = 1$, this reduces to random augmentation.

Outer alignment For a graph pair (E, F), let $\widehat{\sigma}(E, F) \in \mathfrak{S}_N$ denote the permutation returned by the selected inner alignment and define $\widehat{c}(E, F) \coloneqq \|E - \widehat{\sigma}(E, F)^{-1} \cdot F\|^2$. For independently sampled batches $(E^a)_{a=1}^B$ and $(F^b)_{b=1}^B$, we either set $\widehat{\tau} = \mathrm{id}$ or seek to solve

$$\widehat{\tau} \in \arg\min_{\tau \in \mathfrak{S}_B} \sum_{a=1}^{B} \widehat{c}(E^a, F^{\tau(a)}).$$

For GW, we use the relaxed GW objective in place of exact $\widehat{c}$ when computing the outer assignment, avoiding hard projection for all B^2 pairs. For each selected pair, let $\widehat{\sigma}_a \coloneqq \widehat{\sigma}(E^a, F^{\widehat{\tau}(a)})$ and define $\widehat{\pi}_B \coloneqq B^{-1} \sum_{a=1}^{B} \delta_{(E^a, \widehat{\sigma}_a^{-1} \cdot F^{\widehat{\tau}(a)})}$. Thus, $\widehat{\sigma}_a$ aligns nodes within each pair, while $\widehat{\tau}$ selects which graphs are paired.

Algorithm 1 Inner and outer minibatch graph alignment

Require: Source batch $(E^a)_{a=1}^B$, target batch $(F^b)_{b=1}^B$, inner alignment method
if the outer coupling is independent **then**
  $\widehat{\tau}(a) \leftarrow a$ for all $a = 1, \ldots, B$
else
  $C_{ab} \leftarrow \widehat{c}(E^a, F^b)$ for all a, b
  $\widehat{\tau} \leftarrow \arg\min_{\tau \in \mathfrak{S}_B} \sum_a C_{a,\tau(a)}$
end if
for $a = 1, \ldots, B$ **do**
  $G^a \leftarrow F^{\widehat{\tau}(a)}$
  Obtain $\widehat{\sigma}_a$ using the selected inner alignment
  $\widetilde{F}^a \leftarrow \widehat{\sigma}_a^{-1} \cdot G^a$
end for
return $(E^a, \widetilde{F}^a)_{a=1}^B$

Training Given the aligned pairs, training proceeds as in standard flow matching. Write $\widetilde{F}^a = \widehat{\sigma}_a^{-1} \cdot F^{\widehat{\tau}(a)}$, sample $t^a \sim \mathcal{U}[0,1]$, and set $E_{t^a}^a = (1 - t^a)E^a + t^a \widetilde{F}^a$. The continuous and categorical minibatch losses are, respectively,

$$J_B(v^\theta) \coloneqq \frac{1}{B} \sum_{a=1}^{B} \left\| v^\theta(t^a, E_{t^a}^a) - (\widetilde{F}^a - E^a) \right\|^2. \tag{11}$$

$$J_B^{\mathrm{coord}}\big((q_\theta^d)_{d=1}^D\big) \coloneqq -\frac{1}{B} \sum_{a=1}^{B} \sum_{d=1}^{D} \log q_{\theta,t^a}^d\Big(\kappa_d(\widetilde{F}^a) \mid E_{t^a}^a\Big). \tag{12}$$

For one-hot graph features, the categorical loss and velocity are applied blockwise, see Remark B.5. For datasets containing graphs with different numbers of nodes, one graph transformer is shared across all node counts: it maps an N-node input to an N-node prediction, while N may vary between batches. We form each training batch from graphs with the same N. At inference, we sample N from the empirical distribution of node counts in the training set. Our graph transformers are permutation-equivariant. Implementation details: Appendix F.

Remark 4.1 (**Expected Limit Behavior**). *Let us first consider* ***inner alignment*** *only, where* $[E]$ *and* $[F]$ *are matched randomly and the representative* $\widehat{\sigma}(E,F)^{-1}\cdot F$ *is aligned via BoA or GW. For BoA with* $\mathcal{A}$ *sampled permutations, denote the resulting alignment by* $\widehat{\sigma}_{\mathcal{A}}$ *and its realized cost by* $\widehat{c}_{\mathcal{A}}(E,F;\Sigma_{\mathcal{A}})$*. Random alignment corresponds to* $\mathcal{A}=1$*. Assuming a unique optimal GM alignment* σ_{GM}*, we expect* $\widehat{\sigma}_{\mathcal{A}}\to\sigma_{\mathrm{GM}}$ *and* $\widehat{c}_{\mathcal{A}}(E,F;\Sigma_{\mathcal{A}})\to \mathrm{GM}^2_{2,Z}([E],[F])$ *as* $\mathcal{A}$ *increases, with equality when all* $N!$ *permutations are considered. Our GW approximation need not attain this optimum, but we expect better alignment than BoA for moderate* $\mathcal{A}$*.*
For joint ***inner and outer alignment****, let* $\widehat{\tau}_{\mathcal{A}}\in\mathfrak{S}_B$ *denote the outer graph assignment obtained from the BoA costs* $\widehat{c}_{\mathcal{A}}$ *and let* $\widehat{\tau}_{\mathrm{GW}}$ *denote its GW analogue. With exhaustive inner search, the outer assignment minimizes the empirical GM cost. The GW-based assignment need not coincide. Moreover, with exact inner alignment, i.i.d. sampling, and finite second moments, the optimal average minibatch cost converges to* $\mathrm{W}^2_{2,Q}(\boldsymbol{\mu},\boldsymbol{\nu})$ *as* $B\to\infty$*. By* (10)*, Euclidean minibatch optimal transport (MBOT) (Tong et al., 2023) has the same limit when* $\mu=\mu^G$ *and targets are uniformly relabeled upon each sampling, thereby sampling from* ν^G*.*

Remark 4.2 (**Compute Scaling**). *For fixed compute, inner alignment quality trades off with outer batch size. Joint inner and outer BoA evaluates* $\mathcal{A}$ *permutations for each of the* B^2 *graph pairs. Since one comparison costs* $O(C_{\mathrm{e}}N^2+C_{\mathrm{v}}N)$*, as node features* (C_v) *occur only on the diagonal, and the outer assignment costs* $O(B^3)$*, its total complexity is* $O\big(B^2\mathcal{A}(C_{\mathrm{e}}N^2+C_{\mathrm{v}}N)+B^3\big)$*. Thus,* $\mathcal{A}$ *and* B *trade off inner and outer alignment quality. Since the cost is linear in* $\mathcal{A}$ *but cubic in* B*, additional inner alignment becomes relatively cheaper as* B *grows. In comparison, exposing each source to* $\mathcal{A}B$ *candidate target representatives via MBOT requires* $B_{MBOT}\simeq\mathcal{A}B$*, with complexity* $O(B^2_{MBOT}(C_{\mathrm{e}}N^2+C_{\mathrm{v}}N)+B^3_{MBOT})$*. Hence increasing inner search* $\mathcal{A}$ *is computationally cheaper than obtaining the same candidate-set growth solely through the outer batch size* B*. A classical GW solver with* K *iterations costs* $O(KCN^3)$ *per graph pair, making joint alignment practical only for small* B*. Here, a natural quantity is the mean aligned graph distance* $\mathbb{E}_{E\sim\mu^G,F\sim\nu^G}\|E^a-\widehat{\sigma}_a^{-1}\cdot F^{\widehat{\tau}(a)}\|^2$*, where the expectation includes the minibatch-induced randomness of* $\widehat{\tau}$ *and* $\widehat{\sigma}_a$*, i.e., the transport cost. Note that Sinkhorn solvers accelerate MBOT, but return soft plans rather than permutations and entropic bias increases transport cost.*

5 Numerical experiments

We evaluate the coupling constructions in a permutation-equivariant graph flow-matching model. For inner alignment, we compare random node relabeling (Random), BoA with $\mathcal{A}$ permutations (BoA$^{\mathcal{A}}$), and a 10-iteration Frank–Wolfe GW solver (GW). As a permutation-

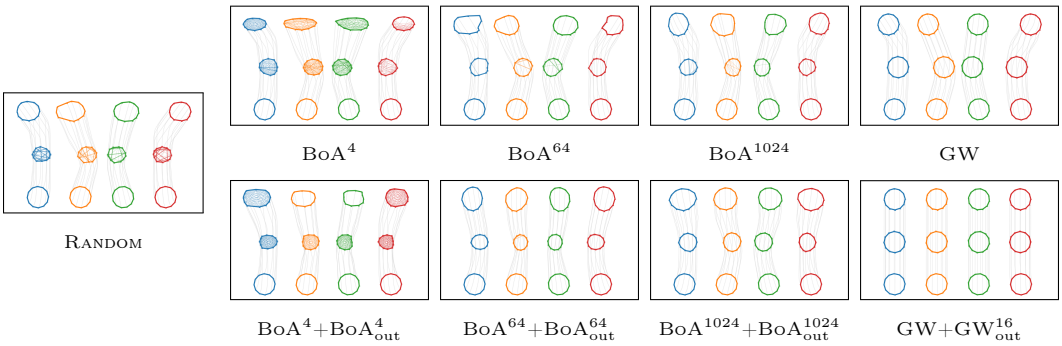


Figure 1: Cycle-graph trajectories ($t=0$: bottom, $t=0.5$: center, $t=1$: top). Colours mark source, lines node paths, and edge width edge value. At $t=0.5$, dense graphs become sparse circles as augmentations $\mathcal{A}$ grow towards GW (top). Outer coupling straightens paths (bottom).

Table 1: Cycle-graphs: Roundness and mean degree are reported as mean_{std}. Best is bold.

Method	Random	MBOT^4	MBOT^{64}	MBOT^{1024}	BoA^4	BoA^{64}	BoA^{1024}	GW	BoA^4 +BoA$^4_{\text{out}}$	BoA^{64} +BoA$^{64}_{\text{out}}$	BoA^{1024} +BoA$^{1024}_{\text{out}}$	GW +GW$^{16}_{\text{out}}$
Roundness (↓)	$12.9_{8.9}$	$18.9_{4.9}$	$8.8_{4.0}$	$7.1_{3.0}$	$26.1_{7.0}$	$16.4_{7.2}$	$7.9_{3.2}$	$0.7_{1.0}$	$11.4_{3.0}$	$5.9_{1.2}$	$5.5_{2.0}$	$\mathbf{0.0}_{0.0}$
Degree (→ 2)	$1.90_{0.28}$	$2.15_{0.09}$	$1.81_{0.14}$	$1.80_{0.12}$	$1.91_{0.09}$	$1.93_{0.13}$	$1.94_{0.03}$	$\mathbf{2.00}_{0.00}$	$1.96_{0.08}$	$1.87_{0.02}$	$1.97_{0.01}$	$\mathbf{2.00}_{0.00}$

blind reference, MBOT^B randomly relabels each target and performs Euclidean minibatch OT over B graphs using the cost $\|E - F\|^2$, without inner alignment. For joint inner–outer alignment, +BoA$^B_{\text{out}}$ assigns B graphs using the BoA$^{\mathcal{A}}$ costs (MBOT$^B \simeq$ BoA^1 +BoA$^B_{\text{out}}$), while +GW$^B_{\text{out}}$ uses the GW costs. The alignment minibatch size B is independent of the network gradient batch size. Aligned minibatches are split or stacked such that every gradient step uses the same number of training pairs. For +GW$^B_{\text{out}}$, we use $B \in \{8, 16\}$ due to its computational cost. Alignment minibatches contain graphs with the same number of nodes. Within each experiment, models share an equivariant graph-transformer backbone and differ only in their training coupling. We evaluate several Euler budgets to assess the effect of alignment on few-step generation. Complete implementation details are given in Appendix F. For variable-size datasets, the population construction first draws the node count $N \sim p_N$ from its empirical distribution. Conditional on N, Random uses the product coupling $\mu_N \otimes \nu_N$, where μ_N and ν_N denote the source and target laws for N nodes. Inner BoA/GW changes only the representation of target graphs, whereas outer variants also optimize the source–target pairing within fixed-N minibatches. See Appendix D for further experiments.

5.1 Continuous target via translated cycle graphs

Inspired by (Piening et al., 2026), we consider randomly relabelled 12-node rings represented by adjacency and planar 2D node positions. Source and target differ only by translation and independent rotation, so they have the same edge law and any intermediate deformation is induced solely by the coupling. To give the source a density and stabilize training, graph sources are perturbed by small Gaussian noise ($\sigma = 0.01$) after coupling. At test time, we start from exact rings and integrate with 125 Euler steps. We report roundness, the standard deviation of node distances to the centroid (rescaled by $\times 10^2$, 0 for an exact ring), and the mean degree $\frac{1}{N}\sum_{i \neq j} \text{Adj}_{ij}$ of the weighted adjacency matrix, which should be 2.

Figure 1 and Table 1 show the results for 125 Euler steps. Under Random, mismatched edges blur rings halfway along the flow, and generated rings are strongly deformed. For BoA, roundness improves steadily as $\mathcal{A}$ grows, though small $\mathcal{A}$ can underperform Random. GW reaches near-exact rings. Adding the outer assignment straightens the paths and improves roundness at every $\mathcal{A}$, and GW+GW$_{\text{out}}$ recovers exact rings. The permutation-blind MBOT also improves with B, but stays far from GW. The two GW-based alignments give the best results and recover exact circles. Inner–outer GW induces continuous translation of circles.

5.2 Categorical target via molecular graph generation

Data We use QM9 ($N \leq 9$) (Ramakrishnan et al., 2014) and ZINC250k ($N \leq 38$) (Jin et al., 2018) in the categorical CatFlow/DiGress representation (Eijkelboom et al., 2024; Vignac et al., 2023). Edges are one-hot over {none, single, double, triple}, while each diagonal node feature concatenates a one-hot atom type with a formal-charge one-hot over $\{-1, 0, +1\}$. QM9 uses atom types {C, N, O, F}, giving $C = 11$. ZINC250k uses {C, N, O, F, Br, Cl, I, P, S}, giving $C = 16$. The network predicts each categorical feature block and minimises the blockwise loss (12). Its velocity uses the blockwise conditional-mean in Remark B.5.

Metrics Following the GDSS/CatFlow evaluation protocols (Eijkelboom et al., 2024; Jo et al., 2022), we report uncorrected molecular validity, the fraction of generated graphs that RDKit sanitizes without correction, and Fréchet ChemNet Distance (FCD). Validity captures chemical correctness and FCD distributional fidelity. Higher is better for validity and lower is better for FCD. We follow (Eijkelboom et al., 2024) and omit novelty since it might reward unfaithful modeling for the selected benchmark datasets. Similarly, we defer uniqueness reporting to Appendix H, as it is saturated and uninformative.

Compute scaling of the alignment For fixed N, we sample 32,768 source graphs with relabeled QM9/ZINC250k targets and record $\mathbb{E}_{E \sim \mu^G, F \sim \nu^G} \|E^a - \widehat{\sigma}_a^{-1} \cdot F^{\widehat{\tau}(a)}\|^2$ against CPU time (Fig. 2), where we split the data into B-sized chunks. The distance depends almost only on the budget $\mathcal{A}B$, but joint splits reach it fastest. On QM9 ($N = 9$), BoA^{64} + BoA$^{512}_{\text{out}}$ matches MBOT^{32768} but is 12× faster. The tradeoff between $\mathcal{A}$ and B depends on N and

the feature dimension C. On QM9 splits ($N \in \{8, 9\}$, $C = 11$), BoA$^{\mathcal{A}}$ approaches exhaustive search over $\mathfrak{S}_N$ and lies below GW. On ZINC250k splits ($N \in \{18, 23\}$, $C = 16$), random search in $\mathfrak{S}_N$ seems hopeless and GW is clearly best.

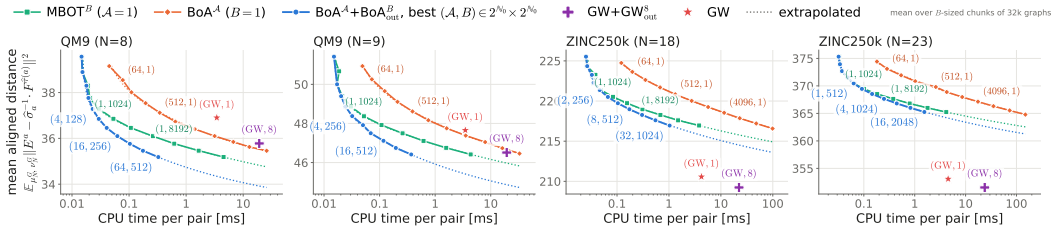


Figure 2: Mean aligned distance vs. CPU time per pair. Labels give $(\mathcal{A}, B)$. Dotted: extrapolated least-squares $\mathcal{A}B$ power-law fits (distance: $c_\infty + a(\mathcal{A}B)^{-p}$, time: $t_0 + b(\mathcal{A}B)^q$).

Limited-budget sweep of endpoint-induced velocity field and flow map We train the shared 2.8M-parameter backbone for 100 epochs under each coupling and evaluate 10,000 samples at 5/25/125 Euler steps. We additionally distill each model into a flow map using the same backbone and (8), evaluated at 1/2/5 steps, see Appendix F. At low ODE budgets (Table 2a), GW alignment consistently improves validity and FCD over random coupling, with the largest gains at five steps. Outer assignment further improves validity, while its FCD benefit is dataset dependent. MBOT is competitive on QM9 but weaker than GW in FCD and on ZINC250k. Differences narrow as the integration budget increases. For the distilled flow map (Table 2b), coupling differences are smaller, but GW-based alignment displays an advantage for two or five steps.

Table 2: Molecular short-budget sweep: (2a) Flow ODE at 5/25/125 Euler steps and (2b) flow map at 1/2/5 steps. Rows are mean$_{\text{std}}$ over 3 evaluation repeats. Best ($\pm 2\sigma$) is bold.

(a) Euler Integration of Learned Flow ODE

Method	Validity (↑) 5	Validity (↑) 25	Validity (↑) 125	FCD (↓) 5	FCD (↓) 25	FCD (↓) 125
QM9 ($N \le 9$)						
Random	$0.8849_{.0020}$	$0.9564_{.0002}$	$0.9648_{.0010}$	$1.731_{.017}$	$0.637_{.022}$	$0.530_{.012}$
MBOT^{1024}	$0.9327_{.0016}$	$\mathbf{0.9716}_{.0018}$	$\mathbf{0.9785}_{.0012}$	$1.581_{.018}$	$\mathbf{0.585}_{.051}$	$0.494_{.019}$
BoA^{4}+BoA$^{256}_{\text{out}}$	$0.9283_{.0041}$	$\mathbf{0.9735}_{.0013}$	$\mathbf{0.9769}_{.0005}$	$1.621_{.031}$	$0.631_{.003}$	$\mathbf{0.511}_{.034}$
GW	$0.9356_{.0020}$	$0.9695_{.0010}$	$0.9744_{.0014}$	$\mathbf{1.278}_{.046}$	$0.589_{.015}$	$\mathbf{0.491}_{.022}$
GW+GW$^{8}_{\text{out}}$	$\mathbf{0.9421}_{.0024}$	$\mathbf{0.9728}_{.0011}$	$\mathbf{0.9783}_{.0009}$	$\mathbf{1.232}_{.017}$	$\mathbf{0.532}_{.014}$	$\mathbf{0.440}_{.016}$
ZINC250k ($N \le 38$)						
Random	$0.5641_{.0048}$	$0.8140_{.0034}$	$0.8597_{.0011}$	$18.448_{.095}$	$13.061_{.162}$	$11.909_{.077}$
MBOT^{1024}	$0.6405_{.0034}$	$0.8343_{.0013}$	$0.8671_{.0003}$	$17.383_{.009}$	$12.628_{.127}$	$11.685_{.033}$
BoA^{2}+BoA$^{512}_{\text{out}}$	$0.6472_{.0035}$	$0.8541_{.0009}$	$0.8815_{.0004}$	$17.366_{.115}$	$12.470_{.165}$	$11.612_{.071}$
GW	$0.6431_{.0029}$	$0.8472_{.0052}$	$0.8742_{.0009}$	$\mathbf{15.024}_{.076}$	$\mathbf{11.993}_{.030}$	$\mathbf{11.365}_{.073}$
GW+GW$^{8}_{\text{out}}$	$\mathbf{0.7280}_{.0055}$	$\mathbf{0.8763}_{.0026}$	$\mathbf{0.8930}_{.0022}$	$15.377_{.048}$	$12.116_{.025}$	$\mathbf{11.370}_{.059}$

(b) Distilled Flow Map for Few-Step Generation

Method	Validity (↑) 1	Validity (↑) 2	Validity (↑) 5	FCD (↓) 1	FCD (↓) 2	FCD (↓) 5
QM9 ($N \le 9$)						
Random	$0.8183_{.0049}$	$0.8715_{.0023}$	$0.9040_{.0029}$	$\mathbf{1.576}_{.011}$	$1.295_{.022}$	$0.839_{.021}$
MBOT^{1024}	$\mathbf{0.8499}_{.0030}$	$0.8955_{.0004}$	$\mathbf{0.9447}_{.0006}$	$\mathbf{1.546}_{.022}$	$1.299_{.042}$	$0.835_{.016}$
BoA^{4}+BoA$^{256}_{\text{out}}$	$0.8353_{.0023}$	$0.8934_{.0010}$	$\mathbf{0.9424}_{.0024}$	$\mathbf{1.579}_{.013}$	$1.224_{.018}$	$0.866_{.038}$
GW	$0.8381_{.0051}$	$\mathbf{0.9056}_{.0029}$	$\mathbf{0.9447}_{.0020}$	$\mathbf{1.555}_{.042}$	$\mathbf{1.187}_{.046}$	$0.827_{.013}$
GW+GW$^{8}_{\text{out}}$	$\mathbf{0.8496}_{.0041}$	$\mathbf{0.9085}_{.0028}$	$\mathbf{0.9474}_{.0020}$	$\mathbf{1.529}_{.049}$	$\mathbf{1.175}_{.015}$	$\mathbf{0.766}_{.016}$
ZINC250k ($N \le 38$)						
Random	$0.4608_{.0008}$	$0.6016_{.0032}$	$0.6611_{.0024}$	$18.561_{.133}$	$16.595_{.064}$	$15.496_{.063}$
MBOT^{1024}	$0.4646_{.0056}$	$0.6308_{.0056}$	$0.6865_{.0022}$	$\mathbf{18.110}_{.104}$	$\mathbf{16.137}_{.076}$	$\mathbf{14.511}_{.038}$
BoA^{2}+BoA$^{512}_{\text{out}}$	$\mathbf{0.4972}_{.0029}$	$\mathbf{0.6758}_{.0001}$	$0.7354_{.0023}$	$\mathbf{18.305}_{.086}$	$16.322_{.047}$	$14.701_{.100}$
GW	$0.4418_{.0039}$	$0.6238_{.0056}$	$0.7302_{.0045}$	$\mathbf{18.198}_{.059}$	$\mathbf{16.257}_{.049}$	$\mathbf{14.402}_{.090}$
GW+GW$^{8}_{\text{out}}$	$0.4699_{.0032}$	$0.6545_{.0033}$	$\mathbf{0.7653}_{.0030}$	$\mathbf{18.200}_{.074}$	$16.555_{.028}$	$\mathbf{14.469}_{.042}$

Full-budget molecular generation We train larger GW-aligned endpoint models without outer alignment and evaluate the resulting GW-CatFlow model with 500 Euler steps. The complete protocol is in Appendix E. Table 3 gives 99.34% validity and 0.115 FCD on QM9, comparable to DeFoG, and 99.01% validity with 0.966 FCD on ZINC250k, the lowest FCD listed. This shows that the underlying model remains competitive with state-of-the-art baselines at full budget. The controlled sweep above isolates the effect of coupling.

6 Conclusion

We study inner and outer alignment in equivariant graph flow matching. Both simplify trajectories and improve few-step generation, while compute scaling reveals a tradeoff between better inner alignment and larger outer batches. Faster GW approximations (Beier et al., 2022; Chowdhury et al., 2021; Piening and Beinert, 2025; Jin et al., 2022; Scetbon et al., 2022) could shift this compute tradeoff, while semi-discrete formulations may allow for outer alignment beyond minibatches (Kong et al., 2025; Mousavi-Hosseini et al., 2025).

Table 3: Molecular full-budget run. Ours: 500 Euler steps, 10,000 samples, mean$_{\text{std}}$ over three repeats. Baselines follow published protocols. GraphAF, MoFlow, and DiGress values are quoted from (Hou et al., 2026), DeFoG values from (Xiong et al., 2026).

Method	QM9 Validity (↑)	QM9 FCD (↓)	ZINC250k Validity (↑)	ZINC250k FCD (↓)
GraphAF (Shi et al., 2020)	67.14	5.246	67.92	16.128
MoFlow (Zang and Wang, 2020)	92.03	4.536	63.76	20.875
GDSS (Jo et al., 2022)	95.72	2.900	97.01	14.656
CatFlow (Eijkelboom et al., 2024)	99.81	0.441	99.21	13.211
DiGress (Vignac et al., 2023)	98.29	0.095	94.98	3.482
GGFlow (Hou et al., 2026)	99.91	0.148	99.63	1.455
DeFoG (Qin et al., 2025)	99.30	0.120	99.22	1.425
VBFN (Xiong et al., 2026)	99.98	0.083	99.63	1.307
GW-CatFlow (ours)	$99.34_{.08}$	$0.115_{.004}$	$99.01_{.03}$	$0.966_{.013}$

AI Use Statement

The research questions, methodological ideas, and scientific conclusions of this work were developed by the authors. Generative AI tools were used as research assistants for language editing and literature search, assistance with proof development and mathematical arguments, programming and debugging, and drafting of supplementary material. All AI-assisted content was critically reviewed and revised by the authors, who take full responsibility for the final work.

References

Michael Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *Proceedings of the ICLR'23*. OpenReview.net, 2023.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005.

Martin Bauer, Facundo Mémoli, Tom Needham, and Mao Nishino. The Z-Gromov-Wasserstein distance. *Journal of Machine Learning Research*, 26(291):1–57, 2025.

Florian Beier, Robert Beinert, and Gabriele Steidl. On a linear Gromov–Wasserstein distance. *IEEE Transactions on Image Processing*, 31:7292–7305, 2022.

Florian Beier, Moritz Piening, Robert Beinert, and Gabriele Steidl. Joint metric space embedding by unbalanced optimal transport with Gromov–Wasserstein marginal penalization. In *Proceedings of ICML'25*. OpenReview.net, 2025.

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional Wasserstein distances with applications in Bayesian OT flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025.

Jannis Chemseddine, Gregor Kornhardt, and Gabriele Steidl. Spherical flows for sampling categorical data. *arXiv preprint arXiv:2605.05629*, 2026.

Ricky TQ Chen and Yaron Lipman. Flow matching on general geometries. In *Proceedings of the ICLR'24*. OpenReview.net, 2024.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog Bits: Generating discrete data using diffusion models with self-conditioning. In *Proceedings of the ICLR'23*. OpenReview.net, 2023.

Samir Chowdhury, David Miller, and Tom Needham. Quantized Gromov–Wasserstein. In *Proceedings of ECML PKDD'21*, pages 811–827. Springer, 2021.

Floor Eijkelboom, Grigory Bartosh, Christian A. Naesseth, Max Welling, and Jan-Willem van de Meent. Variational flow matching for graph generation. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, 2024.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38: 75460–75482, 2025.

Doron Haviv, Aram-Alexandre Pooladian, Dana Pe'er, and Brandon Amos. Wasserstein flow matching: Generative modeling over families of distributions. In *Proceedings of the ICML'25*. OpenReview.net, 2025.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *Proceedings of the ICML'22*. PMLR, 2022.

Xiaoyang Hou, Tian Zhu, Milong Ren, Dongbo Bu, Xin Gao, Chunming Zhang, and Shiwei Sun. Ggflow: A graph flow matching method with efficient optimal transport. *Transactions on Machine Learning Research*, 2026.

Tomaž Hočevar and Janez Demšar. A combinatorial approach to graphlet counting. *Bioinformatics*, 30(4):559–565, 2014. doi: 10.1093/bioinformatics/btt717.

Ka-Hei Hui, Chao Liu, Xiaohui Zeng, Chi-Wing Fu, and Arash Vahdat. Not-so-optimal transport flows for 3d point cloud generation. In *Proceedings of the ICLR'25*. OpenReview.net, 2025.

Keyue Jiang, Jiahao Cui, Xiaowen Dong, and Laura Toni. Bures-Wasserstein flow matching for graph generation. In *Proceedings of the ICLR'26*. OpenReview.net, 2026.

Hongwei Jin, Zishun Yu, and Xinhua Zhang. Orthogonal Gromov–Wasserstein discrepancy with efficient lower bound. In *Proceedings of UAI'22*, volume 180 of *Proceedings of Machine Learning Research*, pages 917–927. PMLR, 2022.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *Proceedings of the ICML'18*, pages 2323–2332. PMLR, 2018.

Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *Proceedings of the ICML'22*. PMLR, 2022.

Alexander Kechris. *Classical descriptive set theory*. Springer Science & Business Media, 2012.

Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. *Advances in Neural Information Processing Systems*, 36:59886–59910, 2023.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: Exact likelihood generative learning for symmetric densities. In *Proceedings of the ICML'20*, pages 5361–5370. PMLR, 2020.

Lingkai Kong, Molei Tao, Yang Liu, Bryan Wang, Jinmiao Fu, Chien-Chih Wang, and Huidong Liu. Alignflow: Improving flow-based generative models with semi-discrete optimal transport. *arXiv preprint arXiv:2510.15038*, 2025.

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

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proceedings of the ICLR'23*. OpenReview.net, 2023.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proceedings of the ICLR'23*. OpenReview.net, 2023.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11:417–487, 2011.

Facundo Mémoli and Tom Needham. Comparison results for Gromov–Wasserstein and Gromov–Monge distances. *ESAIM: Control, Optimisation and Calculus of Variations*, 30: 78, 2024. doi: 10.1051/cocv/2024063. URL `https://www.esaim-cocv.org/articles/cocv/abs/2024/01/cocv230154/cocv230154.html`.

Alireza Mousavi-Hosseini, Stephen Y Zhang, Michal Klein, and Marco Cuturi. Flow matching with semidiscrete couplings. *arXiv preprint arXiv:2509.25519*, 2025.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov–Wasserstein averaging of kernel and distance matrices. In *Proceedings of ICML'16*, volume 48 of *Proceedings of Machine Learning Research*, pages 2664–2672. PMLR, 2016.

Moritz Piening and Robert Beinert. A novel sliced fused Gromov-Wasserstein distance. *arXiv preprint arXiv:2508.02364*, 2025.

Moritz Piening and Robert Beinert. Slicing Wasserstein over Wasserstein via functional optimal transport. In *Proceedings of the ICLR'26*. OpenReview.net, 2026.

Moritz Piening, Richard Duong, and Gabriele Steidl. Generalized wasserstein flow matching: Transport plans, everywhere, all at once. *arXiv preprint arXiv:2605.08424*, 2026.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Proceedings of the ICML'23*. OpenReview.net, 2023.

Yiming Qin, Manuel Madeira, Dorina Thanou, and Pascal Frossard. DeFoG: Discrete flow matching for graph generation. In *Proceedings of the ICML'25*. OpenReview.net, 2025.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1: 140022, 2014.

Daan Roos, Oscar Davis, Floor Eijkelboom, Michael Bronstein, Max Welling, Ismail Ilkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. Categorical flow maps. *arXiv preprint arXiv:2602.12233*, 2026.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs and Modeling*. Birkhäuser, Cham, 2015.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time Gromov-–Wasserstein distances using low-rank couplings and costs. In *Proceedings of ICML'22*, volume 162 of *Proceedings of Machine Learning Research*, pages 19347–19365. PMLR, 2022.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. GraphAF: A flow-based autoregressive model for molecular graph generation. In *Proceedings of the ICLR'20*. OpenReview.net, 2020.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3d molecule generation. *Advances in Neural Information Processing Systems*, 36:549–568, 2023.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2023.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused Gromov–Wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. DiGress: Discrete denoising diffusion for graph generation. In *Proceedings of the ICLR'23*. OpenReview.net, 2023.

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

Christian Wald and Gabriele Steidl. Flow matching: Markov kernels, stochastic processes and transport plans. *Variational and Information Flows in Machine Learning and Optimal Transport*, pages 185–254, 2025.

Asiri Wijesinghe, Sevvandi Kandanaarachchi, Daniel M Steinberg, and Cheng Soon Ong. Flowette: Flow matching with graphette priors for graph generation. *arXiv preprint arXiv:2602.23566*, 2026.

Yida Xiong, Jiameng Chen, Xiuwen Gong, Jia Wu, Shirui Pan, and Wenbin Hu. Variational Bayesian flow network for graph generation. In *Proceedings of the ICML'26*. OpenReview.net, 2026. URL https://openreview.net/forum?id=NNU8cO5wte.

Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin. Gromov–Wasserstein learning for graph matching and node embedding. In *Proceedings of the ICML'19*, pages 6932–6941. PMLR, 2019.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the ICCV'19*, pages 4541–4550. IEEE, 2019.

Jiaxuan You, Rex Ying, Xiang Ren, William L. Hamilton, and Jure Leskovec. GraphRNN: Generating realistic graphs with deep auto-regressive models. In *Proceedings of the ICML'18*. PMLR, 2018.

Chengxi Zang and Fei Wang. MoFlow: An invertible flow model for generating molecular graphs. In *Proceedings of the KDD'20*, pages 617–626. ACM, 2020.

A Proof of representative lifts and quotient geodesics

This section proves Theorem 3.1. We first recall the metric notions used in the geodesic part of the proof. A curve $(\mu_t)_{t\in I}$ belongs to $AC_I^2(\mathcal{P}_2(X))$ if there exists $m \in L^2(I)$ such that

$$\mathrm{W}_{2,X}(\mu_s, \mu_t) \le \int_s^t m(r)\ \mathrm{d}r \qquad \text{for all } 0 \le s \le t \le 1.$$

Let $G \subset O(D)$ be compact and define

$$Q \coloneqq \mathbb{R}^D/G, \qquad [x] \coloneqq \{gx : g \in G\},$$

with quotient map

$$\mathfrak{q} : \mathbb{R}^D \to Q, \qquad \mathfrak{q}(x) = [x].$$

The quotient is equipped with the metric

$$d_Q([x],[y]) \coloneqq \min_{g\in G} \|x - gy\|.$$

Since G is compact and acts by isometries, the minimum is attained and d_Q defines a metric on Q. Since G is compact and acts isometrically, (Q, d_Q) is Polish, and $\mathfrak{q}$ is 1-Lipschitz.

For $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(Q)$, the corresponding Wasserstein distance is

$$\mathrm{W}_{2,Q}^2(\boldsymbol{\mu}, \boldsymbol{\nu}) \coloneqq \min_{\Gamma \in \mathrm{c}_Q(\boldsymbol{\mu},\boldsymbol{\nu})} \int_{Q\times Q} d_Q(\boldsymbol{x}, \boldsymbol{y})^2\ \mathrm{d}\Gamma(\boldsymbol{x}, \boldsymbol{y}), \tag{13}$$

where

$$\mathrm{c}_Q(\boldsymbol{\mu}, \boldsymbol{\nu}) \coloneqq \{\Gamma \in \mathcal{P}_2(Q \times Q) : \mathrm{proj}^0_\sharp \Gamma = \boldsymbol{\mu}, \mathrm{proj}^1_\sharp \Gamma = \boldsymbol{\nu}\}.$$

By $\mathrm{c}_Q^{\mathrm{opt}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ we denote the couplings which minimize (13).

We denote by λ_G the normalized Haar probability measure on the compact group G. Thus, for any integrable $f : G \to \mathbb{R}$,

$$\int_G f(g)\ \mathrm{d}\lambda_G(g)$$

is invariant under left and right multiplication of the argument of f. If G is finite, in particular for permutation groups, λ_G is simply the uniform measure,

$$\int_G f(g)\ \mathrm{d}\lambda_G(g) = \frac{1}{|G|}\sum_{g\in G} f(g).$$

Proposition A.1 (Representative lifting of quotient couplings)**.** *Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(Q)$, fix $\mu \in [\boldsymbol{\mu}]$, and let $\Gamma \in \mathrm{c}_Q(\boldsymbol{\mu}, \boldsymbol{\nu})$. Then*

$$\int_{Q\times Q} d_Q(\boldsymbol{x}, \boldsymbol{y})^2\ \mathrm{d}\Gamma(\boldsymbol{x}, \boldsymbol{y}) = \min_{\substack{\nu\in[\boldsymbol{\nu}],\ \gamma\in \mathrm{c}_{\mathbb{R}^D}(\mu,\nu) \\ (\mathfrak{q},\mathfrak{q})_\sharp\gamma=\Gamma}} \int_{\mathbb{R}^D\times\mathbb{R}^D} \|x - y\|^2\ \mathrm{d}\gamma(x, y).$$

In particular, if $\Gamma \in \mathrm{c}_Q^{\mathrm{opt}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $(\nu_\Gamma, \gamma_\Gamma)$ attains the minimum on the right-hand side, then $\gamma_\Gamma \in \mathrm{c}^{\mathrm{opt}}_{\mathbb{R}^D}(\mu, \nu_\Gamma)$ and $\mathrm{W}^2_{2,\mathbb{R}^D}(\mu, \nu_\Gamma) = \mathrm{W}^2_{2,Q}(\boldsymbol{\mu}, \boldsymbol{\nu})$.

Proof. Let $\nu \in [\boldsymbol{\nu}]$ and $\gamma \in \mathrm{c}_{\mathbb{R}^D}(\mu, \nu)$ satisfy $(\mathfrak{q}, \mathfrak{q})_\sharp\gamma = \Gamma$. Since

$$d_Q(\mathfrak{q}(x), \mathfrak{q}(y)) \le \|x - y\|,$$

we have

$$\int_{Q\times Q} d_Q(\boldsymbol{x}, \boldsymbol{y})^2\ \mathrm{d}\Gamma(\boldsymbol{x}, \boldsymbol{y}) \le \int_{\mathbb{R}^D\times\mathbb{R}^D} \|x - y\|^2\ \mathrm{d}\gamma(x, y).$$

Conversely, disintegrate

$$\mu = \int_Q \mu_{\boldsymbol{x}}\ \mathrm{d}\boldsymbol{\mu}(\boldsymbol{x}), \qquad \Gamma = \int_Q \delta_{\boldsymbol{x}} \otimes \Gamma_{\boldsymbol{x}}\ \mathrm{d}\boldsymbol{\mu}(\boldsymbol{x}),$$

where $\mu_{\boldsymbol{x}}$ is supported on $\mathfrak{q}^{-1}(\boldsymbol{x})$, and define $\eta \in \mathcal{P}(\mathbb{R}^D \times Q)$ by

$$\eta(A \times B) \coloneqq \int_Q \mu_{\boldsymbol{x}}(A)\Gamma_{\boldsymbol{x}}(B)\ \mathrm{d}\boldsymbol{\mu}(\boldsymbol{x}).$$

Then $(\mathfrak{q}, \mathrm{id})_\sharp \eta = \Gamma$.

Since G is compact, the corresponding argmin relation is closed with nonempty compact sections. Hence, a measurable selection theorem (Kechris, 2012, Theorem 18.18) yields a Borel map $Y(x, \boldsymbol{y}) \in \mathfrak{q}^{-1}(\boldsymbol{y})$ satisfying

$$\|x - Y(x, \boldsymbol{y})\| = d_Q(\mathfrak{q}(x), \boldsymbol{y}).$$

Set $\gamma_\Gamma \coloneqq (\mathrm{id}, Y)_\sharp \eta$, and let ν_Γ be its second marginal. Then $(\mathfrak{q}, \mathfrak{q})_\sharp \gamma_\Gamma = \Gamma$ and $\mathfrak{q}_\sharp \nu_\Gamma = \boldsymbol{\nu}$. Moreover, since $G \subset O(D)$, we get

$$\int_{\mathbb{R}^D} \|y\|^2\ \mathrm{d}\nu_\Gamma(y) = \int_Q d_Q(\boldsymbol{y}, [0])^2\ \mathrm{d}\boldsymbol{\nu}(\boldsymbol{y}) < \infty,$$

and hence $\nu_\Gamma \in [\boldsymbol{\nu}]$. Finally,

$$\int_{\mathbb{R}^D \times \mathbb{R}^D} \|x - y\|^2\ \mathrm{d}\gamma_\Gamma(x, y) = \int_{Q \times Q} d_Q(\boldsymbol{x}, \boldsymbol{y})^2\ \mathrm{d}\Gamma(\boldsymbol{x}, \boldsymbol{y}).$$

Thus the minimum is attained. If, in addition, $\Gamma \in \mathrm{c}_Q^{\mathrm{opt}}(\boldsymbol{\mu}, \boldsymbol{\nu})$, then

$$\mathrm{W}_{2,Q}^2(\boldsymbol{\mu}, \boldsymbol{\nu}) \le \mathrm{W}_{2,\mathbb{R}^D}^2(\mu, \nu_\Gamma) \le \int_{\mathbb{R}^D \times \mathbb{R}^D} \|x - y\|^2\ \mathrm{d}\gamma_\Gamma(x, y) = \mathrm{W}_{2,Q}^2(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

Hence all inequalities are equalities, which proves the final claim. □

Symmetrized representative claim in Theorem 3.1. Let Γ be optimal and let $(\nu_\Gamma, \gamma_\Gamma)$ be as in Proposition A.1. We have for $\nu_1, \nu_2 \in [\boldsymbol{\nu}]$ that

$$\begin{aligned}
\int_{\mathbb{R}^D} f(x)\ \mathrm{d}\nu_1^G(x) &= \int_Q \left(\int_G f(gx)\ \mathrm{d}\lambda_G(g) \right) \mathrm{d}(\mathfrak{q}_\sharp \nu_1)([x]) \\
&= \int_Q \left(\int_G f(gx)\ \mathrm{d}\lambda_G(g) \right) \mathrm{d}(\mathfrak{q}_\sharp \nu_2)([x]) \\
&= \int_{\mathbb{R}^D} f(x)\ \mathrm{d}\nu_2^G(x)
\end{aligned}$$

because $x \mapsto \left(\int_G f(gx)\ \mathrm{d}\lambda_G(g) \right)$ is G-invariant and descends to a function on the quotient space. Thus all representatives of the same quotient law have the same G-symmetrization, and we have $\nu_\Gamma^G = \nu^G$ for every $\nu \in [\boldsymbol{\nu}]$. Define

$$\gamma_\Gamma^G \coloneqq \mathbb{E}_{g \sim \lambda_G}[(g, g)_\sharp \gamma_\Gamma].$$

Then $\gamma_\Gamma^G \in \mathrm{c}_{\mathbb{R}^D}(\mu^G, \nu_\Gamma^G)$ and, since $G \subset O(D)$,

$$\int \|x - y\|^2\ \mathrm{d}\gamma_\Gamma^G = \int \|x - y\|^2\ \mathrm{d}\gamma_\Gamma = \mathrm{W}_{2,Q}^2(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

Moreover, $\mathfrak{q}_\sharp \mu^G = \boldsymbol{\mu}$ and $\mathfrak{q}_\sharp \nu_\Gamma^G = \boldsymbol{\nu}$, hence

$$\mathrm{W}_{2,Q}^2(\boldsymbol{\mu}, \boldsymbol{\nu}) \le \mathrm{W}_{2,\mathbb{R}^D}^2(\mu^G, \nu_\Gamma^G) \le \int \|x - y\|^2\ \mathrm{d}\gamma_\Gamma^G.$$

Thus

$$\mathrm{W}_{2,\mathbb{R}^D}(\mu^G, \nu^G) = \mathrm{W}_{2,\mathbb{R}^D}(\mu^G, \nu_\Gamma^G) = \mathrm{W}_{2,Q}(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

□

Geodesic and energy claims in Theorem 3.1. Let Γ be optimal and let $(\nu_\Gamma, \gamma_\Gamma)$ be the pair constructed in Proposition A.1. The proposition gives

$$\gamma_\Gamma \in \mathrm{c}^{\mathrm{opt}}_{\mathbb{R}^D}(\mu, \nu_\Gamma), \qquad \mathrm{W}^2_{2,\mathbb{R}^D}(\mu, \nu_\Gamma) = \mathrm{W}^2_{2,Q}(\boldsymbol{\mu}, \boldsymbol{\nu}). \tag{14}$$

For $0 \le s < t \le 1$, the coupling $(\mathrm{proj}^s, \mathrm{proj}^t)_\sharp \gamma_\Gamma$ and the fact that $\mathfrak{q}$ is 1-Lipschitz give

$$\mathrm{W}_{2,Q}(\boldsymbol{\mu}_s, \boldsymbol{\mu}_t) \le \mathrm{W}_{2,\mathbb{R}^D}(\mu_s, \mu_t) \le (t-s)\, \mathrm{W}_{2,Q}(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

Thus $(\boldsymbol{\mu}_t)_{t\in I}$ belongs to $AC^2_I(\mathcal{P}_2(Q))$. Applying this bound on $[0, s]$, $[s, t]$, and $[t, 1]$, and combining it with the triangle inequality between $\boldsymbol{\mu}_0 = \boldsymbol{\mu}$ and $\boldsymbol{\mu}_1 = \boldsymbol{\nu}$, forces equality in each bound. Hence

$$\mathrm{W}_{2,Q}(\boldsymbol{\mu}_s, \boldsymbol{\mu}_t) = (t-s)\, \mathrm{W}_{2,Q}(\boldsymbol{\mu}, \boldsymbol{\nu}),$$

so $(\boldsymbol{\mu}_t)_{t\in I}$ is a constant-speed geodesic.

Finally, the Euclidean optimal-coupling identity in Section 2.2 and (14) yield

$$\int_I \int_{\mathbb{R}^D} \|v^{\gamma_\Gamma}_t(z)\|^2 \, \mathrm{d}\mu_t(z) \, \mathrm{d}t = \mathrm{W}^2_{2,Q}(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

□

B Proof of equivariant flow matching via symmetrization

This section proves Theorem 3.2. Write λ_G for the normalized Haar probability measure on G. We call $\pi \in \mathrm{c}_{\mathbb{R}^D}(\mu, \nu)$ diagonally G-invariant if $(g, g)_\sharp \pi = \pi$ for every $g \in G$.

Proposition B.1 (Equivariance of the flow-matching minimizer)**.** *Let $\pi \in \mathrm{c}_{\mathbb{R}^D}(\mu, \nu)$ be diagonally G-invariant and let v^π denote the minimizer of J_π in (4). Then v^π admits a G-equivariant representative. More precisely, there exists a jointly Borel field $\bar{v} : I \times \mathbb{R}^D \to \mathbb{R}^D$ such that*

$$\bar{v}_t = v_t^\pi \qquad \mu_t\text{-a.e. for a.e. } t \in I,$$

and

$$\bar{v}_t(gz) = g\bar{v}_t(z) \qquad \text{for every } g \in G,\ z \in \mathbb{R}^D,\ t \in I.$$

Proof. Let $(X, Y) \sim \pi$, $Z_t \coloneqq \mathrm{proj}^t(X, Y)$ and $U \coloneqq Y - X$. Diagonal invariance and linearity of the action give

$$(Z_t, U) \overset{\mathrm{d}}{=} (gZ_t, gU) \qquad \text{for every } g \in G.$$

In particular, each μ_t is G-invariant. For $g \in G$, define

$$v_t^{\pi,g}(z) \coloneqq g^{-1} v_t^\pi(gz).$$

Diagonal invariance and orthogonality give $J_\pi(v^{\pi,g}) = J_\pi(v^\pi)$. By uniqueness of the minimizer in $L^2(\,\mathrm{d}t\ \mathrm{d}\mu_t)$,

$$g^{-1} v_t^\pi(g\,\cdot) = v_t^\pi \qquad \mathrm{d}t\ \mathrm{d}\mu_t\text{-a.e.}$$

for every $g \in G$. Choose a jointly Borel representative of v^π and write

$$F(t, z, h) \coloneqq h^{-1} v_t^\pi(hz).$$

The action is continuous, so F is jointly Borel. Orthogonality of h and G-invariance of every μ_t give

$$\begin{aligned}
&\int_I \int_{\mathbb{R}^D} \int_G \|F(t, z, h)\|^2\ \mathrm{d}\lambda_G(h)\ \mathrm{d}\mu_t(z)\ \mathrm{d}t \\
&\quad = \int_G \int_I \int_{\mathbb{R}^D} \|v_t^\pi(hz)\|^2\ \mathrm{d}\mu_t(z)\ \mathrm{d}t\ \mathrm{d}\lambda_G(h) = \|v^\pi\|^2_{L^2(\,\mathrm{d}t\ \mathrm{d}\mu_t)} < \infty.
\end{aligned}$$

By Tonelli's theorem, the inner $L^2(G)$ integral is finite for $\mathrm{d}t\ \mathrm{d}\mu_t$-almost every (t, z). Since λ_G is a probability measure, Cauchy–Schwarz then gives

$$\int_G \|F(t, z, h)\|\ \mathrm{d}\lambda_G(h) < \infty$$

for almost every (t, z). Hence the vector-valued Haar integral exists outside a $\mathrm{d}t\ \mathrm{d}\mu_t$-null set. Define

$$\bar{v}_t(z) \coloneqq \int_G F(t, z, h)\ \mathrm{d}\lambda_G(h)$$

on this set and set it to zero otherwise. Componentwise integration gives a jointly Borel field $\bar{v}$.

It remains to show that this field represents the same L^2 minimizer. Since the preceding equality holds $\mathrm{d}t\ \mathrm{d}\mu_t$-almost everywhere for every fixed $h \in G$, another application of Tonelli's theorem gives

$$\int_I \int_{\mathbb{R}^D} \int_G \left\|F(t, z, h) - v_t^\pi(z)\right\|^2\ \mathrm{d}\lambda_G(h)\ \mathrm{d}\mu_t(z)\ \mathrm{d}t = 0.$$

Consequently, for $\mathrm{d}t\ \mathrm{d}\mu_t$-almost every (t, z), the integrand vanishes for λ_G-almost every h, and hence $\bar{v}_t(z) = v_t^\pi(z)$. Thus the averaging changes only the representative of the $L^2(\,\mathrm{d}t\ \mathrm{d}\mu_t)$ minimizer.

It remains to check that the averaged representative is exactly equivariant. For $k \in G$, right invariance of Haar measure and the change of variables $r = hk$ give

$$\begin{aligned}\bar{v}_t(kz) &= \int_G h^{-1} v_t^\pi(hkz)\ \mathrm{d}\lambda_G(h) \\ &= \int_G kr^{-1} v_t^\pi(rz)\ \mathrm{d}\lambda_G(r) = k\bar{v}_t(z).\end{aligned}$$

The set where the Haar integral is not finite is itself G-invariant by the same change of variables. Therefore the equality also holds there under the zero convention, and $\bar{v}_t(kz) = k\bar{v}_t(z)$ for every $k \in G$, $z \in \mathbb{R}^D$, and $t \in I$.

□

For $\eta \in \mathcal{P}_2(\mathbb{R}^D)$ and $\pi \in \mathrm{c}_{\mathbb{R}^D}(\mu, \nu)$, define

$$\eta^G \coloneqq \int_G g_\sharp \eta\ \mathrm{d}\lambda_G(g), \qquad \pi^G \coloneqq \int_G (g,g)_\sharp \pi\ \mathrm{d}\lambda_G(g),$$

Then

$$\pi^G \in \mathrm{c}_{\mathbb{R}^D}(\mu^G, \nu^G), \qquad (\mathfrak{q},\mathfrak{q})_\sharp \pi^G = (\mathfrak{q},\mathfrak{q})_\sharp \pi.$$

Haar invariance gives $(h,h)_\sharp \pi^G = \pi^G$ for every $h \in G$. Thus diagonal symmetrization preserves the coupling on the quotient, and $\pi^G \in \mathrm{c}_{\mathbb{R}^D}(\mu,\nu)$ whenever both marginals are G-invariant.

Proposition B.2 (Minimization over equivariant fields)**.** *Let $\pi \in \mathrm{c}_{\mathbb{R}^D}(\mu,\nu)$ be arbitrary. Then the field v^{π^G} associated with π^G minimizes J_π over all G-equivariant fields:*

$$v^{\pi^G} \in \operatorname*{arg\,min}_{\substack{v: I\times\mathbb{R}^D \to \mathbb{R}^D \text{ jointly Borel} \\ v_t(gx) = g v_t(x)\ \forall\, t,g,x}} J_\pi(v).$$

In particular, the constrained problem is the ordinary flow-matching problem for the diagonally symmetrized coupling π^G.

Proof. For every G-equivariant field v and every $g \in G$, orthogonality of the action gives

$$J_{(g,g)_\sharp \pi}(v) = \int_I \int_{\mathbb{R}^D \times \mathbb{R}^D} \left\| v_t(g\mathrm{proj}^t(x,y)) - g(y-x) \right\|^2 \mathrm{d}\pi(x,y)\ \mathrm{d}t = J_\pi(v).$$

Averaging over G therefore yields $J_{\pi^G}(v) = J_\pi(v)$ for every equivariant v. Since π^G is diagonally G-invariant, its unrestricted minimizer v^{π^G} admits a G-equivariant representative. As v^{π^G} minimizes J_{π^G} over all fields, it also minimizes J_π over the equivariant ones. □

Proposition B.3 (Equivariant flows on the quotient)**.** *Let $v : I \times \mathbb{R}^D \to \mathbb{R}^D$ be G-equivariant and assume that* (3) *admits a unique flow γ_t. Then γ_t induces the well-defined quotient flow*

$$\bar{\gamma}_t([x]) \coloneqq [\gamma_t(x)].$$

For every $\eta \in \mathcal{P}_2(\mathbb{R}^D)$ and $t \in I$,

$$\mathfrak{q}_\sharp \gamma_{t,\sharp} \eta = \bar{\gamma}_{t,\sharp} \mathfrak{q}_\sharp \eta.$$

In particular, if $\mathfrak{q}_\sharp \eta = \mathfrak{q}_\sharp \widetilde{\eta}$, then

$$\mathfrak{q}_\sharp \gamma_{t,\sharp} \eta = \mathfrak{q}_\sharp \gamma_{t,\sharp} \widetilde{\eta},$$

so the projected curve is independent of the representative lift at time 0*.*

Proof. Equivariance of v and uniqueness of the ODE imply $\gamma_t(gx) = g\gamma_t(x)$. Thus $\bar{\gamma}_t$ is well-defined and $\mathfrak{q} \circ \gamma_t = \bar{\gamma}_t \circ \mathfrak{q}$, which gives

$$\mathfrak{q}_\sharp \gamma_{t,\sharp} \eta = \bar{\gamma}_{t,\sharp} \mathfrak{q}_\sharp \eta.$$

The final claim follows by applying this identity to two measures with the same quotient push-forward. □

The coordinatewise categorical formulation in Section 2.2 is directly compatible with coordinate permutation actions.

Corollary B.4 (Coordinatewise equivariant endpoint prediction)**.** *Assume that G consists of coordinate-permutation matrices, where $g \cdot d$ is determined by $(gx)_{g\cdot d} = x_d$, and let $\mathcal{A} = \{a_1, \dots, a_M\} \subset \mathbb{R}$. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^D)$ be arbitrary, assume that ν is supported on $\mathcal{A}^D$, and let $\pi \in \mathrm{c}_{\mathbb{R}^D}(\mu, \nu)$. Then the conditional coordinate probabilities $q^{*,\pi^G,d}$ associated with π^G admit a representative satisfying*

$$q_t^{*,\pi^G,g\cdot d}(n \mid gz) = q_t^{*,\pi^G,d}(n \mid z)$$

for every $t \in I$, $g \in G$, $d \in [D]$, $n \in [M]$, and $z \in \mathbb{R}^D$, and minimize J_π^{coord} over all such equivariant coordinate kernels:

$$\big(q^{*,\pi^G,d}\big)_{d=1}^D \in \operatorname*{arg\,min}_{\substack{q^d: I\times\mathbb{R}^D \to \Delta_M \text{ jointly Borel} \\ q_t^{g\cdot d}(n|gz) = q_t^d(n|z)}} J_\pi^{\mathrm{coord}}\big((q^d)_{d=1}^D\big).$$

Writing $\mu_t^G := \mathrm{proj}_\sharp^t \pi^G$, for a.e. $t < 1$ and μ_t^G-a.e. z,

$$\big(v_t^{\pi^G}(z)\big)_d = \frac{1}{1-t}\left(\sum_{n=1}^M a_n q_t^{*,\pi^G,d}(n \mid z) - z_d\right).$$

In particular, this defines a G-equivariant representative of v^{π^G}, without requiring μ or ν to be G-invariant.

Proof. Since G consists of coordinate-permutation matrices, it is finite. Diagonal invariance of π^G allows us to choose the conditional coordinate probabilities equivariantly. If $(X, Y) \sim \pi^G$ and $Z_t = \mathrm{proj}^t(X, Y)$, then $(Z_t, Y) \overset{\mathrm{d}}{=} (gZ_t, gY)$ for every $g \in G$. Consequently, starting from any jointly Borel version q, each $q_t^{g\cdot d}(n \mid gz)$ is a version of the same conditional probability. Define

$$\bar{q}_t^d(n \mid z) := \frac{1}{|G|}\sum_{g\in G} q_t^{g\cdot d}(n \mid gz).$$

This is another version of the same conditional probabilities and satisfies the displayed equivariance identity for every g, d, n, t, z. We use this version below.

For every equivariant coordinate kernel q, relabelling the coordinates and summing over $d \in [D]$ gives

$$J^{\mathrm{coord}}_{(g,g)\sharp\pi}(q) = J_\pi^{\mathrm{coord}}(q).$$

Hence $J^{\mathrm{coord}}_{\pi^G}(q) = J_\pi^{\mathrm{coord}}(q)$ on the equivariant class. By (7), these probabilities minimize the unrestricted coordinatewise objective for π^G, and therefore the constrained objective for π.

For any equivariant coordinate kernel q, define

$$\big(v_t^q(z)\big)_d := \frac{1}{1-t}\left(\sum_{n=1}^M a_n q_t^d(n \mid z) - z_d\right).$$

Then

$$\big(v_t^q(gz)\big)_{g\cdot d} = \frac{\sum_{n=1}^M a_n q_t^{g\cdot d}(n \mid gz) - (gz)_{g\cdot d}}{1-t} = \big(v_t^q(z)\big)_d.$$

Thus $v_t^q(gz) = g v_t^q(z)$, independently of the invariance of the endpoint measures. Taking $q^d = q^{*,\pi^G,d}$ for every $d \in [D]$ and using (6) proves the remaining claims. □

Remark B.5 (Categorical feature blocks). The scalar formulation $\mathcal{A} \subset \mathbb{R}$ above is not restrictive for finite categorical data, but a scalar encoding introduces an arbitrary geometry between classes. In practice, we therefore use one-hot encodings and partition the coordinates into categorical feature blocks $b \in \mathcal{B}$. Let block b have K_b possible values $\{e_1, \dots, e_{K_b}\} \subset \mathbb{R}^{K_b}$,

where e_k denotes the k-th standard basis vector. Writing X'_b and $Z_{t,b}$ for the corresponding target and interpolated blocks, respectively, define

$$q_t^{*,b}(k \mid z) \coloneqq \mathbb{P}_\pi(X'_b = e_k \mid Z_t = z), \qquad k \in [K_b].$$

A softmax head estimates $q_t^{*,b}(\cdot \mid z)$ and is trained with the blockwise cross-entropy

$$J_\pi^{\text{block}}(q) \coloneqq -\int_I \int \sum_{b \in \mathcal{B}} \log q_t^b(\kappa_b(x') \mid (1-t)x + tx') \; \mathrm{d}\pi(x, x') \; \mathrm{d}t,$$

where $X'_b = e_{\kappa_b(X')}$. Its conditional endpoint mean is

$$m_t^{*,b}(z) \coloneqq \mathbb{E}_\pi[X'_b \mid Z_t = z] = \sum_{k=1}^{K_b} e_k \, q_t^{*,b}(k \mid z),$$

and hence, for a.e. $t < 1$ and μ_t-a.e. z, the corresponding velocity block is

$$\big(v_t^\pi(z)\big)_b = \frac{m_t^{*,b}(z) - z_b}{1-t} = \frac{1}{1-t}\left(\sum_{k=1}^{K_b} e_k \, q_t^{*,b}(k \mid z) - z_b\right).$$

Thus different feature types may use blocks of different cardinalities. For molecular graphs, for example, edge type, atom type, and formal charge form separate categorical blocks. Node relabeling permutes blocks within each feature type, so the construction applies equivariantly to one-hot node and edge features without requiring conditional independence between blocks.

C Related Work on Transport-Based Graph Generation

Graph generators commonly enforce node-relabeling symmetry through equivariant architectures (Jo et al., 2022; Vignac et al., 2023; Eijkelboom et al., 2024). Recent graph flow models additionally exploit transport geometry. GGFlow uses minibatch OT for graph pairing (Hou et al., 2026), while BWFlow constructs paths from Bures–Wasserstein transport between graph representations (Jiang et al., 2026). Concurrent Flowette introduces a graph-structured graphette prior and uses fused Gromov–Wasserstein distances for structure-aware minibatch pairing (Wijesinghe et al., 2026). Both GGFlow and Flowette therefore use transport primarily to choose which source and target graphs to pair. We instead study transport at both levels of graph alignment. Inner alignment selects node representatives before interpolation, using either a projected GW plan or Best-of-Augmentation (BoA) over sampled permutations, while outer alignment assigns graphs within a minibatch. Our experiments show that inner alignment can provide substantial gains on its own, with BoA effective for small graphs and GW for larger alignment spaces, while the benefit of outer alignment depends on the dataset and compute budget. This connects our study to GW-based graph matching (Peyré et al., 2016; Vayer et al., 2020; Xu et al., 2019), while deriving both alignment levels from transport on the permutation quotient.

D Continuous target via stochastic block models

Here, we aim to provide further insight into the advantage of GW-type alignment via a synthetic experiment with a multimodal target distribution. Therefore, we consider not only typical graph diagnostics but also GW-type two-sample test statistics for our target samples. To further investigate the role of graph alignment, we additionally test a cheap GW surrogate to see if coarse alignment can already lead to improvements.

D.1 Cheap GW Surrogate via Lower Bound

As a cheaper alternative to GW, let $\mathrm{ecc}_E(i) = (N^{-1}\sum_k \|E_{ik}\|^2)^{1/2}$, where $\mathrm{ecc}_E^{\uparrow}(i)$ is the i-th smallest value among eccentricities of E. This gives the first lower bound (FLB) (Bauer et al., 2025)

$$\mathrm{FLB}^2_{2,Z}([E],[F]) \coloneqq \frac{1}{N}\sum_i |\,\mathrm{ecc}_E^{\uparrow}(i) - \mathrm{ecc}_F^{\uparrow}(i)|^2 \le \mathrm{GW}^2_{2,Z}([E],[F]).$$

This is the squared 2-Wasserstein distance between the empirical eccentricity distributions. We sort the nodes of E and F by their eccentricities and pair the i-th node in one ordering with the i-th node in the other, breaking ties arbitrarily. These pairs define a permutation. Computing the eccentricities takes $O(N^2C)$ operations, and sorting takes $O(N \log N)$. Thus GW retains detailed pairwise structure at higher computational cost, whereas FLB provides an inexpensive inner alignment and is cheap to evaluate for all B^2 candidate pairs in the outer minibatch assignment. Their empirical cost–quality trade-off is reported in Appendix I. Again, we can use the FLB not only for inner alignment of nodes (FLB), but we can also reorder the matching of source and target via FLB as a cost (FLB+$\mathrm{FLB}^B_{\mathrm{out}}$). Again, FLB+$\mathrm{FLB}^8_{\mathrm{out}}$ reorders same-size graphs in sub-batches of at most eight.

D.2 Unconditional continuous-SBM experiment

Data As a controlled continuous testbed, we use an SBM on $N = 10$ nodes with two channels: one symmetric weighted-edge channel in $[0, 1]$ and one scalar node feature on the diagonal. For $K \in \{1, \dots, 5\}$, nodes are assigned to near-balanced blocks; for every $i < j$, we sample and mirror $e^E_{ij} = e^E_{ji}$ from $\mathrm{Beta}(6, 2)$ within a block and $\mathrm{Beta}(2, 6)$ across blocks. A node in block $k \in \{0, \dots, K-1\}$ receives a Beta feature with concentration 8 and mean $\mu_k = (2k+1)/(2K)$, so the diagonal channel also records community structure. We randomize node order, making each tensor an arbitrary representative of its graph orbit. The source has independent $\mathcal{U}[0, 1]$ upper-triangular edges, mirrored across the diagonal, and independent $\mathcal{U}[0, 1]$ node features. Our main experiment pools all K. Appendix D.3 instead supplies K to the model to showcase conditional generation.

Metrics For 100 real and 100 generated graphs, we estimate maximum mean discrepancies (MMDs) of degree, clustering, and four-node graphlet-orbit counts, which distinguish the roles a node can occupy in induced four-node subgraphs (Jo et al., 2022; You et al., 2018). All three require unweighted graphs, obtained by thresholding real and generated edges alike at $e^E_{ij} > 1/2$. To compare the full attributed graphs including their node features, we further report fused-GW nearest-neighbor accuracy (FGW–NNA). This statistic builds on the popular OT-NNA statistic used to compare real and generated sets of point clouds (Yang et al., 2019) by replacing pairwise Wasserstein distance with pairwise FGW. Given 100 generated and 100 real graphs, we pool both samples and label each graph by its origin. We then classify every graph by the label of its nearest neighbor under fused GW, excluding the graph itself. Because fused GW compares attributed graphs up to node relabelling, this is a two-sample test directly on graph orbits rather than on selected descriptors. If real and generated graphs follow the same law, the accuracy approaches 0.5. Larger values mean that the samples are more easily separated. Results are reported over three training runs with different seeds and evaluated at 5/25/125 Euler steps.

Results At five steps (Table 4), GW with outer assignment reduces FGW–NNA from 0.796 to 0.568, degree MMD from 0.114 to 0.018, and clustering MMD from 0.179 to 0.034. The gap narrows by 125 steps, as expected because all matching should result in valid

interpolations. FLB captures part of the low-step gain at much lower cost, while outer assignment adds a smaller improvement. The permutation-blind MBOT already matches or exceeds the FLB variants on most statistics, yet remains clearly behind the GW variants. The conditional experiment likewise favors GW alignment (Appendix D.3). Notably, at five steps the two GW variants are also the only conditions that bring FGW–NNA close to its ideal value of 0.5. Thus, the improvement is visible not merely in selected graph statistics, but in a two-sample test using the underlying relational geometry.

Table 4: Continuous SBM, *unconditional* mixture over $K \in \{1, \ldots, 5\}$ ($N = 10$), at 5/25/125 Euler steps. Rows are mean_{std} over 3 training seeds (5 evaluation repeats each). Best value per column is bold (MMDs (degree, clustering, graphlet orbit) ↓, Best FGW–NNA: ~ 0.5).

	FGW–NNA (→ 0.5)			Degree MMD (↓)			Clustering MMD (↓)			Graphlet-orbit MMD (↓)		
Method	5	25	125	5	25	125	5	25	125	5	25	125
RANDOM	$0.796_{.012}$	$0.581_{.021}$	$0.561_{.015}$	$0.114_{.045}$	$0.067_{.016}$	$0.059_{.011}$	$0.179_{.100}$	$0.085_{.042}$	$0.075_{.028}$	$0.062_{.033}$	$0.046_{.033}$	$0.043_{.025}$
MBOT[16]	$0.687_{.018}$	$0.561_{.016}$	$0.530_{.007}$	$0.031_{.004}$	$0.023_{.006}$	$0.015_{.005}$	$0.090_{.014}$	$0.046_{.010}$	$0.037_{.010}$	$0.006_{.001}$	$0.011_{.001}$	$0.007_{.000}$
FLB	$0.725_{.013}$	$0.558_{.011}$	$0.555_{.022}$	$0.054_{.004}$	$0.037_{.011}$	$0.030_{.010}$	$0.109_{.037}$	$0.042_{.006}$	$0.040_{.011}$	$0.030_{.014}$	$0.027_{.017}$	$0.021_{.015}$
FLB+FLB$^{\text{s}}_{\text{out}}$	$0.691_{.040}$	$0.562_{.019}$	$0.547_{.009}$	$0.034_{.011}$	$\mathbf{0.012}_{.002}$	$0.013_{.007}$	$0.121_{.041}$	$0.045_{.024}$	$0.047_{.021}$	$0.018_{.011}$	$\mathbf{0.007}_{.002}$	$0.010_{.005}$
GW	$0.577_{.011}$	$\mathbf{0.522}_{.004}$	$0.520_{.023}$	$0.027_{.013}$	$0.014_{.002}$	$0.012_{.004}$	$0.049_{.019}$	$\mathbf{0.015}_{.005}$	$0.016_{.007}$	$0.014_{.004}$	$0.010_{.003}$	$0.008_{.001}$
GW+GW$^{\text{s}}_{\text{out}}$	$\mathbf{0.568}_{.008}$	$0.536_{.014}$	$\mathbf{0.504}_{.021}$	$\mathbf{0.018}_{.005}$	$0.014_{.002}$	$\mathbf{0.007}_{.001}$	$\mathbf{0.034}_{.007}$	$0.021_{.004}$	$\mathbf{0.014}_{.002}$	$\mathbf{0.005}_{.001}$	$0.008_{.001}$	$\mathbf{0.004}_{.000}$

D.3 Conditional continuous-SBM experiment

We additionally condition the continuous SBM model on the community count K, form minibatches containing only one value of K, and compare pooled generated and real graphs with matching K-multisets. All other settings and metrics agree with Section D.2. The conditional results in Table 5 reproduce the unconditional ranking: GW alignment has the largest advantage at five Euler steps, and the gap narrows with the integration budget.

Table 5: Continuous SBM, *class-conditional* on $K \in \{1, \ldots, 5\}$ ($N = 10$), evaluated by pooling generated and real graphs with the same ground-truth K multiset. Mean ± std over 3 training seeds (5 evaluation repeats each) at 5/25/125 Euler steps. Descriptor MMDs: lower is better, FGW–NNA: closest to 0.5 is better. Best per column in bold.

	FGW–NNA (→ 0.5)			Degree MMD (↓)			Clustering MMD (↓)			Graphlet-orbit MMD (↓)		
Method	5	25	125	5	25	125	5	25	125	5	25	125
RANDOM	$0.746_{.009}$	$0.580_{.008}$	$0.544_{.021}$	$0.043_{.009}$	$0.008_{.002}$	$0.005_{.001}$	$0.076_{.010}$	$0.029_{.007}$	$0.020_{.004}$	$0.028_{.002}$	$0.006_{.004}$	$0.003_{.003}$
FLB	$0.714_{.010}$	$0.543_{.017}$	$0.558_{.023}$	$0.023_{.004}$	$0.005_{.001}$	$0.005_{.001}$	$0.047_{.006}$	$0.021_{.003}$	$0.017_{.003}$	$0.018_{.003}$	$0.005_{.001}$	$0.003_{.002}$
FLB+FLB$^{\text{s}}_{\text{out}}$	$0.699_{.007}$	$0.552_{.018}$	$0.539_{.007}$	$0.013_{.002}$	$0.007_{.002}$	$0.007_{.001}$	$0.049_{.009}$	$0.030_{.005}$	$0.027_{.010}$	$\mathbf{0.009}_{.006}$	$0.005_{.003}$	$0.005_{.004}$
GW	$0.584_{.002}$	$\mathbf{0.523}_{.010}$	$\mathbf{0.524}_{.003}$	$0.011_{.001}$	$\mathbf{0.005}_{.001}$	$\mathbf{0.004}_{.001}$	$\mathbf{0.016}_{.002}$	$\mathbf{0.011}_{.001}$	$\mathbf{0.010}_{.001}$	$0.016_{.001}$	$0.005_{.002}$	$0.003_{.001}$
GW+GW$^{\text{s}}_{\text{out}}$	$\mathbf{0.561}_{.017}$	$0.526_{.028}$	$0.529_{.014}$	$\mathbf{0.008}_{.001}$	$0.005_{.001}$	$0.005_{.002}$	$0.018_{.001}$	$0.014_{.001}$	$0.014_{.004}$	$0.012_{.001}$	$\mathbf{0.004}_{.001}$	$\mathbf{0.003}_{.001}$

E Full-budget molecular training protocol

Protocol For the full-budget evaluation, we retain the categorical endpoint objective and use GW inner alignment with independent outer pairing. We increase the model capacities and training budgets toward those used by DeFoG (Qin et al., 2025). In particular, we augment the transformer inputs with structural graph statistics and relative random-walk probabilities (RRWP), and use self-conditioning. For our self-conditioning implementation, on 50% of optimization steps, the model is conditioned on a detached preliminary endpoint prediction (Chen et al., 2023). During sampling, each Euler step receives the preceding step's endpoint prediction. As regularization during training, Gaussian noise with standard deviation $0.5\,t(1-t)$ is added to the interpolated network input while leaving the endpoint target unchanged.

The categorical objective consists of a bond cross-entropy term and a node-feature term, the latter averaging the atom-type and formal-charge cross-entropies. These terms are weighted equally for QM9. For ZINC250k, we multiply the bond term by 5 relative to the node-feature term, following the stronger emphasis on edge prediction used by DeFoG and CatFlow.

Both QM9 and ZINC250k are trained in two stages. We first train without dropout, then restore the model and EMA weights and fine-tune with dropout 0.1 using a new AdamW optimizer. All reported full-budget numbers in this section use the final EMA checkpoint after dropout fine-tuning. The dropout-free checkpoints are only used to initialize the second stage.

The QM9 model has node, edge, and global widths 256/64/64, feed-forward widths 256/128/128, depth 9, 8 attention heads, and 12 RRWP channels, for approximately 6.0M parameters. It is first trained for 1000 epochs with batch size 1024 without dropout, and then fine-tuned for 50 additional epochs with dropout 0.1. The ZINC250k model has widths 256/64/128, feed-forward widths 256/128/256, depth 12, 8 heads, and 20 RRWP channels, for approximately 10.8M parameters. It is first trained for 300 epochs using microbatches of 128 with two-step gradient accumulation, giving an effective batch size of 256, and then fine-tuned for 30 additional epochs with dropout 0.1. The initial training runs use AdamW with learning rate 2×10^{-4}, weight decay 10^{-4}, gradient-norm clipping at 1.0, cosine learning-rate decay, and EMA decay 0.999.

The dropout fine-tuning stages restore both model and EMA weights, use a new AdamW optimizer at learning rate 5×10^{-5}, and do not use cosine decay. We report only the final fine-tuned EMA checkpoints, evaluated using 500 Euler steps.

ZINC250k ablation. To isolate the contribution of the main additions used in the full-budget ZINC model, we run the cumulative ablation in Table 6 on ZINC250k using the smaller 100-epoch architecture from the base molecular experiments. This model has node, edge, and global widths 128/64/128, depth 6, 8 attention heads, and about 2.8M parameters. All rows use GW inner alignment, independent outer pairing, formal-charge features, the categorical endpoint objective, batch size 32, and are trained for 100 epochs. We evaluate final EMA checkpoints with the standard molecular protocol using 10,000 samples and 3 repeats, but only at 25 Euler steps. The expanded computational budget for the full run may contribute to the remaining performance gap.

Table 6: Cumulative ZINC250k ablation at 25 Euler steps and reduced training budget. Higher validity is better and lower FCD is better.

Setting	Validity	FCD
Base + RRWP	0.8846 ± 0.0030	12.3371 ± 0.1229
+ latent noise	0.9080 ± 0.0012	11.2470 ± 0.0704
+ self-conditioning	0.9307 ± 0.0010	8.8572 ± 0.0763
+ edge weight 5	0.9606 ± 0.0019	6.9861 ± 0.0231

F Reproducibility details

This appendix collects the architecture, data, noise, solver, and metric specifications needed to reproduce the experiments of Section 5. All coupling conditions within an experiment share every model and optimization setting listed here and differ only in their training coupling, namely the inner solver $\widehat{\sigma}_a$ and outer reordering $\widehat{\tau}$ of Section 4, with MBOT replacing both by its permutation-blind minibatch assignment.

Architecture and optimization Every model builds on the same G_N^{graph}-equivariant graph transformer with node (X), edge (E), and global (y) streams, following the XEy architecture of Vignac et al. (2023). Its node, edge, and global widths are $d_X = 128$, $d_E = 64$, and $d_y = 128$, respectively, with feed-forward widths 256/128/256, depth 6, 8 attention heads, and $\sim$2.8M parameters. Training uses AdamW (weight decay 10^{-4}), gradient-norm clipping at 1.0, a cosine learning-rate schedule with base rate 2×10^{-4}, and an exponential moving average of the weights with decay 0.999. The EMA weights are used for all evaluation and checkpointing. The per-experiment channel count C, node count N, batch size, epoch budget, and training head are summarized in Table 7. The SBM models use an MSE velocity loss (11). The molecular models use blockwise categorical endpoint heads with the cross-entropy objective (12) and recover the velocity from blockwise conditional means as described in Appendix B. All experiments use at most a single NVIDIA GeForce RTX 5090 GPU with 32 GB of VRAM.

Few-step distillation Each flow-map student uses the backbone above with an additional start-time embedding, passed through its own linear layer and added to the global-stream initialization (8,320 parameters). This layer is zero-initialized and all other weights are copied from the frozen EMA teacher, so the student initially coincides with the teacher. Pairs (x, x') are drawn independently for every coupling condition, so conditions differ only through the teacher. ρ_0 is the logit-normal distribution LogitNormal$(-0.4, 1)$, i.e. the law of $1/(1 + e^{-\xi})$ with $\xi \sim \mathcal{N}(-0.4, 1)$. Under ρ, $t \sim \rho_0$, $s \mid t \sim \mathcal{U}(0, t)$, and $u = (s + t)/2$, except that a quarter of the samples use $(s, t) = (0, 1)$. Each minibatch is split 75%/25% between the diagonal and composition terms of (8). Each term is averaged over its part and the two are summed, with KL terms implemented as soft cross-entropies. Training runs for 100 epochs with AdamW (learning rate 10^{-4}, weight decay 10^{-12}), gradient-norm clipping at 1.0, a cosine schedule, and EMA 0.999, at batch size 512 (QM9) and 256 (ZINC250k). Sampling uses the uniform grid $t_k = k/K$ with $K \in \{1, 2, 5\}$.

Cycle-graph data Source and target are both distributions of cycle graphs on $N = 12$ nodes. A sample places twelve nodes at equispaced angles, offset by a rotation drawn uniformly from $[0, 2\pi)$, on the circle of radius 1 centred at (h, y), and joins two nodes exactly when they are neighbours along that circle. Nodes are then relabelled by a uniformly random permutation, so the node index carries no information about the circular order. The tensor $E \in \mathbb{R}^{12\times12\times3}$ has adjacency in $\{0, 1\}$ as its first channel, while the remaining two channels store the planar node positions on the diagonal, $E_{ii,2:3} \in \mathbb{R}^2$. Positions are stored divided by 8, which puts the position and adjacency channels on comparable scales in both the GW

Table 7: Per-experiment settings for the controlled coupling comparisons, shared across all coupling conditions. Architecture, optimizer, EMA, gradient clipping, and learning-rate schedule are identical throughout (see text), where only the entries below vary. The larger-model protocol is specified separately in Appendix E. "Head" is the flow-matching parametrization: MSE on the velocity field, or cross-entropy on the predicted categorical endpoint.

	Cycle graphs	Continuous SBM	QM9	ZINC250k
Channels C	3	2	11	16
Nodes N	12	10	≤ 9	≤ 38
Gradient batch ($\neq B$)	16	16	64	32
Epochs	≈ 819 (204,800 steps)	1000	100	100
Time embedding dim	32	32	64	64
Head	velocity (MSE)	velocity (MSE)	endpoint (cross-entropy)	endpoint (cross-entropy)

cost and the MSE loss. The source takes $y = 0$ and the target $y = 8$, with $h \sim \mathcal{U}[-6, 6]$ drawn independently on each side. The two distributions therefore agree up to a translation and have identical edge distributions. To give the source a density, each training source is perturbed after coupling by small symmetric Gaussian noise ($\sigma = 0.01$) on all adjacency entries and node positions. We draw 4000 target graphs and a fresh source per pair, use $\lambda_{\text{edge}} = \lambda_{\text{node}} = 0.5$, and train for 204,800 steps at batch size 16 with the velocity (MSE) head. GW_{out} reorders chunks of 16 pairs. For evaluation, we integrate 125 Euler steps from 1024 exact source rings.

Continuous SBM data A graph on $N = 10$ nodes with K communities assigns the nodes to fixed near-balanced blocks (remainder placed in the first blocks, e.g. $K = 3 \to 4/3/3$) in a randomized order. Parametrizing $\text{Beta}(a, b)$ by its mean μ and concentration s via $a = s\mu$, $b = s(1-\mu)$, every unordered pair $i < k$ carries an edge weight $\text{Beta}(s_e\mu_{\text{in}}, s_e(1-\mu_{\text{in}}))$ when i, k share a block and $\text{Beta}(s_e\mu_{\text{out}}, s_e(1-\mu_{\text{out}}))$ otherwise, with $\mu_{\text{in}} = 0.75$, $\mu_{\text{out}} = 0.25$, and $s_e = 8$ (i.e. $\text{Beta}(6, 2)$ and $\text{Beta}(2, 6)$). The diagonal node feature of a node in block k is $\text{Beta}(s_n\mu_k, s_n(1-\mu_k))$ with evenly spaced means $\mu_k = (2k+1)/(2K)$ and $s_n = 8$. For $K = 2$ this reproduces the binary SBM feature model exactly. In the unconditional variant graphs with $K \in \{1, \dots, 5\}$ are pooled and in the conditional variant K is the class label. In the unconditional variant, 2000 graphs are drawn for each $K \in \{1, \dots, 5\}$ and pooled (10 000 in total). In the conditional variant, the same per-class sets are used, with K supplied to the model as the class label.

Molecular data Molecules are encoded from their RDKit representation after kekulization, so bond orders are integer-valued. A molecule on N heavy atoms becomes a tensor $E \in \mathbb{R}^{N\times N\times C}$ whose edge channel is a one-hot over $\{\text{none}, \text{single}, \text{double}, \text{triple}\}$ and whose diagonal node feature concatenates an atom-type one-hot with a formal-charge one-hot over $\{-1, 0, +1\}$. QM9 uses heavy atoms $N \le 9$ over $\{\text{C}, \text{N}, \text{O}, \text{F}\}$, giving $C = 11$. ZINC250k uses $N \le 38$ over $\{\text{C}, \text{N}, \text{O}, \text{F}, \text{Br}, \text{Cl}, \text{I}, \text{P}, \text{S}\}$, giving $C = 16$. For ZINC250k the 28-way PyTorch Geometric atom vocabulary—which encodes charged species—is remapped to the nine element types listed above and to a signed formal charge in $\{-1, 0, +1\}$. 31.8% of ZINC250k molecules carry a formal charge, and modelling it explicitly is required for those atoms to pass valence on decode. QM9 uses the GDSS train/test split, matched on raw QM9 record identifiers (`valid_idx_qm9.json`). ZINC250k uses the standard PyTorch Geometric split with 220,011 training, 24,445 validation, and 5,000 test molecules. FCD is evaluated against the respective test split.

Variable node count For the molecular datasets, one set of graph-transformer parameters is shared across all node counts. An input with N nodes produces an output with N nodes. Source noise for each target is drawn at the target's own node count, so no padding or masking is used. During training, minibatches are formed by a sampler that groups dataset indices by exact N and yields same-N batches. This is required because the outer cost matrices of Section 4 compare same-size tensors. In the conditional SBM variant, each batch additionally contains only graphs with a single value of K, so graphs with different community counts are never coupled. At generation time N is sampled from the empirical training-set node-count distribution and the Euler solver is initialized from noise of shape (N, N, C). Euler sampling evaluates the field at $t_k = k/K$, $k = 0, \dots, K-1$, and steps to $t = 1$, so the network is not evaluated at the singular endpoint. The final tensor is decoded by argmax within each categorical block.

Source noise For the molecular datasets the source is a symmetric Gaussian. For each edge channel and each $i < j$, we draw $U_{ij}, U_{ji} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\text{adj}})$ and set $A_{ij} = A_{ji} = \frac{U_{ij}+U_{ji}}{\sqrt{2}}$, so that each off-diagonal edge entry satisfies $A_{ij} \sim \mathcal{N}(0, \sigma^2_{\text{adj}})$. The edge diagonal is zero. Each diagonal node feature is drawn independently from $\mathcal{N}(0, \sigma^2_{\text{node}})$, with $\sigma_{\text{adj}} = \sigma_{\text{node}} = 0.5$. For the continuous SBM the source is independent $\mathcal{U}[0, 1]$ on both channels. The upper-triangular edge entries are mirrored across the diagonal, and the diagonal node features are drawn independently. This matches the $(0, 1)$ scale of the dense Beta target. The edge diagonal is again zero.

Compute scaling For each fixed node count ($N \in \{8, 9\}$ for QM9, $N \in \{18, 23\}$ for ZINC250k) we draw $M = 32{,}768$ source–target pairs once: targets are training graphs with N nodes (with replacement where fewer than M exist), each uniformly relabeled, and sources

are drawn from the training source law at the same N. All methods couple this identical set of pairs. For every budget $AB \in \{2^5, \dots, 2^{15}\}$ and every power-of-two split (A, B), the M pairs are partitioned into M/B disjoint minibatches of B pairs. Within each minibatch we evaluate $\hat{c}_A(E^a, F^b; \Sigma_A)$ for all B^2 combinations, solve the outer assignment $\hat{\tau}$ by the Hungarian algorithm, and keep for every assigned target its minimizing permutation $\hat{\sigma}_A$. Thus $B = 1$ is inner BoA^A alone (pairing unchanged), $A = 1$ is MBOT^B, $A = B = 1$ is Random, and $B = M$ is a single assignment over all pairs. Here Σ_A consists of the identity and $A - 1$ distinct further permutations drawn uniformly without replacement for QM9, so that at large A no budget is spent on repeated candidates (for QM9 with $N = 8$, $A = 2^{15}$ covers 81% of S_8). For ZINC, we simply sample with replacement due to the negligible probability of sampling the same permutation twice. Each point reports the mean aligned distance over all M pairs and the single-threaded CPU wall-clock time per coupled pair, including permutation sampling, cost evaluation and assignment. All timings, including GW, are measured on the same machine. The simple GW reference applies the 10-iteration Frank–Wolfe solver with Hungarian rounding to the same M pairs with $\hat{\tau} = \mathrm{id}$. Unlike joint inner-outer GW alignment, its per-pair cost and time do not depend on B. Outer GW assignment approximates the cost between two graphs via GW approximation with 5 iterations.

Coupling solvers All fused solvers use the combined graph entries from Section 4, whose squared norms are

$$\|E_{ik}\|^2 = \lambda_{\mathrm{edge}} \|e^E_{ik}\|^2 + \tfrac{\lambda_{\mathrm{node}}}{C_{\mathrm{v}}} \|f^E_i\|^2 \mathbf{1}_{\{i=k\}}, \qquad \lambda_{\mathrm{edge}} = \lambda_{\mathrm{node}} = \tfrac{1}{2},$$

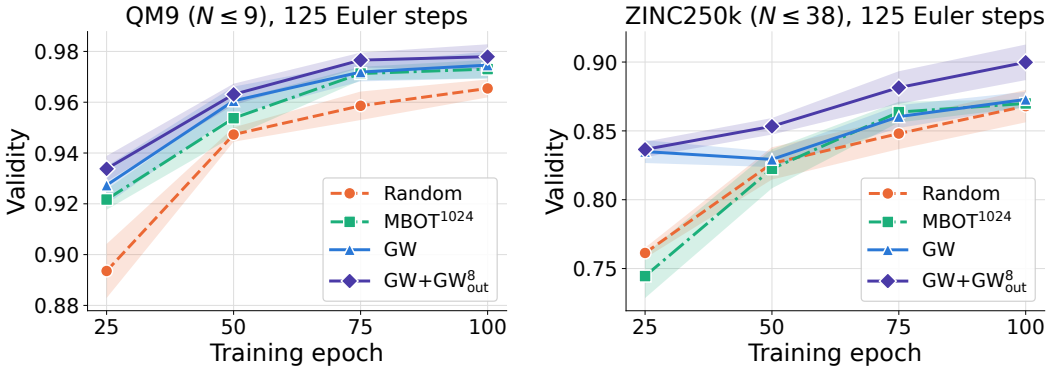
where node features enter only on the diagonal. The inner Gromov–Wasserstein solver (GW) runs 10 Frank–Wolfe iterations, each linearized by an exact optimal-transport (Hungarian) step, and the final soft plan is projected to the Frobenius-nearest scaled permutation matrix using the Hungarian assignment in Section 4. When GW is used for outer batch matching, we compute the pairwise batch-cost matrix with 5 Frank–Wolfe iterations before applying the Hungarian assignment over the batch. BoA^A relabels each target uniformly at random and keeps, among this labeling and $A-1$ further random permutations, the one with the smallest squared Euclidean distance of the combined entries. For BoA^A+$\mathrm{BoA}^B_{\mathrm{out}}$, this best-of-$A$ cost is computed for all B^2 pairs of a same-N pool, the Hungarian assignment is solved on it, and each assigned target keeps its minimizing permutation ($A = 1$ recovers MBOT^B). Pools larger than the gradient batch are assigned once and then split into gradient batches, which leaves the number of steps per epoch unchanged. During training, permutations are drawn i.i.d. with replacement: for $A \le 32$ the expected fraction of repeated candidates is at most $(A-1)/(2\,N!)$, below 0.05% for $N \ge 8$. The first lower bound (FLB) ranks nodes by the root-mean-square eccentricity $\mathrm{ecc}_E(i) = (N^{-1} \sum_k \|E_{ik}\|^2)^{1/2}$ defined in Appendix D. It sorts the nodes in each graph by this value and pairs nodes at the same position in the two orderings. Computing the eccentricities and sorting costs $O(N^2 C + N \log N)$ per pair. For outer reordering with FLB, each same-N training batch is partitioned into sub-batches of at most eight graphs like GW reordering. Within each sub-batch, we build the pairwise FLB cost matrix and apply the Hungarian algorithm to permute the target graphs.

Evaluation metrics For cycle graphs, roundness is $\sigma_r = \mathrm{std}_i \|p_i - \bar{p}\|$, the standard deviation of the node distances to their centroid $\bar{p} = \frac{1}{N} \sum_i p_i$ (reported $\times 10^2$, 0 for an exact circle). The positions p_i are in unscaled plot units, i.e. the stored values multiplied back by 8. Mean degree is $\frac{1}{N} \sum_{i \neq j} A_{ij}$ on the raw, symmetrized continuous adjacency, which equals 2 for an exact 12-cycle. It does not round the adjacency, so partially present edges count with their fractional value. For molecules, validity is the fraction of generated graphs that decode to a sanitizable RDKit molecule without valence correction or other postprocessing. For FCD, we retain the largest connected fragment of each sanitizable generated molecule, canonicalize it, remove duplicates, and compare the ChemNet embeddings of these unique generated SMILES with the corresponding unique test-split SMILES. The descriptor MMDs follow the GraphRNN/GDSS conventions (Jo et al., 2022; You et al., 2018) with the biased estimator and kernels of the form $k(x, y) = \exp[-d(x, y)^2/(2\sigma^2)]$. Degree uses earth mover's distance (EMD) with $\sigma = 1.0$ on per-graph degree histograms. Clustering uses EMD over 100 bins on $[0, 1]$ with $\sigma = 10$ in raw-bin units (equivalently $\sigma = 0.1$ in the $[0, 1]$-normalized EMD units of GraphRNN/GDSS). Graphlet orbit uses Euclidean distance in a Gaussian radial basis function kernel with $\sigma = 30.0$ on the 15-dimensional, per-node-averaged four-node

graphlet-orbit counts computed with the ORCA graphlet-counting program (Hočevar and Demšar, 2014). FGW–NNA pools the n generated and n real graphs and classifies each by the label of its nearest neighbor (self excluded) under the fused Gromov–Wasserstein distance with $\lambda_{\text{edge}} = \lambda_{\text{node}} = \frac{1}{2}$. Under the null this accuracy tends to $1/2$.

G Training convergence of the molecular sweep

Table 2a reports only the final checkpoints. We additionally evaluate the EMA checkpoints every 25 epochs, using 5 sets of 1024 molecules at 125 Euler steps. Seeds are matched across couplings and epochs. We use 125 steps to reduce integration error and isolate training progress. The epoch-100 results agree with Table 2a up to the smaller evaluation sample.



(a) QM9 ($N \leq 9$)

(b) ZINC250k ($N \leq 38$)

Figure 3: Validity during training at 125 Euler steps, averaged over 5 repeats of 1024 molecules. The shaded bands show ±1 standard deviation.

Figure 3 shows that GW alignment accelerates early training: validity differences between couplings are largest at the first evaluated checkpoint and much smaller by the end of training. On QM9, the GW+GW$_{\text{out}}$ advantage over Random likewise shrinks substantially. This hints at faster convergence with aligned couplings, presumably because consistent node correspondences reduce the variability in the loss.

H Uniqueness

Uniqueness is the fraction of distinct largest-fragment canonical SMILES among the valid generated molecules, which we can again evaluate on 10,000 generated samples per repeat, averaged over 3 repeats. However, uniqueness (Table 8) does not separate the couplings. Essentially, it is saturated on ZINC250k, and on QM9 the only visible effect is a slight GW advantage at the lowest Euler step count that vanishes at higher step counts and is within noise for the flow maps. Because uniqueness only counts distinct valid SMILES, it ignores whether a sample is chemically sensible, unlike validity or FCD, so it cannot serve as evidence of sample quality. What it does show is that the gains of GW alignment in validity and FCD are not bought at the expense of diversity.

Table 8: Molecular short-budget sweep, uniqueness (same runs and evaluations as Table 2): (8a) Flow ODE at 5/25/125 Euler steps and (8b) flow map at 1/2/5 steps. Rows are mean_{std} over 3 evaluation repeats. Best is bold (up to $\pm 2\sigma$ of evaluation variability).

(a) Euler Integration of Learned Flow ODE

	Uniqueness (↑)		
Method	5	25	125
QM9 ($N \leq 9$)			
Random	$0.9342_{.0036}$	$\mathbf{0.9678}_{.0027}$	$\mathbf{0.9713}_{.0010}$
MBOT^{1024}	$0.9426_{.0028}$	$0.9647_{.0011}$	$0.9661_{.0007}$
BoA^{4}+BoA$^{256}_{\text{out}}$	$0.9464_{.0019}$	$\mathbf{0.9668}_{.0016}$	$0.9679_{.0007}$
GW	$0.9531_{.0009}$	$0.9650_{.0010}$	$0.9685_{.0006}$
GW+GW$^{8}_{\text{out}}$	$\mathbf{0.9585}_{.0022}$	$\mathbf{0.9684}_{.0007}$	$0.9685_{.0005}$
ZINC250k ($N \leq 38$)			
Random	$\mathbf{0.9999}_{.0001}$	$\mathbf{1.0000}_{.0001}$	$\mathbf{1.0000}_{.0001}$
MBOT^{1024}	$\mathbf{0.9998}_{.0002}$	$\mathbf{1.0000}_{.0000}$	$\mathbf{0.9999}_{.0001}$
BoA^{2}+BoA$^{512}_{\text{out}}$	$\mathbf{0.9999}_{.0001}$	$\mathbf{1.0000}_{.0001}$	$\mathbf{0.9999}_{.0001}$
GW	$\mathbf{1.0000}_{.0000}$	$\mathbf{1.0000}_{.0001}$	$\mathbf{0.9999}_{.0001}$
GW+GW$^{8}_{\text{out}}$	$\mathbf{1.0000}_{.0000}$	$0.9997_{.0001}$	$\mathbf{0.9999}_{.0000}$

(b) Distilled Flow Map for Few-Step Generation

	Uniqueness (↑)		
Method	1	2	5
QM9 ($N \leq 9$)			
Random	$\mathbf{0.9641}_{.0015}$	$\mathbf{0.9686}_{.0011}$	$\mathbf{0.9702}_{.0015}$
MBOT^{1024}	$0.9613_{.0009}$	$0.9620_{.0011}$	$\mathbf{0.9662}_{.0022}$
BoA^{4}+BoA$^{256}_{\text{out}}$	$\mathbf{0.9643}_{.0015}$	$0.9624_{.0019}$	$\mathbf{0.9689}_{.0007}$
GW	$\mathbf{0.9663}_{.0020}$	$\mathbf{0.9667}_{.0013}$	$\mathbf{0.9685}_{.0011}$
GW+GW$^{8}_{\text{out}}$	$\mathbf{0.9639}_{.0010}$	$\mathbf{0.9650}_{.0019}$	$\mathbf{0.9710}_{.0014}$
ZINC250k ($N \leq 38$)			
Random	$\mathbf{1.0000}_{.0000}$	$\mathbf{0.9999}_{.0001}$	$\mathbf{0.9999}_{.0001}$
MBOT^{1024}	$\mathbf{0.9999}_{.0001}$	$\mathbf{1.0000}_{.0000}$	$\mathbf{1.0000}_{.0000}$
BoA^{2}+BoA$^{512}_{\text{out}}$	$\mathbf{1.0000}_{.0000}$	$\mathbf{1.0000}_{.0000}$	$\mathbf{1.0000}_{.0001}$
GW	$\mathbf{1.0000}_{.0000}$	$\mathbf{1.0000}_{.0000}$	$\mathbf{0.9999}_{.0001}$
GW+GW$^{8}_{\text{out}}$	$\mathbf{1.0000}_{.0000}$	$\mathbf{0.9999}_{.0001}$	$\mathbf{0.9997}_{.0003}$

I Solver cost and correlation with GW

Table 9 compares GW (10 Frank–Wolfe iterations) with cheaper pair costs on the pairs seen in training. These are FLB, the Euclidean distance $\|E - F\|^2$ used by MBOT, and BoA$^{\mathcal{A}}$ with $\mathcal{A} = 32$ and $\mathcal{A} = 1024$. For each dataset we draw 300 pairs of a source sample and a training graph with fixed N ($N = 10$ for the continuous SBM, $N = 9$ for QM9, $N = 23$ for ZINC250k). Sources follow the training source law, uniform for the continuous SBM and Gaussian for the molecules. We set $\lambda_{\text{edge}} = \lambda_{\text{node}} = \frac{1}{2}$.

Table 9: Runtime in ms per pair (CPU, single thread) and correlation with the GW cost, Pearson (Spearman), over 300 pairs of a training source sample and a same-N training graph. Euclidean is the MBOT cost $\|E - F\|^2$ and BoA$^{\mathcal{A}}$ its minimum over $\mathcal{A}$ random relabellings.

	Continuous SBM $N = 10$		QM9 $N = 9$		ZINC250k $N = 23$	
Method	ms/pair	Corr. to GW	ms/pair	Corr. to GW	ms/pair	Corr. to GW
GW	1.30	—	2.21	—	2.66	—
FLB	0.053	0.64 (0.56)	0.061	−0.38 (−0.34)	0.065	−0.45 (−0.46)
Euclidean	0.004	0.42 (0.38)	0.004	0.77 (0.77)	0.006	0.90 (0.89)
BoA^{32}	0.036	0.72 (0.67)	0.043	0.91 (0.91)	0.16	0.95 (0.94)
BoA^{1024}	0.44	0.80 (0.74)	0.60	0.94 (0.94)	5.61	0.96 (0.96)

FLB is 24 to 41 times cheaper than GW, and the Euclidean distance more than 100 times. BoA$^{\mathcal{A}}$ evaluates $\mathcal{A}$ Euclidean distances, so its cost grows linearly in $\mathcal{A}$. On all datasets BoA^{32} correlates better with GW than the Euclidean distance, and BoA^{1024} better still, but it is slower than GW on ZINC250k. The gain from relabelling is largest on the continuous SBM. FLB is negatively correlated with GW on molecules, because the one-hot targets have constant eccentricities, so FLB only measures the source. See (Piening and Beinert, 2025) for further lower bounds.

Table 10 reports the cost per training batch. Inner GW is 48 to 78 times as expensive as MBOT, and outer GW multiplies this by about 5.4. BoA$^{\mathcal{A}}$+BoA$^{B}_{\text{out}}$ evaluates $B^2\mathcal{A}$ distances and is the slowest non-GW method on molecules, but still 14 to 133 times cheaper than GW+GW$^{8}_{\text{out}}$.

Table 10: Alignment cost per training batch in ms (CPU, single thread). Outer FLB and GW reorder in sub-batches of eight. MBOT and BoA$_{\text{out}}$ assign over the full minibatch.

	Continuous SBM $N = 10$, $B = 16$, $\mathcal{A} = 32$	QM9 $N = 9$, $B = 64$, $\mathcal{A} = 32$	ZINC250k $N = 23$, $B = 32$, $\mathcal{A} = 32$
MBOT^{B}	0.30	1.74	1.71
FLB	0.75	3.39	1.94
FLB+FLB$^{8}_{\text{out}}$	1.48	6.54	3.68
BoA$^{\mathcal{A}}$	0.51	2.62	5.07
BoA$^{\mathcal{A}}$+BoA$^{B}_{\text{out}}$	0.82	9.82	31.14
GW	20.21	136.27	82.76
GW+GW$^{8}_{\text{out}}$	109.50	734.24	442.41