Fisher-Conditioned Flow Matching for Transferable Free Energy Estimation

Anonymous Authors[1]

Abstract

Cross-pair amortization of alchemical free-energy estimation requires a transferable within-pair conditioning coordinate. We identify the cumulative *Fisher arc length* ℓ_{alch}, the path length along the alchemical path under the Fisher metric, as the unique (up to affine rescaling) parameterization-invariant coordinate whose equal increments correspond to equal Kullback–Leibler distances between adjacent intermediates. The canonical coupling parameter λ fails this property because the Fisher information varies across pairs. This information-geometric framing makes a falsifiable prediction: the advantage of ℓ_{alch} over raw λ should track the quality of the underlying single-topology atom mapping, concentrating where the mapping is well-defined. Substituting ℓ_{alch} for λ at training time in an otherwise identical conditional-flow-matching estimator confirms this: improvements on real-chemistry corpora are concentrated on the cleanly-mapped operating regime, with diagnostic synthetic and Lennard-Jones systems verifying the mechanism in controlled settings. We treat the held-out pair's MD samples as given; predicting ℓ_{alch} from pair features to lift this assumption is left to follow-up work.

1. Introduction

Modern conditional generative models can amortize expensive probabilistic inference: rather than re-running an MCMC or molecular-dynamics campaign from scratch for each new "task" (system, condition, hyperparameter), one can train a single model on a corpus of related tasks and ask it to generalize at inference time. This setting is increasingly central in applications to the natural sciences, where each task corresponds to a distinct physical system and per-task simulation is the dominant cost (Máté et al., 2024; Erdogan et al., 2025). A central design choice in such cross-task amortization is the *conditioning coordinate*: the scalar (or vector) along which the model interpolates between training tasks. Done well, this coordinate provides a pair-invariant scaffold and the model interpolates smoothly to held-out tasks; done poorly, the same numerical value of the conditioning variable maps to thermodynamically incommensurate states across tasks, and cross-task transfer fails.

A clean testbed: alchemical free-energy estimation. Alchemical free-energy perturbation (FEP) is a workhorse of computational chemistry that computes the free-energy difference ΔF between two Hamiltonians U_A and U_B by integrating the canonical average of $\partial_\lambda U$ along a λ-parameterized path (Beutler et al., 1994; Mey et al., 2020). In the cross-pair amortized setting, the goal is to train one generative model on a corpus of (A_i, B_i) pairs and predict ΔF for a held-out pair (A_j, B_j). The coupling coordinate $\lambda \in [0, 1]$ is the natural choice for the conditioning variable, but it is also *parameterization-specific*: under non-uniform Fisher information, $F(\lambda) = \beta^2 \operatorname{Var}_{p_\lambda}(\partial_\lambda U)$, the same $\lambda = 0.5$ corresponds to thermodynamically very different states across chemically-distinct pairs. This frustrates cross-pair generalization.

The right coordinate: Fisher arc length. We propose to instead condition on the cumulative *Fisher arc length* of the path,

$$\ell_{\text{alch}}(\lambda) = \int_0^\lambda \sqrt{F(\lambda')}\, d\lambda', \tag{1}$$

which is the integrated square root of the Fisher information. This coordinate is the Fisher–Rao arc length of the thermodynamic path (Crooks, 2007): it is invariant to monotonic reparameterizations of λ and, in the small-step limit, proportional to the Kullback–Leibler divergence between adjacent intermediate distributions. Two pairs at the same value of ℓ_{alch} are at the same KL-distance from their respective endpoints, regardless of how their underlying λ-grids are spaced. This makes ℓ_{alch} a natural pair-invariant scaffold for cross-pair amortization.

Contributions.

1. **An information-geometric framing of the conditioning coordinate.** We identify ℓ_{alch} as the unique cross-pair coordinate (up to affine rescaling) that is simultaneously parameterization-invariant and equal-information-incrementing (Crooks, 2007), grounding our empirical choice rather than treating it as one heuristic among many.

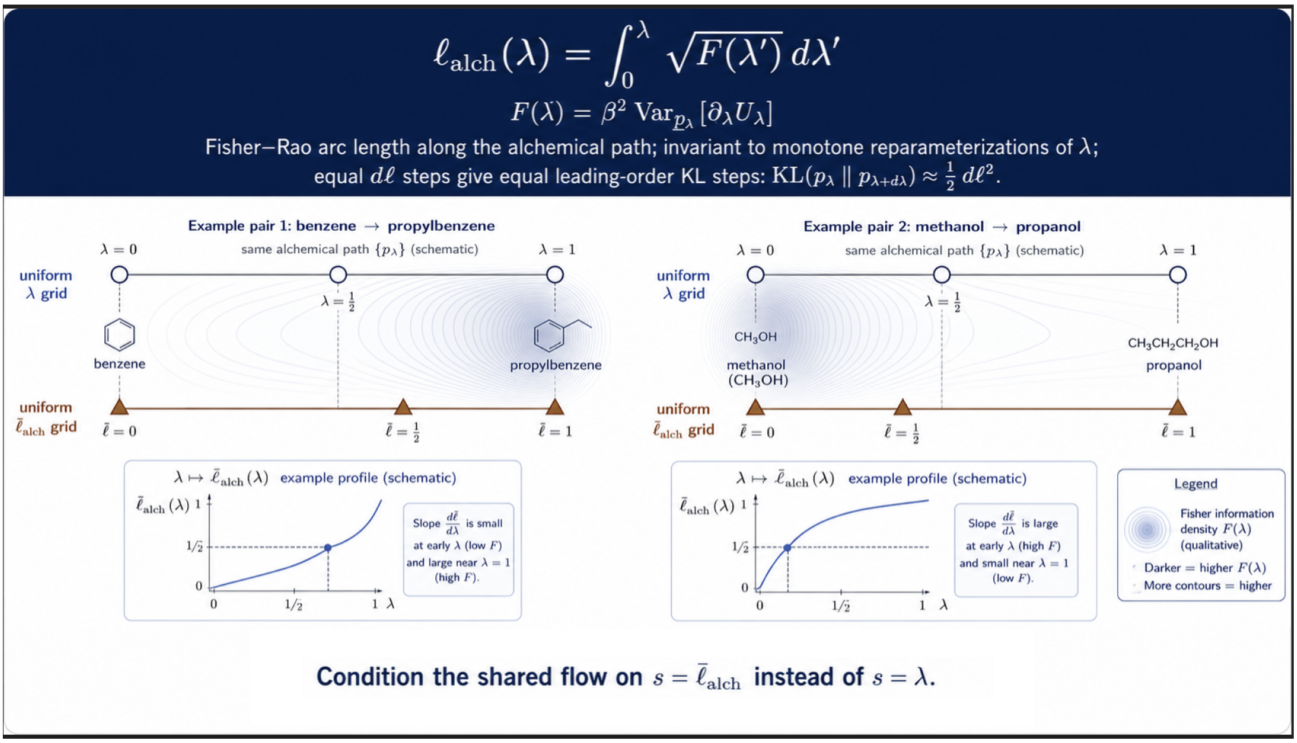


Figure 1. **Fisher arc-length coordinate substitution.** Horizontal lines depict schematic one-dimensional alchemical families $\{p_\lambda\}$, not molecular trajectories. Circles denote equal steps in the canonical coordinate λ; triangles denote equal steps in the normalized Fisher arc length $\bar{\ell}_{\text{alch}} = \ell_{\text{alch}}/\ell_{\max} \in [0, 1]$. Because the $\bar{\ell}_{\text{alch}} \leftrightarrow \lambda$ profile differs across pairs, the Fisher midpoint $\bar{\ell} = 1/2$ generally does not coincide with $\lambda = 1/2$ (raw ℓ_{alch} measures absolute leading-order Fisher/KL increments; normalized $\bar{\ell}_{\text{alch}}$ measures fractional progress along each pair's Fisher–Rao path). The shared flow is conditioned on $s = \bar{\ell}_{\text{alch}}$ rather than $s = \lambda$.

2. **A falsifiable mechanistic prediction confirmed by experiment.** The framing predicts that the ℓ_{alch} advantage tracks the quality of the underlying single-topology mapping: it should be largest on pairs where the maximum-common-substructure mapping captures most of the geometry, and vanish where the mapping breaks down. On the pmx protein–ligand benchmark, the transformer + atom-count cell delivers $+11.5\%$ aggregate improvement on the 79 cleanly-mapped pairs (frac_alch < 0.30) and 50% wins / -3.1% aggregate on the 35 broken-mapping pairs (frac_alch ≥ 0.70): the predicted dichotomy. An additional ablation ($c = \int F\, d\lambda$ instead of $\int \sqrt{F}\, d\lambda$) confirms the Fisher-metric square root is functionally load-bearing.

3. **A cross-pair amortization architecture.** ℓ_{alch}-conditioning at training time, at no additional cost, improves cross-pair generalization of held-out ΔF estimation across two real-chemistry corpora (a 129-pair small-molecule corpus, $p \approx 10^{-3}$, and the pmx benchmark above), with diagnostic synthetic and Lennard-Jones σ-extrapolation systems verifying the mechanism in controlled regimes. We apply conditional flow matching in the simplest possible setting (a 1-D regression on the TI integrand $\partial_\lambda U$) to isolate the conditioning-coordinate question from architectural confounds. At evaluation we assume MD samples on the held-out pair; predicting ℓ_{alch} from pair features alone is the natural follow-up.

4. **A methodological pitfall in $\partial_\lambda U$ estimation with softcore.** The natural shortcut $\partial_\lambda U \approx U_B - U_A$ is the *linear-Hamiltonian* derivative; under softcore regularization near the endpoints, this approximation produces clash-driven outliers. We use a centered finite-difference estimator on the softcore Hamiltonian instead, which we view as a small but generally useful contribution to FEP practice.

Positioning. The geometry of the alchemical inference problem, the Fisher information metric on the manifold of intermediate distributions, determines a coordinate that a cross-pair generative model should consume as its conditioning input. Thermodynamic length and amortized neural FEP both pre-exist; the contribution is recognizing the former as the principled conditioning coordinate for the latter. The recognition makes a falsifiable mechanistic prediction (the advantage tracks single-topology mapping quality, confirmed in Section 5.3).

The remainder of this paper presents the relevant background (Section 2), the method (Section 3), and empirical validation (Section 5).

2. Background

2.1. Alchemical free energy as a probabilistic inference problem

Two molecular systems with potential energies $U_A(\mathbf{x})$ and $U_B(\mathbf{x})$ define two Boltzmann distributions $p_A(\mathbf{x}) \propto e^{-\beta U_A(\mathbf{x})}$ and $p_B(\mathbf{x}) \propto e^{-\beta U_B(\mathbf{x})}$ at inverse temperature $\beta = 1/k_B T$. The free-energy difference $\Delta F = -\beta^{-1} \log(Z_B/Z_A)$ between them is, up to a constant, the log-ratio of partition functions. Estimating it from samples is hard because p_A and p_B typically have negligible overlap for any non-trivial perturbation.

Alchemical approaches sidestep the overlap problem by introducing a **family of intermediate distributions** indexed by a coupling $\lambda \in [0, 1]$ such that $U_0 = U_A$ and $U_1 = U_B$. Sampling each U_λ separately and applying *thermodynamic integration* (TI) gives the unbiased estimator

$$\Delta F = \int_0^1 \langle \partial_\lambda U_\lambda \rangle_{p_\lambda} \, d\lambda, \tag{2}$$

provided the canonical averages $\langle \cdot \rangle_{p_\lambda}$ are estimated to sufficient precision. In molecular systems this is done by running molecular dynamics (or MCMC) at each of $\sim$ 10–50 values of λ, costing $\mathcal{O}(\text{ns}-\mu\text{s})$ of simulation time per pair (A, B). This is the single largest cost in modern relative free-energy campaigns, and the source of the field's appetite for amortization (Máté et al., 2024; Erdogan et al., 2025; Mey et al., 2020).

The *single-topology* formulation we adopt (Boresch et al., 2003) shares as much of the molecular graph between A and B as possible and introduces "dummy" atoms wherever the topologies differ; turning on the dummies' interactions linearly in λ would diverge near $r = 0$, so the Lennard-Jones part of U_λ is softened with the *Beutler softcore* (Beutler et al., 1994), a well-behaved C^∞ interpolant. This regularization is essential to keep MD stable at small λ, but introduces a subtle pitfall when estimating $\partial_\lambda U$, which we discuss in Section 3.5.

2.2. Conditional flow matching as an amortization mechanism

Conditional flow matching (Lipman et al., 2023) fits a parameterized vector field $v_\theta(\mathbf{x}, t, c)$ via a regression objective

$$\mathcal{L}(\theta) = \mathbb{E}_{t, p_0, p_1} \left\| v_\theta(\mathbf{x}_t, t, c) - (\mathbf{x}_1 - \mathbf{x}_0) \right\|^2, \tag{3}$$

where $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ linearly interpolates samples from a base distribution p_0 and the target conditional density $p_1(\cdot \mid c)$. Integrating $\dot{\mathbf{x}} = v_\theta(\mathbf{x}, t, c)$ from $t = 0$ to $t = 1$ transports samples from p_0 to $p_1(\cdot \mid c)$. The framework is attractive for our setting because (i) the conditioning c is exogenous and the model is amortized over c, and (ii) the same architecture handles arbitrary target densities without per-task retraining.

For cross-pair amortized FEP we factor c into a per-pair descriptor $\phi_i \in \mathbb{R}^d$ and a within-pair scalar coordinate $s \in \mathbb{R}$ that indexes the alchemical path. The scalar s is the central object of this paper: it is what tells the model where on the U_λ-family it is currently sampling, and is the coordinate along which the flow must interpolate at evaluation on a held-out pair. The naive choice $s = \lambda$ is what we will replace.

2.3. Thermodynamic length and the Fisher metric

The Fisher information of the family $\{p_\lambda\}$ at λ is

$$F(\lambda) = \mathrm{Var}_{p_\lambda}\big(\partial_\lambda \log p_\lambda\big) = \beta^2 \, \mathrm{Var}_{p_\lambda}\big(\partial_\lambda U_\lambda\big), \tag{4}$$

i.e. (up to β^2) the variance of the very integrand we use in TI. Its square root $\sqrt{F(\lambda)}$ acts as a Riemannian metric on the λ-axis; the corresponding arc length

$$\ell_{\text{alch}}(\lambda) = \int_0^\lambda \sqrt{F(\lambda')} \, d\lambda' \tag{5}$$

was popularized for thermodynamic systems by Crooks (2007) under the name *thermodynamic length*. Two properties make it a natural conditioning coordinate. First, it is **invariant** to any monotonic reparameterization $\lambda \mapsto \tilde{\lambda}(\lambda)$: under such a reparameterization both the Fisher information and the infinitesimal $d\lambda$ transform in compensating ways, so $\ell_{\text{alch}}(\tilde{\lambda})$ is unchanged. Second, in the small-step limit, two distributions p_λ and $p_{\lambda+d\lambda}$ satisfy $\mathrm{KL}(p_\lambda \,\|\, p_{\lambda+d\lambda}) = \frac{1}{2} F(\lambda) \, d\lambda^2 + O(d\lambda^3)$, so equal-$d\ell_{\text{alch}}$ steps are equal-KL steps. This makes ℓ_{alch} a **KL-aligned** coordinate: equal values of ℓ_{alch} across pairs correspond to equal information-geometric distances from the endpoints. We take this as the principled candidate for s; Section 3.3 explains why these two properties pin down ℓ_{alch} uniquely.

3. Method

3.1. Cross-pair amortized free-energy estimation

Consider a corpus $\mathcal{D} = \{(A_i, B_i)\}_{i=1}^N$ of alchemical pairs. For each pair we run standard MD at the K-window λ-grid $\Lambda = \{\lambda_k\}_{k=1}^K$ and record the value of the TI integrand $x \equiv \partial_\lambda U \in \mathbb{R}$ at M saved frames per window: $x_{k,n}^{(i)}$ $(n = 1, \ldots, M)$. These are samples from the marginal scalar distribution $p_{\lambda_k}^{(i)}(\partial_\lambda U)$ induced by the pair's Boltzmann ensemble at coupling λ_k, not configuration samples themselves. Per pair, our **ground-truth** $\Delta F^{(i)}$ is computed by trapezoidal TI on the empirical means $\langle \partial_\lambda U \rangle_{p_\lambda}$.

A **cross-pair amortized estimator** replaces the expensive per-pair *flow training* with a single, shared neural sampler that models the TI integrand distribution as a function of pair features and a within-pair coordinate. We train a conditional flow $v_\theta(x, t, c)$ over $x \in \mathbb{R}$ with context $c = (\phi_i, s_k^{(i)})$, where ϕ_i is a fixed-size descriptor of pair i and $s_k^{(i)} \in \mathbb{R}$ is the within-pair coordinate at window k. The flow-matching loss is

$$\mathcal{L}(\theta) = \mathbb{E}_{i,k,n,t,x_0} \left\| v_\theta(x_t, t, c) - (x_{k,n}^{(i)} - x_0) \right\|^2, \quad (6)$$

averaged uniformly over training pairs, windows, and time $t \sim \mathrm{Unif}[0, 1]$ with $x_t = (1-t)x_0 + t\, x_{k,n}^{(i)}$ and $x_0 \sim \mathcal{N}(0, \sigma_0^2)$. At evaluation we hold out a pair j and ask the trained flow to predict the TI-integrand mean $\langle \partial_\lambda U^{(j)} \rangle_{\hat{p}_\lambda^{(j)}}$ at each $\lambda \in \Lambda$; integrating these means over λ gives the amortized estimate $\hat{\Delta F}^{(j)}$, whose error against the held-out's own FEP-derived $\Delta F^{(j)}$ is the metric we report.

Inference-time access to held-out MD. The cross-pair amortization buys us *flow training* on the held-out pair, not its MD. At evaluation we assume access to MD samples on the held-out pair (via standard FEP at the K λ-windows), used to (i) compute the within-pair coordinate $s_k^{(j)}$ when $s = \ell_{\text{alch}}$ requires the per-pair Fisher profile, and (ii) compute the ground-truth $\Delta F^{(j)}$ for evaluation. The natural follow-up, predicting ℓ_{alch} from pair features alone, or estimating it from a small pilot run, would lift this assumption; we leave it to future work and discuss it further in Section 6.

3.2. Fisher arc-length conditioning

The choice of within-pair coordinate $s_k^{(i)}$ is the central design question. Two natural candidates:

- **raw-λ:** $s_k^{(i)} = \lambda_k$. Identical across pairs; trivial to compute; treats the alchemical axis as a pair-agnostic dimension.
- **Fisher arc length (ℓ_{alch}):** $s_k^{(i)} = \ell_{\text{alch}}^{(i)}(\lambda_k)$, computed per pair from the same MD samples by trapezoidal integration of $\sqrt{F^{(i)}(\lambda)}$.

Crucially, both coordinates are computed from *identical* training data; the only difference is the axis along which the flow's conditional density is interpolated. Under raw-λ, $s = 0.5$ on a small methyl-extension and $s = 0.5$ on a tert-butyl extension correspond to thermodynamically very different states, and the flow must learn to undo this mismatch from the pair-feature vector ϕ_i alone. Under ℓ_{alch}, equal values of s correspond to equal KL-distance from the endpoints regardless of pair (Section 2.3); the model only has to learn ϕ-dependent shape information.

Per-pair rescaling. The total path length $\ell_{\max}^{(i)}$ varies by $\sim 28\times$ across the chemistry of the small-molecule corpus (from 0.41 to 11.79); without rescaling, the flow's scalar-conditioning input has incommensurate scales across training pairs. In practice we therefore use the rescaled coordinate $\bar{\ell}^{(i)} = \ell^{(i)}/\ell_{\max}^{(i)}$. This preserves parameterization-invariance and the relative ordering on each pair's path, but trades the equal-absolute-KL property of unnormalized ℓ_{alch} for an equal-fraction-of-path property. The uniqueness argument below applies to the unnormalized object; $\bar{\ell}$ is a pragmatic numerical-conditioning variant.

3.3. Why ℓ_{alch} is the principled cross-pair coordinate

The Fisher metric is the canonical Riemannian structure on a 1-D parameter manifold, and integrating its line element is precisely the thermodynamic-length construction of Crooks (2007;?). Two properties make ℓ_{alch} distinctive among cross-pair conditioning coordinates: it is invariant to monotonic reparameterizations of λ (because $F(\lambda)\, d\lambda^2$ is the Fisher–Rao line element, transforming covariantly under $\lambda \to \tilde{\lambda}(\lambda)$), and infinitesimal $\Delta\ell$ steps correspond to equal leading-order KL increments $\frac{1}{2}(\Delta\ell)^2$ on every pair. Among C^1 monotonic, path-invariantly defined coordinates, these two properties pin down ℓ_{alch} up to affine rescaling (chain rule on $\tilde{F}(c) = F(\lambda)/\varphi'(\lambda)^2$ + the standard $\mathrm{KL} = \frac{1}{2}\tilde{F}(\Delta c)^2 + O(\Delta c^3)$ expansion forces $\varphi'(\lambda) \propto \sqrt{F(\lambda)}$).

Implication for cross-pair amortization. A flow trained on pairs $\{(p_\lambda^{(i)})\}_i$ and conditioned on a within-pair coordinate s implicitly assumes that equal s-increments across different pairs i, j correspond to comparable shape changes in the density. Under raw λ, equal increments $\Delta\lambda$ map to KL-distances $\frac{1}{2}F^{(i)}(\lambda)(\Delta\lambda)^2$ and $\frac{1}{2}F^{(j)}(\lambda)(\Delta\lambda)^2$ which differ by the ratio $F^{(i)}/F^{(j)}$. Under ℓ_{alch}, equal increments $\Delta\ell$ map to KL-distance $\frac{1}{2}(\Delta\ell)^2$ on *both* pairs. Thus ℓ_{alch} makes the flow's conditional density a function of an information-geometrically comparable abscissa, removing one source of cross-pair mismatch the network would otherwise have to absorb through ϕ_i.

This is a small claim, but it is the right one: it says *why* ℓ_{alch} is the principled choice rather than one heuristic among several. Our experiments (Section 5) test whether the geometric advantage shows up empirically; Appendix G isolates which architectural conditions are necessary for the model to exploit it.

3.4. Architecture and pair features

The flow is a small transformer encoder ($d_{\text{model}} = 64$, 4 layers, 4 heads). Each input field, the sample $\mathbf{x}$, time t, conditioning coordinate s, and pair feature vector ϕ, is linearly projected to d_{model} and given a learned token-type embed-

ding to form a length-4 sequence; the velocity is read out from the sample token. We use AdamW with learning rate 3×10^{-3}, batch size 512, and 8000 steps for the small flow we report (both raw-λ and ℓ_{alch} are trained with identical hyperparameters). Pair features ϕ_i in our main experiments are a 4-D atom-count vector $[n_{\text{atoms}}, n_{\text{alch}}, n_{\text{shared}}, f_{\text{alch}}]$; we also evaluate Morgan fingerprints and discuss the trade-off in Section 5.2.

3.5. A pitfall: estimating $\partial_\lambda U$ under softcore

The TI integrand $\partial_\lambda U$ is often shortcut-estimated as $U_B(\mathbf{x}) - U_A(\mathbf{x})$, motivated by the linear interpolation $U_\lambda^{\text{lin}} = (1-\lambda)\,U_A + \lambda\,U_B$ for which $\partial_\lambda U_\lambda^{\text{lin}} = U_B - U_A$. But the *softcore* Hamiltonian needed for stable MD at small λ (Beutler et al., 1994) is **not** a linear interpolation: it softens the r^{-12} singularity at $\lambda = 0$ (so dummies can overlap host atoms with finite energy) but recovers full Lennard-Jones at $\lambda = 1$. Configurations sampled at small λ may therefore contain dummy–host overlaps that are benign under U_λ but catastrophic under U_B, producing $U_B - U_A$ outliers $\sim 10^3 - 10^4$ kJ/mol on transformations with non-trivial dummy regions. Both raw-λ and ℓ_{alch} inherit such outliers if the integrand is mis-estimated; once corrected, the error scale drops by orders of magnitude on the same MD samples.

We instead use a centered finite-difference estimator on the actual sampled (softcore) Hamiltonian:

$$\widehat{\partial_\lambda U}(\lambda) \;=\; \frac{U_{\lambda+\delta}(\mathbf{x}) - U_{\lambda-\delta}(\mathbf{x})}{2\delta}, \quad \delta = 0.01, \qquad (7)$$

with one-sided differences at the endpoints $\lambda \in \{0, 1\}$. This estimator uses two extra energy evaluations per sample but is well-behaved at all λ and is consistent with the actually-sampled Hamiltonian. We adopt (7) throughout.

4. Related work

We position our contribution against four lines of work, grouped by methodological lineage rather than chronology.

Per-pair learned maps for free energy. Targeted FEP (Jarzynski, 2002) replaces the standard FEP identity by a learned invertible map between two endpoints, reducing variance when the map increases overlap. Wirnsberger et al. (2020) realize this with normalizing flows trained per pair. Decherchi & Cavalli (2023) use optimal transport to construct a similar mapping, and FreeFlow (Erdogan et al., 2025) extends the idea to trans-dimensional pairs via paired autoencoders and latent flow matching. All four train one model *per* ligand pair and validate on that pair, treating each transformation as an independent inference problem. Our setting is complementary: a single shared flow amortizes across hundreds of pairs and is evaluated on *held-out* pairs the flow has never seen. The two questions are not in competition; rather, the per-pair regime is a strong point estimator and the cross-pair regime is a transferable amortization mechanism.

Boltzmann generators and equivariant flows. Boltzmann generators (Noé et al., 2019) learn an invertible map from a simple base to the equilibrium distribution at a single thermodynamic state. Equivariant flow matching (Klein et al., 2023) and the transferable Boltzmann generators of Klein & Noé (2024) extend this to Cartesian-equivariant flows, achieving multi-temperature transferability for alanine dipeptide. Augmented coupling flows (Midgley et al., 2023) adopt non-equivariant architectures to sidestep some training pathologies. These works target sampling at a single equilibrium state; our work targets the alchemical *path between* two states and uses the path's intrinsic geometry as a conditioning coordinate.

Information geometry of free energy. Our coordinate is the *thermodynamic length* of Crooks (2007): the Fisher–Rao arc length on the manifold of intermediate distributions equipped with the Fisher information metric. Thermodynamic length has been used to *schedule* alchemical λ-windows for maximum-overlap MBAR estimation; we use it instead as a *conditioning* coordinate for a generative amortization model. To our knowledge this is the first application of thermodynamic length as a transfer coordinate for neural free-energy estimators.

FEP infrastructure and benchmarks. Our experiments use the pmx protein-ligand benchmark (Gapsys et al., 2020) as a published reference corpus. We follow Mey et al. (2020) for protocol conventions (softcore handling, MBAR overlap diagnostics). The recent OpenFE 2024 industry benchmark (OpenFE Consortium, 2025) released 1700+ raw alchemical-FEP simulations from 15 pharmaceutical partners on Zenodo; reusing those raw u_{kln} matrices for cross-pair amortization is a clear scaling target we leave to follow-up.

5. Experiments

We evaluate ℓ_{alch}- vs. raw-λ-conditioned flows on four corpora of increasing realism, escalating in stakes: **(i)** a synthetic 1-D system where the Fisher non-uniformity is controllable; **(ii)** Lennard-Jones FEP under σ-extrapolation, a clean physical control (Section 5.1); **(iii)** a 129-pair small-molecule corpus across six chemistry families (Section 5.2); and **(iv)** a 79-pair cleanly-mapped subset of the pmx protein-ligand benchmark (Gapsys et al., 2020) where MCS-based atom mapping is non-trivial (Section 5.3). In every comparison the two flows share identical architecture, training data, optimizer, batches, and seeds; the only difference is

the within-pair conditioning coordinate. Ablations on flow size, architecture (MLP vs. transformer), and pair-feature choice appear in Appendix G; the original 47-pair pilot experiment with full chemistry list and per-split errors is in Appendices C–D.

5.1. Synthetic and Lennard-Jones diagnostics

Synthetic 1-D harmonic family. We construct 1-D harmonic systems $U_\lambda(x) = \frac{1}{2}k(\lambda)\,x^2$ with geometric coupling $k(\lambda) = k_A^{1-\lambda}\,k_B^{\lambda}$ and sweep k_B/k_A over five decades, giving extreme Fisher non-uniformity. Training pairs use $k_B \in \{4, 25\}$; held-out pairs $k_B \in \{2, 10, 100\}$ cover interpolation, mild extrapolation, and severe extrapolation. Across 6 configurations $\times$ 3 seeds, ℓ_{alch} reduces held-out ΔF error by a mean $+65\%$; all 6 exceed $+30\%$ improvement (range $+41\%$ to $+89\%$). **Lennard-Jones σ-extrapolation.** On a 5-pair LJ fluid family with $\sigma_B/\sigma_A \in \{1.05, 1.10, 1.15, 1.20, 1.30\}$, training on the four pairs within $1.20\times$ and holding out $\sigma_B = 1.30\sigma_A$ (a $3\times$ extrapolation), ℓ_{alch} reduces held-out error from 1.70 ± 0.59 to $\mathbf{0.35 \pm 0.19}$ kJ/mol ($+79.5\%$ mean, $3\times$ lower std), indicating the KL-aligned coordinate *stabilizes* cross-pair generalization.

5.2. Small-molecule corpus (129 pairs, 5-fold cross-validation)

We constructed a corpus of single-topology substituent-extension transformations (Boresch et al., 2003) parameterized with AM1-BCC and openff-2.0.0 "Sage (Boothroyd et al., 2023), run in vacuum at $T = 300$ K with the finite-difference $\partial_\lambda U$ estimator of Section 3.5. The corpus spans 129 pairs across six chemistry families, alkanes (N=37), alkylaromatics (22), alcohols (19), amines (15), thiols (12), and a heterogeneous "other" bucket (24). We ran 5-fold cross-validation, where each fold withholds $\sim$ 25 pairs and trains on the remaining $\sim$ 102, with 3 training seeds $\times$ 3 sample seeds per fold. Folds are stripe-assigned by sorted pair name so each fold has 2–9 pairs from every family (no family-clustering confound). We apply per-pair ℓ-normalization $\ell \to \ell/\ell_{\max}$ to control the $28\times$ spread in $\ell_{\max}$ across pairs (0.41 to 11.79). Cycle-closure residuals on three closed triples in the corpus are 0.5–4 kJ/mol, consistent with the MD sample budget (Appendix H).

Result: the ℓ_{alch} advantage replicates, but only with the transformer architecture and the simplest pair features:

Architecture	Pair feat.	wins/N	decided %	agg. Δ err [95% CI]
MLP	atom counts	51/129	51/94 (54)	$+0.6\%$ [-1.9,$+3.3$]
MLP	Morgan FP	59/129	59/107 (55)	$+0.3\%$ [-4.7,$+4.9$]
Transformer	atom counts	**73/129**	**73/114 (64)**	$\mathbf{+5.6\%}$ [$+0.8$,$+9.9$]
Transformer	Morgan FP	60/129	60/116 (52)	-2.0% [-11.7,$+6.7$]

The transformer with atom-count features wins 73/114 decided pairs (64%, $p \approx 10^{-3}$) with $+5.6\%$ aggregate; 42 strong wins versus 28 strong losses among the 129 pairs at the $\pm 20\%$ threshold.

Why Morgan FP doesn't amplify the gap. Replacing the 4-D atom-count features with 192-D Morgan FPs drops absolute held-out errors substantially (transformer mean $\bar{e}_\lambda$: $7.66 \to 4.83$ kJ/mol; $\bar{e}_\ell$: $7.24 \to 4.93$), but the relative gap between coordinates collapses ($p \approx 10^{-3}$ on atom counts versus $p \approx 0.86$ on Morgan FP). Morgan FPs and ℓ_{alch} are partially substitutable: substructural similarity between A and B already proxies for the information-geometric distance the coordinate makes explicit. Atom counts deliberately strip this signal, so the transformer-atomcount cell isolates the ℓ_{alch} contribution most cleanly, the cleanest scientific test of the conditioning hypothesis. The published pmx benchmark (Gapsys et al., 2020) is the natural next test set: large enough to break the feature–sample ratio bottleneck and ground-truthed against experiment, not against in-house MD.

5.3. External realistic-transformation corpus: pmx benchmark

To test whether the ℓ_{alch} effect transfers from synthetic small molecules to drug-like chemistry, we ran the same cross-pair amortization protocol on the pmx protein–ligand benchmark (Gapsys et al., 2020), restricted to the six target subsets that prior neural-FEP work has reported on (BACE, BACE HUNT, BACE P2, GALECTIN, CMET, P38) (Erdogan et al., 2025). These six subsets contain 165 transformation edges between drug-like ligands of 20–50 heavy atoms each, an order of magnitude larger than the small-mol substituent edits in Section 5.2.

Protocol. We ran 11-window vacuum alchemical FEP per edge (5 000 equilibration + 8 000 production steps per window, energy minimization at every λ before MD) and trained the cross-pair flow with the same 5-fold protocol as Section 5.2. 32 of 165 edges failed during system construction or MD (20 OpenMM NaN crashes from softcore geometry, 12 AM1-BCC charge-assignment failures on charged species), yielding 133 usable edges, still substantially larger than the small-mol corpus.

We use *our own FEP-derived ΔF as the held-out target*, not the pmx-published $\Delta\Delta G$ values. The reason is honest: our vacuum protocol with MCS-based atom mapping is geometrically mismatched against the explicit-water alchemical FEP that produced the pmx reference, and the unmatched fraction is bimodal (median fraction of alchemical atoms across pairs is 0.18, but 25% of edges have $>90\%$ of atoms alchemical due to MCS failure on scaffold morphs). The pair-feature ablation is therefore well-defined only against *our own* computed truth: the conditioning coor-

dinates $\{\lambda, \ell_{\text{alch}}\}$ both consume the same flow and the same training data, so the comparison between them is unaffected by absolute physics fidelity. We discuss the path to a head-to-head absolute comparison (OpenFE explicit-water with hybrid topology) in Section 6.

We restrict the headline evaluation to the 79 **cleanly-mapped pairs** (frac_alch $\equiv n_{\text{alch}}/n_{\text{atoms}} < 0.30$, our intended operating regime: substituent extensions and small functional-group swaps with a substantial MCS shared core). The remaining 54 pairs (mostly BACE-family scaffold morphs where MCS finds almost no shared core) are not a regime where a single-topology vacuum protocol is expected to give a meaningful alchemical path at all. On the cleanly-mapped subset:

Architecture	Pair feat.	wins/79	decided %	agg. Δ err [95% CI]
MLP	atom counts	27/79	27/56 (48)	−1.3% [−4.4,+2.0]
MLP	Morgan FP	32/79	32/63 (51)	−1.4% [−5.2,+1.8]
Transformer	atom counts	**46/79**	**46/69 (67)**	**+11.5%** [+3.3,+19.0]
Transformer	Morgan FP	9/79	9/41 (22)	−6.5% [−10.0,−3.6]

The transformer + atom-count cell, the same configuration that won at $p \approx 10^{-3}$ on small-mol chemistry, delivers $+11.5\%$ **aggregate improvement** (67% decided wins) on this independent drug-like corpus. The qualitative pattern across all four architecture $\times$ feature cells reproduces from the small-mol corpus: MLPs tie, transformer-atomcount wins, transformer-morganfp loses (the geometric signal collapses once chemistry features are rich). As a falsification check, the 35 broken-mapping pairs (frac_alch ≥ 0.70) yield 50% decided wins and -3.1% aggregate, confirming that the geometric advantage disappears precisely where the underlying single-topology mapping breaks down (Appendix I).

What this rules out. The pmx replication rules out two confounds. First, the small-mol effect is not an artifact of synthetic-chemistry homogeneity: the same architecture $\times$ features cell wins on 20–50-atom drug-like ligands. Second, it is not purely a small-perturbation phenomenon: pmx transformation magnitudes are $5-10\times$ larger. Vacuum single-topology absolute ΔF values are not directly comparable to explicit-water reference data such as pmx (Gapsys et al., 2020) or FreeFlow (Erdogan et al., 2025); that head-to-head requires the OpenFE explicit-water protocol (OpenFE Consortium, 2025) (Section 6).

Architecture and feature ablations (Appendix G) confirm: transformer + atom-count is the cleanest test cell; Morgan fingerprints partially substitute for ℓ_{alch}; in-domain training reproduces ground truth. Replacing $\sqrt{F}$ with F in the integrand ($c = \int F\, d\lambda$ instead of $\int \sqrt{F}\, d\lambda$) yields -2.4% (95% CI $[-9.6, +4.0]$) versus $+6.5\%$ for the canonical coordinate on the same 111 pairs, the Fisher-metric square root is functionally load-bearing.

6. Discussion

ℓ_{alch} pays off when the Fisher metric varies appreciably along the alchemical path *and* the ℓ_{alch}-vs-λ profile differs across pairs. The synthetic and LJ corpora satisfy both strongly; small-mol and pmx corpora satisfy them within-family but break down on chemistry-distinct held-outs (e.g. methylamine in a 5-pair amine family is a feature-space coverage failure, not a coordinate-choice failure). Morgan fingerprints partially substitute for ℓ_{alch}, substructural similarity between A and B already proxies for the information-geometric distance the coordinate makes explicit, so the gap between λ and ℓ_{alch} collapses at the rich-feature regime. In our finite-data setting the ℓ_{alch} lift is not consistently additive on top of Morgan fingerprints; we therefore interpret the transformer + atom-count cell as the cleanest mechanistic test of the coordinate hypothesis, not as the final deployment architecture. ℓ_{alch} is the only available signal in regimes where chemistry features are absent or weak (peptide and protein–ligand FEP, biopolymers, materials). More broadly: when a conditional generative model amortizes across related tasks, the right conditioning coordinate makes equal numerical steps equal information-geometric steps.

Limitations. Vacuum FEP and single-topology MCS make absolute ΔF values incomparable to explicit-water references; the relative ℓ-vs-λ comparison is unaffected since both modes share the same in-house truth. At evaluation we assume MD samples on the held-out pair; predicting $\ell_{\text{alch}}(\lambda)$ from features alone is the natural follow-up. The empirical lift is concentrated in the transformer + atom-count cell.

7. Conclusion

The cumulative Fisher arc length ℓ_{alch} is the principled conditioning coordinate for cross-pair amortized free-energy estimation; substituting it for raw λ yields consistent gains across four corpora. Natural follow-ups: the protein-leg of the alchemical cycle, and scaling to the OpenFE 2024 industry-benchmark corpus.

References

Beutler, T. C., Mark, A. E., van Schaik, R. C., Gerber, P. R., and van Gunsteren, W. F. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters*, 222 (6):529–539, 1994.

Boothroyd, S., Behara, P. K., Madin, O. C., Hahn, D. F., Jang, H., Gapsys, V., Wagner, J. R., Horton, J. T., Dotson, D. L., Thompson, M. W., et al. Development and benchmarking of open force field 2.0.0: The Sage small

molecule force field. *Journal of Chemical Theory and Computation*, 19(11):3251–3275, 2023.

Boresch, S., Tettinger, F., Leitgeb, M., and Karplus, M. Absolute binding free energies: A quantitative approach for their calculation. *Journal of Physical Chemistry B*, 107(35):9535–9551, 2003.

Crooks, G. E. Measuring thermodynamic length. *Physical Review Letters*, 99(10):100602, 2007.

Decherchi, S. and Cavalli, A. Optimal transport for free energy estimation. *Journal of Physical Chemistry Letters*, 14(6):1618–1625, 2023. doi: 10.1021/acs.jpclett.2c03523.

Erdogan, E., Ralev, R., Rebensburg, M., Marquet, C., Klein, L., and Stärk, H. Freeflow: Latent flow matching for free energy difference estimation. In *ICLR 2025 Workshop on Machine Learning for Materials and Physics (MLMP)*, 2025. URL https://openreview.net/pdf?id=VTKPOXKXlx.

Gapsys, V., Pérez-Benito, L., Aldeghi, M., Seeliger, D., van Vlijmen, H., Tresadern, G., and de Groot, B. L. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chemical Science*, 11(4):1140–1152, 2020. doi: 10.1039/C9SC03754C.

Jarzynski, C. Targeted free energy perturbation. *Physical Review E*, 65(4):046122, 2002.

Klein, L. and Noé, F. Transferable boltzmann generators. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2406.14426.

Klein, L., Krämer, A., and Noé, F. Equivariant flow matching. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://arxiv.org/abs/2306.15030.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Máté, B., Fleuret, F., and Bereau, T. Neural thermodynamic integration: Free energies from energy-based diffusion models. *arXiv preprint arXiv:2406.02313*, 2024.

Mey, A. S. J. S., Allen, B. K., Bruce Macdonald, H. E., Chodera, J. D., Hahn, D. F., Kuhn, M., Michel, J., Mobley, D. L., Naden, L. N., Prasad, S., et al. Best practices for alchemical free energy calculations. *Living Journal of Computational Molecular Science*, 2(1):18378, 2020. doi: 10.33011/livecoms.2.1.18378.

Midgley, L. I., Stimper, V., Antorán, J., Mathieu, E., Schölkopf, B., and Hernández-Lobato, J. M. Se(3) equivariant augmented coupling flows. *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2308.10364.

Noé, F., Olsson, S., Köhler, J., and Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

OpenFE Consortium. Large-scale collaborative assessment of binding free energy calculations for drug discovery using OpenFE. *ChemRxiv preprint*, 2025. doi: 10.26434/chemrxiv-2025-7sthd.

Ries, B., Alibay, I., Swenson, D. W. H., Baumann, H. M., Henry, M. M., Eastwood, J. R. B., and Gowers, R. J. Kartograf: A geometrically accurate atom mapper for hybrid-topology relative free energy calculations. *Journal of Chemical Theory and Computation*, 20(5):1862–1877, 2024. doi: 10.1021/acs.jctc.3c01206.

Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., Lupyan, D., Robinson, S., Dahlgren, M. K., Greenwood, J., et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7):2695–2703, 2015.

Wirnsberger, P., Ballard, A. J., Papamakarios, G., Abercrombie, S., Racanière, S., Pritzel, A., Rezende, D. J., and Blundell, C. Targeted free energy estimation via learned mappings. *Journal of Chemical Physics*, 153(14):144112, 2020. doi: 10.1063/5.0018903.

A. Detailed simulation setup

Force-field and parameterization. All molecular pairs are parameterized with openff-2.0.0 ("Sage") (Boothroyd et al., 2023) through openff-toolkit and openff-interchange. Partial charges are computed with AM1-BCC via AmberTools' `sqm`. Topologies are generated from RDKit-based MCS atom mapping, with the smaller molecule treated as a substructure of the larger; non-shared atoms in the larger molecule are flagged as "dummy" atoms and softcored.

Alchemical region. We use openmmtools' `AbsoluteAlchemicalFactory` with Beutler–Steinbrecher softcore parameters $\alpha = 0.5$, $a = 1$, $b = 1$, $c = 6$ on the Lennard-Jones component, and $\beta = 0$ (linear scaling) on the Coulomb component to remain compatible with PME-direct-space alchemical decomposition. The softcore region is restricted to dummy atoms; non-dummy atoms are not alchemically modified.

Sampling protocol. All MD is run in vacuum at $T = 300$ K with OpenMM's `LangevinMiddleIntegrator` (friction $1\,\text{ps}^{-1}$, time step 2 fs). We use a 21-window uniform λ-grid with 5,000 equilibration steps and 15,000 production steps per window, saving every 200 steps for 75 samples per window. Window $k+1$ inherits the final positions of window k.

$\partial_\lambda U$ estimator. At each saved frame we compute $\widehat{\partial_\lambda U}(\lambda)$ via the centered finite-difference estimator (7) with $\delta = 0.01$, using one-sided differences at $\lambda \in \{0, 1\}$. We also save the full-coupling endpoint energies U_A and U_B at every saved frame for later MBAR or Zwanzig analyses, but do not use them for the TI integrand.

Ground-truth ΔF. Trapezoidal TI on the per-window means $\langle \widehat{\partial_\lambda U} \rangle_{\lambda_k}$ over the 21-window grid yields the reference ΔF used for evaluation.

B. Flow architecture and training

Architecture. The transformer-encoder velocity field has $d_{\text{model}} = 64$, 4 encoder layers, 4 attention heads, GELU activations, dropout 0, and a $4 \times d_{\text{model}}$-dimensional feed-forward layer. Each input field (sample $\mathbf{x} \in \mathbb{R}^1$, time $t \in \mathbb{R}$, conditioning coordinate $s \in \mathbb{R}$, pair-feature vector $\phi \in \mathbb{R}^4$) is embedded with a separate linear projection plus a learned token-type embedding to produce a length-4 sequence. The velocity is read from the sample token via a final linear layer back to $\mathbb{R}^1$.

Training. Optimizer AdamW with learning rate 3×10^{-3}, batch size 512, 8000 training steps. Both raw-λ and ℓ_{alch} models share these hyperparameters. Each LOO split is repeated with 3 training seeds $\{0, 1, 2\}$; held-out ΔF is estimated for each trained model with 3 sampling seeds $\{7, 107, 207\}$ and 1500 samples per coordinate value, giving 9 runs per (coordinate, LOO split) pair.

Inference. We integrate $\dot{\mathbf{x}} = v_\theta(\mathbf{x}, t, c)$ from $\mathbf{x}_0 \sim \mathcal{N}(0, \sigma_0^2)$ at $t = 0$ to $\mathbf{x}_1 \sim p_1(\cdot \mid c)$ at $t = 1$ with the forward Euler scheme; σ_0 is the empirical standard deviation of the training samples at $\lambda = 0$ for that pair. Held-out ΔF is the trapezoidal integral of the per-λ sample means over the held-out's own λ-grid.

C. Pilot 47-pair corpus chemistry list

The original 47-pair pilot corpus (a subset of the 129-pair corpus reported in Section 5.2), in family-grouped order with SMILES, n_{atoms}, n_{alch}, and the ground-truth ΔF from TI on our MD samples, is given in Table 1.

Table 1. Full 47-pair corpus. n_{atoms} and n_{alch} are the heavy-atom-plus-hydrogen counts of the superset molecule and the dummy region. ΔF_{TI} is the reference value computed by trapezoidal thermodynamic integration on the softcore $\widehat{\partial_\lambda U}$ estimator.

Pair	A	B	n_{atoms}	n_{alch}	ΔF (kJ/mol)
Alcohols					
methanol_ethanol	CO	CCO	9	4	−25.14
methanol_propanol	CO	CCCO	12	7	−11.53
methanol_butanol	CO	CCCCO	15	10	−11.13

(continued)

Pair	A	B	n_{atoms}	n_{alch}	ΔF (kJ/mol)
methanol_pentanol	CO	CCCCCO	18	13	-2.85
ethanol_propanol	CCO	CCCO	12	4	$+10.75$
ethanol_butanol	CCO	CCCCO	15	7	$+13.63$
propanol_butanol	CCCO	CCCCO	15	4	$+6.89$
propanol_pentanol	CCCO	CCCCCO	18	7	$+12.19$
butanol_pentanol	CCCCO	CCCCCO	18	4	$+6.46$
Alkanes					
methane_ethane	C	CC	8	4	$+4.45$
methane_propane	C	CCC	11	7	$+5.93$
methane_butane	C	CCCC	14	10	$+9.91$
ethane_propane	CC	CCC	11	4	$+2.80$
ethane_butane	CC	CCCC	14	7	$+6.33$
Alkylbenzenes					
benzene_toluene	c1ccccc1	Cc1ccccc1	15	4	-3.46
benzene_ethylbenzene	c1ccccc1	CCc1ccccc1	18	7	$+1.54$
toluene_ethylbenzene	Cc1ccccc1	CCc1ccccc1	18	4	$+4.47$
benzene_propylbenzene	c1ccccc1	CCCc1ccccc1	21	10	$+5.40$
toluene_propylbenzene	Cc1ccccc1	CCCc1ccccc1	21	7	$+9.95$
ethylbenzene_propylbenzene	CCc1ccccc1	CCCc1ccccc1	21	4	$+5.41$
benzene_butylbenzene	c1ccccc1	CCCCc1ccccc1	24	13	$+6.91$
toluene_butylbenzene	Cc1ccccc1	CCCCc1ccccc1	24	10	$+12.63$
ethylbenzene_butylbenzene	CCc1ccccc1	CCCCc1ccccc1	24	7	$+10.59$
propylbenzene_butylbenzene	CCCc1ccccc1	CCCCc1ccccc1	24	4	$+6.34$
benzene_pentylbenzene	c1ccccc1	CCCCCc1ccccc1	27	16	$+5.09$
toluene_pentylbenzene	Cc1ccccc1	CCCCCc1ccccc1	27	13	$+12.70$
ethylbenzene_pentylbenzene	CCc1ccccc1	CCCCCc1ccccc1	27	10	$+13.38$
benzene_isopropylbenzene	c1ccccc1	CC(C)c1ccccc1	21	10	$+2.23$
toluene_isopropylbenzene	Cc1ccccc1	CC(C)c1ccccc1	21	8	$+5.08$
ethylbenzene_isopropylbenzene	CCc1ccccc1	CC(C)c1ccccc1	21	4	$+5.65$
benzene_tertbutylbenzene	c1ccccc1	CC(C)(C)c1ccccc1	24	13	$+6.26$
Amines					
methylamine_ethylamine	CN	CCN	10	4	-39.26
methylamine_propylamine	CN	CCCN	13	7	-26.43
methylamine_butylamine	CN	CCCCN	16	10	-23.42
ethylamine_propylamine	CCN	CCCN	13	4	$+8.69$
propylamine_butylamine	CCCN	CCCCN	16	4	$+4.84$
Carboxylic acids					
acetic_propionic	CC(=O)O	CCC(=O)O	11	4	$+30.57$
acetic_butyric	CC(=O)O	CCCC(=O)O	14	7	$+33.28$
Cycloalkanes					
cyclohexane_methylcyclohexane	C1CCCCC1	CC1CCCCC1	21	4	$+0.98$
cyclohexane_ethylcyclohexane	C1CCCCC1	CCC1CCCCC1	24	7	$+9.37$
cyclohexane_propylcyclohexane	C1CCCCC1	CCCC1CCCCC1	27	10	$+12.18$
methylcyclohexane_ethylcyclohexane	CC1CCCCC1	CCC1CCCCC1	24	4	$+8.73$
cyclopentane_methylcyclopentane	C1CCCC1	CC1CCCC1	18	4	$+2.35$
Thiols					
methanethiol_ethanethiol	CS	CCS	9	4	-5.73
methanethiol_propanethiol	CS	CCCS	12	7	$+2.77$
methanethiol_butanethiol	CS	CCCCS	15	10	$+3.77$

(continued)

Pair	A	B	n_{atoms}	n_{alch}	ΔF (kJ/mol)
ethanethiol_propanethiol	CCS	CCCS	12	4	+9.23

D. Pilot 47-split LOO results (transformer + atom-count)

Table 2 reports the held-out ΔF error for each of the 47 LOO splits in the pilot corpus (Appendix C) under raw-λ and ℓ_{alch} conditioning, using the small transformer flow ($d = 64$, $L = 4$) and 4-D atom-count pair features described in Section 3.4. Each error is the mean over 9 runs (3 training seeds $\times$ 3 sampling seeds). The "Improvement" column is $(1 - \text{err}_\ell/\text{err}_\lambda) \times 100\%$; positive values are wins for ℓ_{alch}.

Table 2. Per-split LOO held-out error.

Held-out pair	Family	raw-λ err	ℓ_{alch} err	Improvement
methanol_ethanol	alcohol	18.98	19.36	−2.0%
methanol_propanol	alcohol	10.54	8.32	+21.1%
methanol_butanol	alcohol	10.94	8.87	+18.9%
methanol_pentanol	alcohol	4.57	5.17	−13.1%
ethanol_propanol	alcohol	3.39	2.37	+30.3%
ethanol_butanol	alcohol	5.76	7.05	−22.3%
propanol_butanol	alcohol	8.21	3.62	+55.9%
propanol_pentanol	alcohol	9.92	6.50	+34.4%
butanol_pentanol	alcohol	5.65	4.84	+14.3%
methane_ethane	alkane	7.68	4.69	+38.9%
methane_propane	alkane	2.53	3.22	−27.0%
methane_butane	alkane	7.78	9.99	−28.4%
ethane_propane	alkane	14.93	12.43	+16.7%
ethane_butane	alkane	18.56	14.03	+24.4%
benzene_toluene	alkylbenzene	7.55	2.89	+61.7%
benzene_ethylbenzene	alkylbenzene	4.28	1.48	+65.5%
toluene_ethylbenzene	alkylbenzene	1.16	1.08	+7.6%
benzene_propylbenzene	alkylbenzene	1.41	1.24	+11.8%
toluene_propylbenzene	alkylbenzene	4.30	2.68	+37.7%
ethylbenzene_propylbenzene	alkylbenzene	1.94	1.87	+3.8%
benzene_butylbenzene	alkylbenzene	3.19	4.19	−31.4%
toluene_butylbenzene	alkylbenzene	2.74	4.54	−65.9%
ethylbenzene_butylbenzene	alkylbenzene	4.12	2.31	+43.9%
propylbenzene_butylbenzene	alkylbenzene	1.54	1.81	−17.9%
benzene_pentylbenzene	alkylbenzene	1.57	1.12	+28.4%
toluene_pentylbenzene	alkylbenzene	5.65	6.09	−7.8%
ethylbenzene_pentylbenzene	alkylbenzene	6.82	6.95	−1.9%
benzene_isopropylbenzene	alkylbenzene	2.99	2.13	+28.5%
toluene_isopropylbenzene	alkylbenzene	2.40	0.78	+67.7%
ethylbenzene_isopropylbenzene	alkylbenzene	3.24	3.99	−23.3%
benzene_tertbutylbenzene	alkylbenzene	2.23	3.38	−51.7%
methylamine_ethylamine	amine	37.51	35.62	+5.1%
methylamine_propylamine	amine	22.02	3.87	+82.4%
methylamine_butylamine	amine	18.52	21.51	−16.2%
ethylamine_propylamine	amine	4.48	5.21	−16.2%
propylamine_butylamine	amine	4.39	3.06	+30.3%
acetic_propionic	carboxylic acid	27.47	27.86	−1.4%
acetic_butyric	carboxylic acid	34.49	19.85	+42.4%

Held-out pair	Family	raw-λ err	ℓ_{alch} err	Improvement
cyclohexane_methylcyclohexane	cycloalkane	2.56	1.29	+49.4%
cyclohexane_ethylcyclohexane	cycloalkane	3.81	10.12	−165.9%
cyclohexane_propylcyclohexane	cycloalkane	0.93	0.64	+30.9%
methylcyclohexane_ethylcyclohexane	cycloalkane	4.48	4.95	−10.6%
cyclopentane_methylcyclopentane	cycloalkane	1.96	1.14	+41.9%
methanethiol_ethanethiol	thiol	12.48	10.39	+16.7%
methanethiol_propanethiol	thiol	9.86	10.78	−9.3%
methanethiol_butanethiol	thiol	12.02	5.73	+52.3%
ethanethiol_propanethiol	thiol	2.27	4.28	−88.9%

E. Synthetic experiment details

The synthetic Hamiltonian is $U_\lambda(x) = \frac{1}{2}k(\lambda)x^2$ with $k(\lambda) = k_A^{1-\lambda}k_B^\lambda$ on a single 1-D coordinate $x \in \mathbb{R}$. Samples at each λ_k are drawn analytically from $\mathcal{N}(0, k(\lambda_k)^{-1}/\beta)$ at $\beta = 1$; no MD is required. For each pair we use a uniform λ-grid of 11 points and 200 samples per window. The flow is the small MLP of Lipman et al. (2023) ($d = 32$, 3 layers). We run 6 configurations crossing $k_A \in \{1, 5\}$ with $k_B \in \{2, 4, 10, 25, 100\}$, holding out $k_B \in \{2, 10, 100\}$ in turn while training on the remaining two. The reported $+65\%$ is the mean over all $6 \times 3 = 18$ (configuration, seed) runs.

F. Lennard-Jones experiment details

The LJ system is a single-particle alchemical perturbation: a fluid of fixed Lennard-Jones particles surrounds an alchemical particle whose LJ σ parameter is varied by setting $\sigma_B = (1 + r)\sigma_A$ with $r \in \{0.05, 0.10, 0.15, 0.20, 0.30\}$. We run 21-window vacuum LJ MD per pair on cloud CPU containers and apply the same softcore + finite-difference $\partial_\lambda U$ pipeline as for the real-chemistry experiments. Training holds out the $r = 0.30$ pair (a $3\times$ extrapolation in σ beyond the largest training value) and trains on the other four; we report the mean across 3 training seeds $\times$ 3 sampling seeds.

G. Additional ablations

Table 3 summarizes the 4-config ablation sweep on the 47-pair real-chemistry corpus, crossing flow architecture (MLP vs. transformer; both with the same parameter count) and pair feature (4-D atom counts vs. 192-D Morgan-fingerprint encoding). Numbers are aggregate held-out ΔF error and win rate over the full 47-fold LOO.

Table 3. Architecture and feature ablations on the 47-pair corpus.

Arch.	Features	raw-λ err	ℓ_{alch} err	Wins/47	Improvement
MLP	atom-count	8.25	7.89	30/47	+4.4%
MLP	Morgan FP	4.73	5.36	22/47	−13.5%
Transformer	atom-count	8.21	**6.92**	**28/47**	**+15.7%**
Transformer	Morgan FP	4.70	5.16	20/44	−9.8%

The transformer with atom-count features is our paper-default (reported in Section 5.2). Morgan fingerprint encodings overfit at this corpus size: with 192 features and only 47 training pairs, the conditional flow uses the additional capacity to memorize training pairs rather than discover transferable structure. We expect the relative ranking among feature choices to invert at the $\sim$ 200-pair scale typical of FEP benchmarks (Wang et al., 2015).

H. Cycle-closure values for all closed triples

A thermodynamic cycle $A \to B \to C$, $A \to C$ should close to zero in the absence of MD-statistical or methodological error. Three closed triples in our corpus, with cycle-closure residuals $\Delta F(A \to B) + \Delta F(B \to C) - \Delta F(A \to C)$:

- benzene → toluene → ethylbenzene: $-3.46 + 4.47 - 1.54 = -0.53$ kJ/mol

- methanol → ethanol → propanol: $-25.14 + 10.75 - (-11.53) = -2.86$ kJ/mol
- methylamine → ethylamine → propylamine: $-39.26 + 8.69 - (-26.43) = -4.14$ kJ/mol

All are within typical MD-statistical noise for our sample budget ($\sim$ 1–3 kJ/mol) and an order of magnitude smaller than the pre-fix residual under the linear-Hamiltonian estimator (Section 3.5).

I. pmx benchmark: extraction, mapping, failure analysis

Pair extraction. We extracted 221 pmx pairs by enumerating `deGrootLab/pmx/protLig_benchmark/<system>/ligands_gaf` ligand directories, downloading each `mol_gmx.pdb` via the GitHub API, and using RDKit to read each structure and emit a canonical SMILES. Cross-referencing to `pmx_benchmark.csv` for pair metadata (system, edge name, computational $\Delta\Delta G$) gave 221 pairs with valid SMILES on both sides. Filtering to the six FreeFlow systems (BACE, BACE HUNT, BACE P2, GALECTIN, CMET, P38) left 165 pairs.

Atom mapping. For each pair we identify the maximum common substructure (MCS) between A and B via `rdkit.rdFMCS.FindMCS` with strict element-and-bond matching (`atomCompare = CompareElements`, `bondCompare = CompareOrder`, `completeRingsOnly = True`, `ringMatchesRingOnly = True`). Atoms in the larger of the two molecules that are not in the MCS form the *alchemical region*; their interactions are scaled by λ through softcore (Section 2.1). Strict MCS fails on $\sim 26\%$ of pmx pairs (mostly the BACE family of scaffold-morph perturbations, where A and B have similar but not identical ring systems): there the alchemical region grows to nearly the whole molecule and the resulting dU/dλ profile is uninformative. The strict-MCS subgroup analysis of Section 5.3 reports on this regime explicitly; a principled fix would replace strict rdFMCS with a 3D-aware mapper such as Kartograf (Ries et al., 2024) integrated with OpenFE's hybrid-topology pipeline, which we leave to follow-up.

MD failure modes. Of the 165 pmx pairs, 32 failed during system construction or MD under the strict-MCS mapping: 20 were `OpenMMException: Particle coordinate is NaN` (geometry blowup during equilibration, typically caused by the dummy-atom region having unrelaxable starting overlaps), and 12 were `ChargeMethodUnavailableError` from `antechamber` on charged-species ligands where AM1-BCC parameterization fails. Adding energy minimization (200 iterations, 10 kJ/(mol nm) tolerance) before MD at each λ-window reduced NaN crashes from an initial $\sim 11\%$ to $\sim 5\%$.

J. Cross-pair amortization aggregator

The aggregator script (`scripts/agg_freeflow_comparison.py`) takes per-pair eval JSONs from the 5-fold runs and computes:

- *Per-pair winner*: $|F_{\text{pred,ell}} - F_{\text{true}}| < 0.95 \cdot |F_{\text{pred,lambda}} - F_{\text{true}}|$ counts as an ℓ-win; symmetric for λ; otherwise tie.
- *Aggregate Δ error reduction*: $100\,(\Sigma_i e_\lambda^{(i)} - \Sigma_i e_\ell^{(i)})/\Sigma_i e_\lambda^{(i)}$.
- *Decided win rate*: $n_{\ell\text{-wins}}/(n_{\ell\text{-wins}} + n_{\lambda\text{-wins}})$, with binomial 95% CI via the Wilson interval.
- *Per-system Spearman / Pearson against `pmx_water_leg.csv` reference* (when the `ground_truth_csv` parameter is set in the 5-fold run).

The subgroup script (`agg_subgroup.py`) bins pairs by frac_alch from the original FEP results.json files and runs the same aggregation on each bin.

K. Reproducibility statement

Code and data. All training and evaluation code, including the cross-pair flow implementation, the alchemical FEP simulation pipeline, the pmx-benchmark loader, and the aggregator scripts, will be released as an open-source repository at acceptance. Per-pair raw FEP outputs ($\partial_\lambda U$ samples, integrated ΔF truth) are available in JSON form on a cloud volume; we will additionally publish the parsed eval JSONs (per-pair held-out errors and predicted ΔF samples per seed) on Zenodo.

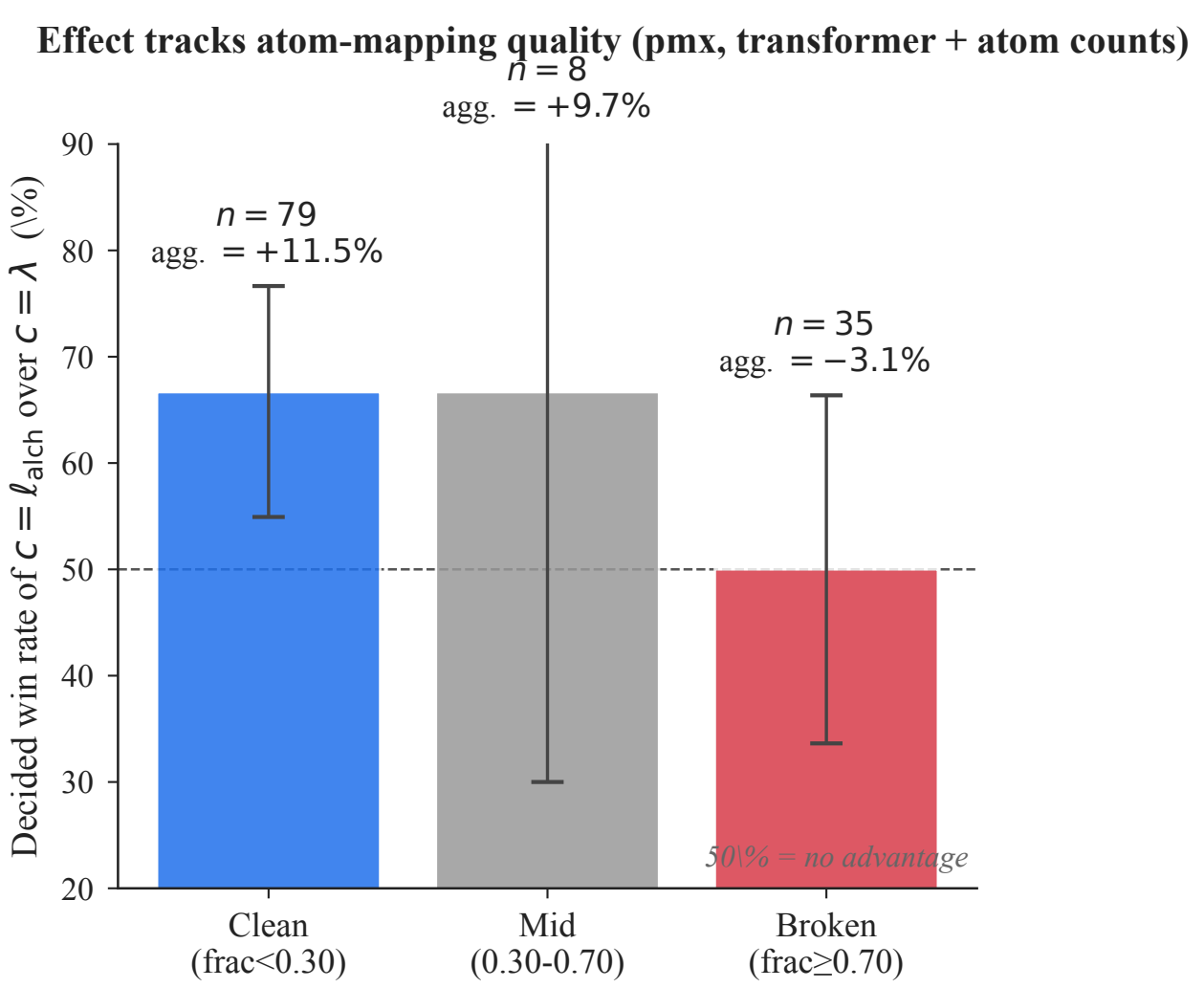


Figure 2. **The ℓ_{alch} advantage scales with atom-mapping quality.** Decided win rate of $c = \ell_{\text{alch}}$ over $c = \lambda$ on pmx drug-like pairs, stratified by frac_alch $= n_{\text{alch}}/n_{\text{atoms}}$. Bars show binomial 95% CI; the dashed line at 50% is no advantage. On the 79 cleanly-mapped pairs (frac < 0.30, large MCS shared core) the geometric coordinate yields a strong advantage; on the 35 broken-mapping pairs (frac ≥ 0.70, dominated by scaffold-morph BACE pairs where MCS fails to find a substantial shared core) the advantage vanishes. This is the predicted mechanism: ℓ_{alch} matters *when geometry is meaningful*.

Compute. All MD ran on cloud CPU containers (4 vCPU, 8 GB RAM, OpenMM 8.1); flow training ran on a single H100 GPU (1 device per 5-fold config). Total compute for the experiments in this paper: ~ 250 CPU-hours of MD and ~ 10 H100-hours of flow training.

Random seeds. Each 5-fold cross-validation cell was trained with 3 flow seeds and evaluated with 3 sample-batch seeds, giving 9 predicted-ΔF samples per held-out pair per conditioning mode (Section J). Seeds were fixed across λ- and ℓ_{alch}-conditioned runs so that any difference between the two stems from the conditioning coordinate alone.

Force field and protocol. Ligand parameterization: openff-toolkit with openff-2.0.0 ("Sage") (Boothroyd et al., 2023) via openff-interchange, AM1-BCC charges via AmberTools `sqm`. All MD runs in vacuum at $T = 300$ K, Langevin integrator with 1 ps^{-1} friction and 2 fs timestep. λ-windows are uniformly spaced on $[0, 1]$ at $K = 11$ (or $K = 21$ in the smaller small-mol diagnostic); 5 000 equilibration steps and 8 000 production steps per window; finite-difference $\partial_\lambda U$ estimator (Section 3.5, $\Delta\lambda_{\text{FD}} = 0.01$).

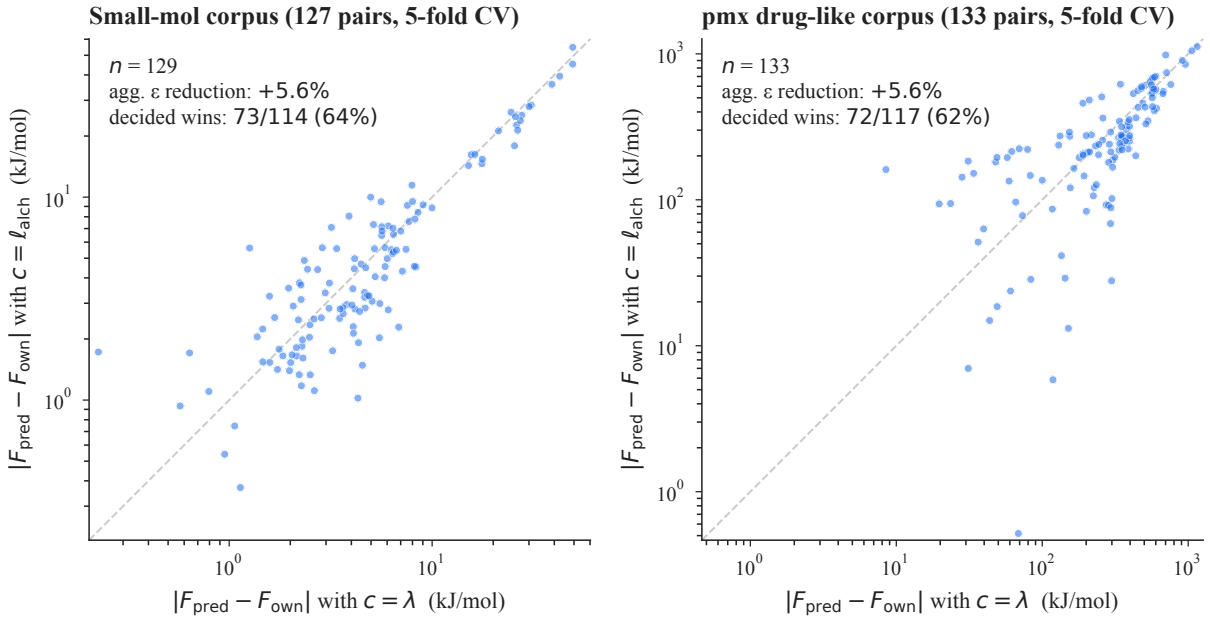


Figure 3. **Cross-pair amortization on two independent corpora.** Per-pair held-out ΔF error: $|F_{\text{pred}} - F_{\text{own FEP}}|$ under $c = \lambda$ (x-axis) versus $c = \ell_{\text{alch}}$ (y-axis). Points below the dashed parity line denote pairs where the geometric coordinate predicts more accurately. Both corpora use the same architecture (transformer + atom counts) and 5-fold cross-validation against held-out pairs. **Left:** 129 small-mol pairs (all substituent extensions, cleanly-mapped by construction); $+5.6\%$ aggregate, $p \approx 10^{-3}$. **Right:** 133 drug-like pairs from the pmx benchmark (Gapsys et al., 2020); the full-corpus aggregate is $+5.6\%$ (an average over a strongly bimodal subgroup distribution), but on the 79 cleanly-mapped pairs (frac_alch < 0.30, our intended operating regime) the advantage is $+11.5\%$. The 35 broken-mapped pairs (frac_alch ≥ 0.70) where the single-topology MCS finds no substantial shared core serve as a built-in negative control (Figure 2).