LATENTSPRING: HARMONIC SOURCES AND PHYSICAL CORRECTIONS FOR MOLECULAR GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating three-dimensional molecules from atomic composition requires learning both connectivity and geometry. We present LatentSpring, a flow-matching model that combines a harmonic-mixture source with learned geometric and physical corrections. The source samples atom positions along random auxiliary trees using element-dependent length scales, without requiring a chemical bond graph. Recovery training teaches the flow to restore perturbed geometries, while two small equivariant networks refine its predicted endpoints using reference structures and force feedback. On 64 unseen compositions, LatentSpring achieves a graph-and-force yield of 35.35% across five fits, compared with 8.75% for GAGA (Qu et al., 2026) and 8.15% for EDM (Hoogeboom et al., 2022), at a GFN2-xTB force-RMS threshold of 5 eV/Å. Adding local geometry checks gives a yield of 9.63%, compared with 4.43% for the initial force-corrected flow. This improvement holds in all three additional training repetitions, and the complete method outperforms a continuation control with matched backbone updates. The endpoint formulation also allows the trained correction networks to improve EDM and GAGA without retraining. Sampling requires only neural-network evaluations, with no energy queries or geometry optimization.

1 INTRODUCTION

Molecular generation from atomic composition requires a model to determine how the atoms connect and how they are arranged in three dimensions. Unlike graph-conditioned conformer generation, the input provides no chemical connectivity. We study this task conditioned on atom identities, total charge, and spin multiplicity, generating coordinates from which bonds are subsequently inferred for evaluation.

Flow matching transports samples from a source distribution to a target distribution (Lipman et al., 2023). The usual isotropic Gaussian source treats all atoms on a common spatial scale. Harmonic sources introduce geometric correlations through virtual springs, but existing graph-conditioned methods construct these springs from known bonds. Composition-conditioned generation therefore calls for a structured source that does not rely on an input molecular graph.

LatentSpring addresses this problem with a mixture of random auxiliary trees. Each tree supplies independent Gaussian edge displacements whose scales depend on the atoms' covalent radii. Adding these displacements along tree paths gives the initial coordinates; averaging over trees defines a normalized harmonic mixture. The tree is discarded after initialization, leaving the flow free to generate a different connectivity, including rings. An equivariant velocity network with geometry self-conditioning then transforms the source into molecular coordinates.

A structured source alone does not ensure accurate local geometry. In our experiments, many outputs pass chemical-graph checks and a residual-force threshold yet retain distorted local arrangements. We address these errors through geometric recovery training and two endpoint-correction networks. Recovery training teaches the backbone to restore perturbed reference geometries. A geometric network then corrects the predicted endpoint, followed by a physical network trained on force feedback. Learning the force correction offline avoids evaluating an atomic potential at every sampling step.

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.

Both networks operate on the generator's predicted final coordinates rather than its internal features. Their coordinate displacements can therefore be converted into either a flow-velocity correction or a diffusion-noise correction. We test this interface by directly reusing FM-trained correction weights in EDM (Hoogeboom et al., 2022) and GAGA (Qu et al., 2026).

Our contributions are:

- A normalized harmonic-mixture source that introduces element-dependent geometric correlations without requiring chemical bonds. We derive its density and compare it with isotropic, shell-tree, and covariance-matched sources.
- A geometric–physical refinement method that combines recovery training with two small equivariant endpoint-correction networks. A common coordinate interface applies the learned corrections to flow matching and diffusion.
- Improved raw molecular geometry across five fits, with gains in all three additional training repetitions and over a continuation control with matched backbone updates. The correction networks also improve EDM (Hoogeboom et al., 2022) and GAGA (Qu et al., 2026) without target-specific retraining.

2 Related work

Molecular flows and source distributions. ConfGF and GeoDiff generate conformations conditioned on a chemical graph (Shi et al., 2021; Xu et al., 2022). HarmonicFlow combines harmonic sources with self-conditioned flow matching for ligand docking, and FlowSite extends the approach to binding-site design (Stark et al., 2024). ET-Flow likewise constructs its harmonic source from a supplied molecular graph (Hassan et al., 2024). Our source instead averages over auxiliary trees conditioned only on atomic composition. Other source-adaptation methods guide a frozen vector field (Wang et al., 2026b) or learn a condition-dependent source (Kim et al., 2026). Our main experiments use the EGNN underlying EDM (Hoogeboom et al., 2022); the earlier source experiments use FlowMol3's coordinate network and self-conditioning modules (Dunn & Koes, 2026).

Distributions over trees. Factored tree distributions support tractable probabilities and learning (Meila & Jaakkola, 2000; Duan & Dunson, 2023). Random spanning-tree differences have also been used to define Gaussian mixtures for imaging inverse problems (Everink, 2026). We use isotropic edge displacements with atomic length scales to define a distribution on centered molecular coordinates. Isotropic Gaussian, shell-tree, and covariance-matched Gaussian controls test the effect of this source construction.

Physical feedback and nonequilibrium work. OMol25 provides molecular structures, energies, and forces across electronic conditions (Levine et al., 2025). Physical information can enter generative training through energy-weighted flow matching (Dern et al., 2025), importance sampling (Midgley et al., 2023), physical-reward fine-tuning (Zhou et al., 2025), or simulation-based protein fine-tuning (Lu et al., 2024). Residual physical correction has also been studied for frozen video flow fields (Wang et al., 2026a). In molecular modeling, FLEXDOCK regularizes predicted docking endpoints with a geometric potential (Corso et al., 2025); ConfDiff learns force guidance for protein conformations (Wang et al., 2024); and Lai et al. (2026) apply MMFF94 guidance during flow and diffusion sampling. Our physical network learns bounded coordinate corrections from forces at generated endpoints while the backbone is fixed.

We also examine local target reweighting through nonequilibrium work identities (Jarzynski, 1997; Hummer & Szabo, 2001; Vaikuntanathan & Jarzynski, 2008; Jarzynski, 2002). Learned escorting has been used for multistate free-energy estimation (Holdijk et al., 2026). Our supplementary studies test local distribution recovery on controlled molecular rotors and its effect on subsequent neural training.

3 A harmonic source with latent connectivity

In our task, the input condition $c = ((z_i)_{i=1}^N, Q, S)$ specifies each atom's atomic number z_i, the total charge Q in elementary-charge units, and spin multiplicity S. For example, neutral singlet C_2H_6O

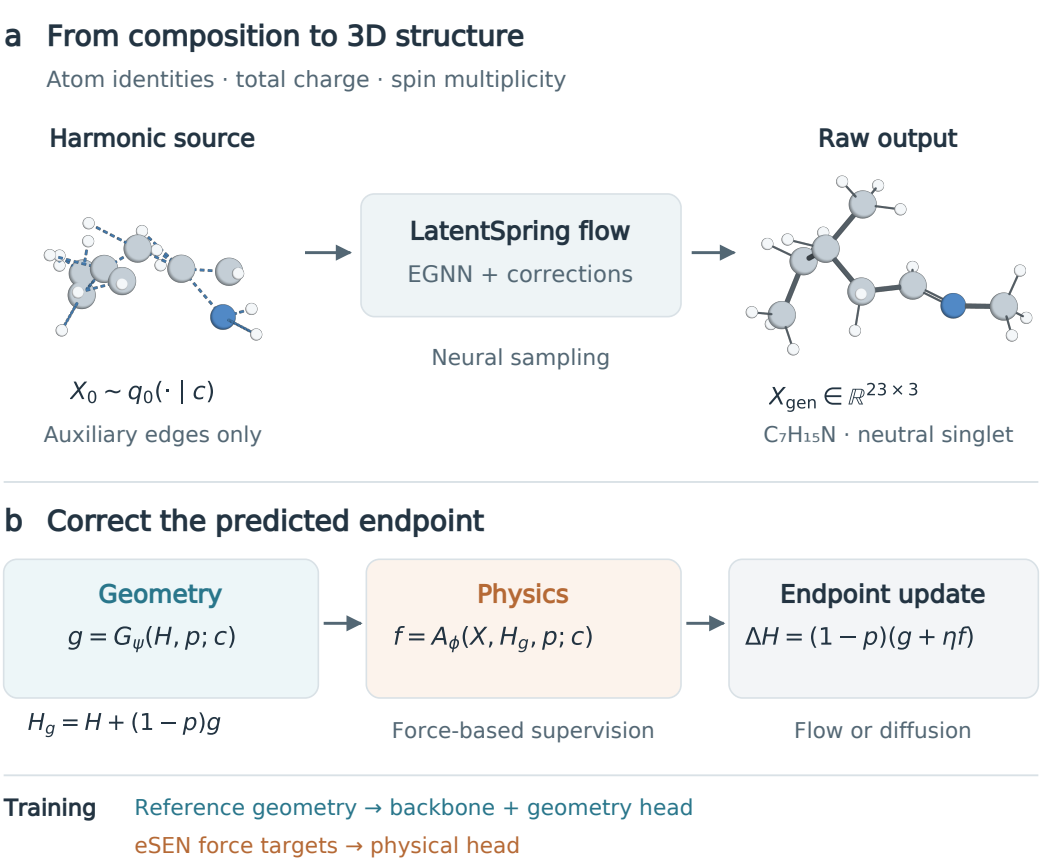


Figure 1: LatentSpring architecture and training. (a) Replayed harmonic source and raw neutral-singlet $C_7H_{15}N$ output, with the same camera and scale. The example comes from the 12-composition common-element validation panel, separately from the main 64-composition benchmark (Appendix D). Dashed links are the sampled auxiliary tree; output bonds are the recorded, coordinate-inferred graph. (b) The geometric head changes the predicted endpoint H to $H_g = H + (1-p)g$. The physical head then receives (X, H_g), giving the total displacement $\Delta H = (1-p)(g + \eta f)$ with $\eta = 4$. Reference geometry trains backbone recovery and the geometric head; eSEN forces at initial-flow endpoints train the physical head. All networks are fixed during sampling, with no energy queries or geometry optimization.

has $Q = 0$ and $S = 1$, but permits several connectivities. Chemical bonds are inferred for evaluation and are not training labels.

The output is a coordinate matrix $X \in \mathbb{R}^{N\times 3}$, with positions x_i in angstroms. Subtracting the mean position removes translation and places X in the $3(N-1)$-dimensional centered space $\mathcal{H}_N = \{X : \sum_i x_i = 0\}$.

Tree distribution. We sample a spanning tree T over the atoms, assigning each possible edge a positive symmetric affinity $a_{ij} = a_{ji}$. The probability of a tree is proportional to the product of its affinities:

$$p(T \mid c) = \frac{\prod_{(i,j)\in T} a_{ij}}{\tau(a)}, \qquad \tau(a) = \sum_T \prod_{(i,j)\in T} a_{ij}. \tag{1}$$

Here $\tau(a)$ sums the weights of all spanning trees. We use $a_{ij} = u_i u_j$, with $u_i = 0.05$ for H and F/Cl/Br/I and $u_i = \max(1, \text{default valence}(z_i) - 1)$ otherwise. Higher values favor more incident edges, while the small hydrogen and halogen weights favor terminal positions. These affinities use element-level valences rather than molecule-specific connectivity.

Coordinate sampling. Each edge has an independent Gaussian displacement $d \in \mathbb{R}^3$. Let ℓ_{ij} be the sum of the two atoms' covalent radii and set $\sigma_{ij}^2 = \ell_{ij}^2 \exp(s^2)/3$, with $s = 0.2$. The edge density is

$$h_{ij}(d) = (2\pi\sigma_{ij}^2)^{-3/2} \exp\left[-\frac{\|d\|^2}{2\sigma_{ij}^2}\right]. \tag{2}$$

Thus σ_{ij} is the standard deviation along each axis, and the covariance matches that of the lognormal shell control. We add edge vectors along tree paths and center the resulting coordinates. The FlowMol implementation uses weighted Wilson walks; the shared EGNN uses weighted Prüfer sequences to sample the same factorized tree law. Drawing a new tree for each sample yields a source invariant to rotations and permutations of identical atoms.

Normalized mixture density. Marginalizing the auxiliary tree gives a mixture of harmonic networks:

Proposition 1 (Normalized source density)**.** *For normalized isotropic edge densities,*

$$q_0(X \mid c) = N^{3/2} \frac{\tau\big(a_{ij} h_{ij}(x_i - x_j)\big)}{\tau(a)} \tag{3}$$

integrates to one over the centered coordinate space $\mathcal{H}_N$.

The numerator weights each tree edge by its displacement density. The Jacobian factor $N^{3/2}$ converts edge vectors to centered atom coordinates. Both tree sums can be evaluated by the matrix-tree theorem; Appendix Q derives the normalization. Training requires only the tree-and-edge sampler, not determinant evaluation. The Gaussian edge densities correspond to quadratic spring energies, hence the term *harmonic source*.

4 Integration with molecular flow matching

For each reference molecule X_1, we sample a starting configuration $X_0 \sim q_0(\cdot \mid c)$ from the harmonic-mixture source and draw $t \sim \mathcal{U}(0, 1)$. We construct an intermediate configuration $X_t = (1-t)X_0 + tX_1$ and train the equivariant network $v_\theta(X_t, t; c) \in \mathbb{R}^{N\times 3}$ to predict the displacement $V^* = X_1 - X_0$. The network receives the current coordinates X_t, the time embedding t, and encoded atom identities from the molecular condition c. The experiments fix total charge $Q = 0$ and spin multiplicity $S = 1$. It outputs one three-dimensional velocity vector per atom.

At inference time, no reference structure is available. We first sample $X_0 \sim q_0(\cdot \mid c)$ using a new auxiliary tree, then integrate $\frac{dX_t}{dt} = v_\theta^{\mathrm{corr}}(X_t, t; c)$ from $t = 0$ to $t = 1$, where the corrected velocity includes the geometric and physical endpoint updates defined below. The final coordinates X_{gen} are the generated three-dimensional structure. The auxiliary tree is used only to sample X_0; it is not passed to the network and does not constrain the generated chemical connectivity.

Training pairs. We pair a source sample X_0 with an OMol25 reference X_1, matching atoms of the same type and aligning the coordinates by rotation. A shared uniform random rotation and random permutation of identical atoms then preserve the source marginal (Appendix R). Along the linear interpolation $X_t = (1-t)X_0 + tX_1$, the regression target is $V^* = X_1 - X_0$. Here $t \in [0, 1]$ denotes interpolation progress.

Geometry self-conditioning. The EGNN is evaluated twice with shared weights. Its first velocity prediction $v^{(0)}$ defines a provisional endpoint $\widehat{X}_1 = X_t + (1-t)v^{(0)}$. Pairwise distances from this detached geometry enter an invariant edge channel in the second pass, which predicts $v^{(1)}$. Both predictions are supervised:

$$\mathcal{L}_{\mathrm{FM}} = \frac{1}{2}\mathbb{E}\left[\frac{\|v^{(0)} - V^*\|_F^2 + \|v^{(1)} - V^*\|_F^2}{3N}\right]. \tag{4}$$

The expectation is over reference structures, source samples, and interpolation times. Normalization by $3N$ gives each molecule equal weight. Gradients train both passes but do not propagate through the distance feedback. The one-pass ablation omits the second evaluation; Appendix F describes the earlier FlowMol implementation.

Geometric recovery training. To improve recovery from sampling errors, we continue backbone training on perturbed late interpolants. Perturbations add local atom noise or displace a spatial region as a group. The recovery objective compares the predicted endpoint coordinates, local distances, and angles with the original reference, alternating with the FM loss. Supervision uses coordinates and element identities. Training-time geometric distortion has also benefited molecular flow models (Dunn & Koes, 2026). Appendix C compares the complete recovery procedure with unperturbed endpoint continuation using the same number of backbone updates.

Sampling. We use 32 midpoint steps. Each step evaluates the velocity at the current state and again at the half-step state; two self-conditioning passes per evaluation give $32 \times 2 \times 2 = 128$ backbone calls. Applying the two endpoint corrections below adds 64 calls to each small network. We return the raw coordinates without terminal noise, energy-based selection, or geometry optimization.

4.1 Geometric and physical endpoint corrections

We train separate networks for geometric recovery and force-based refinement (Figure 1). Both act on the provisional endpoint: the generator's current prediction of its final coordinates. For a flow with current state X, velocity v, and schedule progress $p = t$, this estimate is $H = X + (1 - p)v$.

Geometric correction. For a reference X_{ref} that passes the geometric checks, we construct a perturbed endpoint $\widetilde{H}$ and train $G_\psi(\widetilde{H}, p; c)$ toward $(X_{\text{ref}} - \widetilde{H})/(1 - p)$. Uniform rescaling caps each atom's target norm at $4p^2$ Å per unit progress; unperturbed references receive zero targets. The network uses element identities, pair distances, local coordination, and neighboring element embeddings. Symmetric pair features and opposite vector messages ensure rotation/permutation equivariance and zero total velocity. Appendix C gives the reference screen, perturbations, and architecture.

Physical correction. We collect predicted endpoints from the initial flow and evaluate their centered, mirror-averaged eSEN forces $F_+(H)$. The target coordinate shift is

$$\delta(H) = \frac{\sigma_H^2}{k_B T_{\text{phys}}} F_+(H), \qquad T_{\text{phys}} = 300 \text{ K}. \tag{5}$$

Under a first-order energy approximation, Boltzmann reweighting of a Gaussian centered at H shifts its mean by $\delta(H)$. We choose its local width $\sigma_H \leq 0.03$ Å to keep the molecular displacement norm within 0.1 Å (Appendix U). This local width is independent of the diffusion-noise scale.

The physical network $A_\phi(X, H, p; c)$ predicts a velocity correction from the current coordinates, endpoint estimate, atom identities, and progress. As in the geometric network, symmetric coefficients and opposite pair vectors give an equivariant, zero-sum field. We minimize squared error to $\delta(H)/(1-p)$, with both targets and predictions bounded by $2p^2$ Å per atom per unit progress. Training includes every recorded state, including invalid endpoints, and initializes the final layer to zero. The backbone is fixed during this stage. We subsequently reuse the physical weights with the geometrically refined backbone and in transfer experiments.

Sequential endpoint refinement. The geometric network predicts $g = G_\psi(H, p; c)$, giving $H_g = H + (1 - p)g$. The physical network then evaluates $f = A_\phi(X, H_g, p; c)$. With physical strength η, the total endpoint displacement is $(1 - p)(g + \eta f)$.

For a diffusion model, $H = (X - \sigma_t \epsilon)/\alpha_t$, where ϵ is the predicted noise and α_t, σ_t are the schedule's signal and noise scales. The corresponding native-output corrections are

$$v^{\text{corr}} = v + g + \eta f, \qquad \epsilon^{\text{corr}} = \epsilon - \frac{\alpha_t(1 - p)}{\sigma_t}(g + \eta f). \tag{6}$$

Both produce the specified endpoint displacement. GAGA uses $p = 1 - t/t_{\max}$ with $t_{\max} = 0.65$. Transfer preserves the target's source distribution, reverse transitions, and final observation noise. All sampling-time corrections are neural evaluations.

Appendices N and O detail the architectures, bounds, and fitting protocol. Separate studies examine paired backbone updates (Appendix M) and work-weighted local targets (Appendix T).

Optional hydrogen readout. A conditional flow trained on perturbed reference hydrogens adjusts hydrogen coordinates while preserving relative heavy-atom positions. A detached-hydrogen contact test activates the readout, which uses four midpoint steps and eight network calls (Appendix O.4).

5 Experiments

5.1 Experimental setup

We train the main EGNN on 20,000 OMol25 structures and evaluate 64 unseen neutral organic singlet compositions with 17–40 atoms. Each fit generates 16 structures per composition. The comparison includes Gaussian-source FM, EDM (Hoogeboom et al., 2022), GAGA (Qu et al., 2026), and LatentSpring, which combines the harmonic source, geometric recovery, and endpoint corrections. The main comparison and component ablation exclude the optional hydrogen readout.

We report three nested criteria. *Graph validity* requires an inferred connected, sanitized chemical graph with the requested total charge. *Graph-and-force yield*, abbreviated *joint yield*, additionally requires GFN2-xTB atomic force RMS at most 5 eV/Å. *Geometry-and-force yield* also requires zero assigned radicals and the specified bond-length, angle, internal-distance, and planarity checks. Every rate uses all generation attempts as its denominator, and all structures are evaluated at their raw coordinates. Confidence intervals resample compositions and, where stated, fitted models. Appendices B and G give implementation and evaluation details.

5.2 Comparison with generative baselines

Figure 2 compares raw outputs from LatentSpring, Gaussian-source FM, EDM, and GAGA on the same compositions. The latter three are evaluated without LatentSpring's recovery training or endpoint corrections.

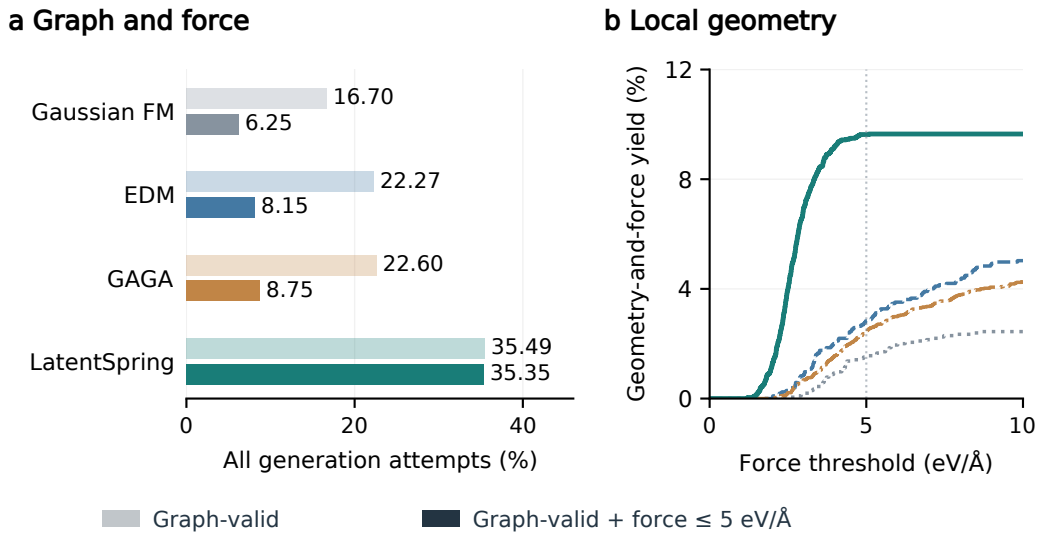


Figure 2: Generation on 64 unseen compositions. (a) Paired horizontal bars show graph validity (pale) and graph-and-force yield (solid) at a force-RMS threshold of 5 eV/Å. (b) Geometry-and-force yield versus force threshold, with a vertical reference at 5 eV/Å. Method colors match across panels. Gaussian FM and EDM have two fits and 2,048 outputs each; GAGA and LatentSpring have five fits and 5,120 outputs. All rates include every attempt, without hydrogen readout.

LatentSpring achieves 35.49% graph validity and 35.35% graph-and-force yield. Joint yields for GAGA, EDM, and Gaussian FM are 8.75%, 8.15%, and 6.25%, giving gains of 26.60, 27.20, and 29.10 percentage points. Under the stricter geometry-and-force criterion, LatentSpring reaches 9.63%, compared with 2.40%, 2.83%, and 1.51%, respectively.

Geometric refinement raises the five-fit geometry-and-force yield from 4.43% for the initial force-corrected FM to 9.63%. All three additional fits improve, with a mean gain of 5.08 points and a crossed fit/composition 95% interval of $[3.09, 7.52]$. In the two-fit control, the complete method reaches 10.25%, compared with 5.76% for unperturbed endpoint continuation with matched backbone updates: a 4.49-point gain, with crossed interval $[2.44, 6.45]$. The control tests the complete refinement procedure, including its geometric head (Appendix C).

On the complementary 12-composition common-element organic panel, geometry-and-force yield is 35.42% for LatentSpring, 20.57% for the initial force-corrected FM, 17.71% for GAGA, 14.84% for EDM, and 16.41% for Gaussian FM. The 14.84-point improvement over the initial FM has crossed interval $[7.81, 25.52]$. Appendix D reports the panel selection, all 384 attempts per method, geometric checks, and charge-assignment diagnostics.

5.3 Ablation of the source and endpoint corrections

We compare Gaussian and harmonic sources with the geometric–physical correction block switched off or on. Both one-pass EGNN models receive the same frozen correction weights within each fit, trained with the self-conditioned FM. This holds correction fitting fixed while testing source choice.

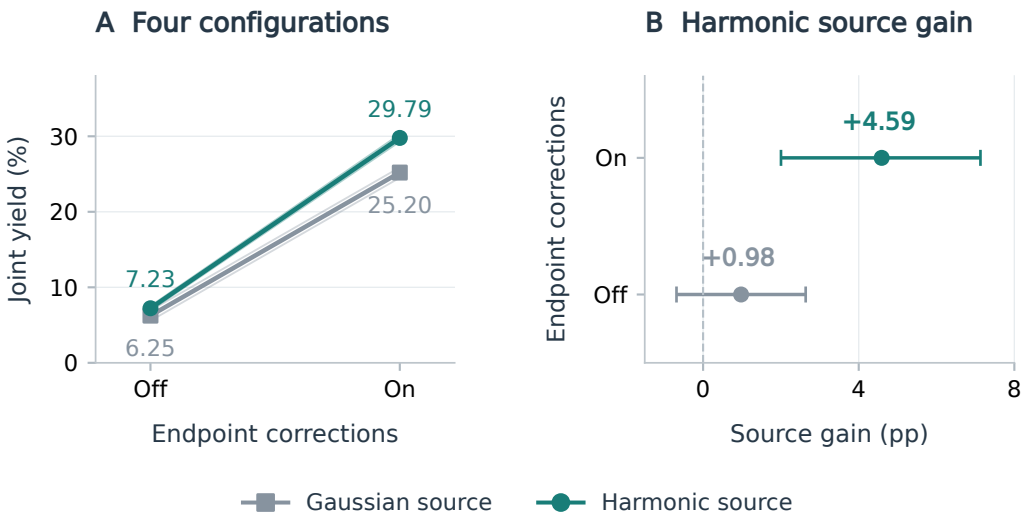


Figure 3: Source and correction ablation. (A) Mean graph-and-force yield, with both fits shown as faint lines. (B) Harmonic-minus-Gaussian gains with nominal crossed fit/composition 95% intervals. Each setting contains 2,048 outputs on 64 compositions.

The corrections increase joint yield from 6.25 to 25.20% for Gaussian FM and from 7.23 to 29.79% for harmonic FM. With correction, the harmonic source adds 4.59 points (crossed 95% interval $[2.00, 7.13]$); without it, the gain is 0.98 points ($[-0.68, 2.64]$).

Geometry-and-force yield increases from 1.51 to 6.40% for Gaussian FM and from 2.10 to 8.11% for harmonic FM. The corrected harmonic–Gaussian difference is 1.71 points, with interval $[-1.46, 4.69]$. Thus the source advantage is resolved for graph-and-force yield but remains uncertain under the stricter geometry criterion. Appendix B reports both fits, and Appendix C separates geometric and physical feedback.

5.4 Transfer across generative models

We transfer both FM-trained correction networks to EDM and GAGA, fixing the correction weights and strengths. Each target retains its native source, backbone, and sampling schedule. Equation (6) converts the endpoint displacement into a noise-prediction correction.

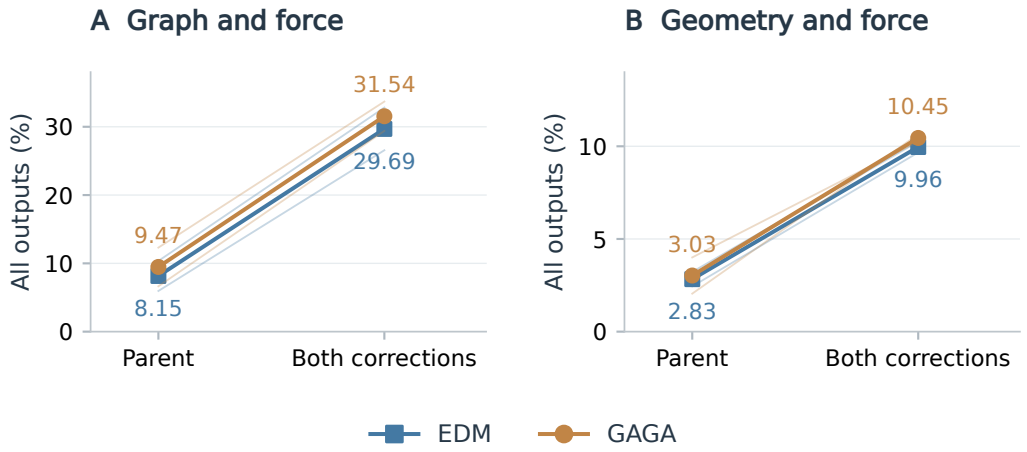


Figure 4: Direct transfer of both correction networks: (A) graph-and-force yield and (B) geometry-and-force yield. Bold lines show two-fit means and faint lines show individual fits. Each setting contains 2,048 attempts on 64 compositions, with the target source and backbone fixed and hydrogen readout disabled.

Joint yield increases from 8.15 to 29.69% for EDM and from 9.47 to 31.54% for GAGA. Geometry-and-force yield increases from 2.83 to 9.96% and from 3.03 to 10.45%, respectively. Both metrics improve in both fits for each target. Together with the one-pass FM results in Figure 3, these experiments show direct correction-weight reuse across generative objectives within the shared EGNN architecture.

Physical-head transfer. In the earlier five-pair study, transferring the physical head alone raised GAGA joint yield from 8.75 to 25.27%, compared with 25.37% using GAGA-trained heads. Across the three additional pairs, the mean transferred-head gain was 16.54 points (crossed 95% interval $[12.66, 20.67]$). Reverse transfer raised initial FM yield from 7.21 to 25.59%. Appendix O.2 reports all fits.

Source transfer and hydrogen readout. Retraining EDM and GAGA with harmonic corruption noise and adding the physical head alone gave joint yields of 0.00% and 0.05%, compared with native-parent yields of 8.15% and 9.47% (Appendix E). In contrast, the successful weight transfers above preserve the target source. The optional hydrogen readout raised joint yield of the initial force-corrected generators to 26.33% for FM and 25.74% for GAGA; it is excluded from the main experiments.

6 Discussion

LatentSpring combines an element-aware harmonic source with learned geometric and physical endpoint corrections. Across five fits, the complete method improves raw graph-and-force and geometry-and-force yields; it also outperforms unperturbed endpoint continuation with matched backbone updates. Direct weight reuse in EDM and GAGA shows that endpoint corrections can remain useful across generative objectives within the shared architecture.

Chemical quality remains the main limitation. Fragmentation, nonplanarity, and unusual inferred charge patterns persist even after geometric refinement. Improving global connectivity and charge assignment is therefore necessary for more reliable composition-conditioned generation.

Ethics statement

This study uses existing molecular datasets and computational experiments. It involves no human participants, personal data, or animal experiments. The supplementary code retains the notices and licenses of included third-party software. Before laboratory use, generated candidates should be assessed for chemical hazards, toxicity, and environmental impact.

Reproducibility statement

Code, trained model weights, and archived per-output scores are provided in the supplementary material. The accompanying README gives commands to generate structures and reproduce the main comparisons. Appendices B and C specify the training and sampling settings, and Appendix Q gives the source-density derivation. Checkpoint and result-file checksums identify the released artifacts.

AI use statement

In this work, we have not used generative AI tools for any of the tasks with required disclosure (including generating synthetic datasets, formulating mathematical claims or proofs, designing re-search methodologies, implementing code, or interpreting results). Additionally, we used generativeAI tools solely for the task with recommended disclosure, specifically to edit the research paper toimprove readability, grammar, and phrasing. All AI-assisted text edits were carefully reviewed andverified by the authors to ensure that the original technical meaning was strictly preserved withoutintroducing any misrepresentations or factual errors. We take full responsibility for the final content of this work, including all text and claims.

References

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. GFN2-xTB — an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15 (3):1652–1671, 2019. doi: 10.1021/acs.jctc.8b01176.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15:3130–3139, 2024. doi: 10.1039/D3SC04185A.

Austin H. Cheng, Chong Sun, and Alán Aspuru-Guzik. Scalable autoregressive 3d molecule generation. *arXiv preprint arXiv:2505.13791*, 2025. doi: 10.48550/arXiv.2505.13791. URL https://arxiv.org/abs/2505.13791.

Gabriele Corso, Vignesh Ram Somnath, Noah Getz, Regina Barzilay, Tommi Jaakkola, and Andreas Krause. Composing unbalanced flows for flexible docking and relaxation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/451dbb8f4fca0327ac4e6782786673bf-Paper-Conference.pdf.

Niclas Dern, Lennart Redl, Sebastian Pfister, Marcel Kollovieh, David Lüdke, and Stephan Günnemann. Energy-weighted flow matching: Unlocking continuous normalizing flows for efficient and scalable boltzmann sampling. *arXiv preprint arXiv:2509.03726*, 2025. URL https://arxiv.org/abs/2509.03726.

Leo L. Duan and David B. Dunson. Bayesian spanning tree: Estimating the backbone of the dependence graph. *Journal of Machine Learning Research*, 24(397):1–44, 2023. URL https://www.jmlr.org/papers/v24/22-0252.html.

Ian Dunn and David R. Koes. FlowMol3: Flow matching for 3D de novo small-molecule generation. *Digital Discovery*, 5(5):2052–2066, 2026. doi: 10.1039/D5DD00363F. URL https://doi.org/10.1039/D5DD00363F.

Jasper Marijn Everink. Random spanning tree markov random field priors for bayesian inverse problems in imaging. *arXiv preprint arXiv:2605.18619*, 2026. URL https://arxiv.org/abs/2605.18619.

Majdi Hassan, Nikhil Shenoy, Jungyoon Lee, Hannes Stärk, Stephan Thaler, and Dominique Beaini. ET-Flow: Equivariant flow-matching for molecular conformer generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4091. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e8bd617e7dd0394ceadf37b4a7773179-Abstract-Conference.html.

Lars Holdijk, Nithishwer Mouroug Anand, Michael Bronstein, and Max Welling. Learning escorted protocols for multistate free-energy estimation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Da8PJXp0js.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887, 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

Gerhard Hummer and Attila Szabo. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proceedings of the National Academy of Sciences*, 98(7):3658–3661, 2001. doi: 10.1073/pnas.071034098.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.04089.

Christopher Jarzynski. Equilibrium free energy differences from nonequilibrium measurements: a master equation approach. *Physical Review E*, 56:5018–5035, 1997. doi: 10.1103/PhysRevE.56.5018. URL https://arxiv.org/abs/cond-mat/9707325.

Christopher Jarzynski. Targeted free energy perturbation. *Physical Review E*, 65:046122, 2002. doi: 10.1103/PhysRevE.65.046122. URL https://arxiv.org/abs/cond-mat/0109461.

Junwan Kim, Jiho Park, Seonghu Jeon, and Seungryong Kim. Better source, better flow: Learning condition-dependent source distribution for flow matching. *arXiv preprint arXiv:2602.05951*, 2026. URL https://arxiv.org/abs/2602.05951.

Helen Lai, Tingyu Wang, Hassan Sirelkhatim, Joe Eaton, Howard Huang, Brad Rees, Ola Engkvist, Jon Paul Janet, Xiaoyun Wang, and Alessandro Tibo. Improving protein-ligand complex generation with force field guidance. *Journal of Cheminformatics*, 18:55, 2026. doi: 10.1186/s13321-026-01198-2. URL https://doi.org/10.1186/s13321-026-01198-2.

Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The Open Molecules 2025 (OMol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762*, 2025. URL https://arxiv.org/abs/2505.08762.

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

Jiarui Lu, Zuobai Zhang, Bozitao Zhong, Chence Shi, and Jian Tang. Fusing neural and physical: Augment protein conformation sampling with tractable simulations. In *GEM Workshop at ICLR*, 2024. URL https://arxiv.org/abs/2402.10433.

Marina Meila and Tommi Jaakkola. Tractable bayesian learning of tree belief networks. Technical Report CMU-RI-TR-00-15, Carnegie Mellon University, 2000. URL https://publications.ri.cmu.edu/tractable-bayesian-learning-of-tree-belief-networks.

Laurence Illing Midgley, Vincent Stimper, Gregor N. C. Simm, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Flow annealed importance sampling bootstrap. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2208.01893. arXiv:2208.01893.

Jingxiang Qu, Wenhan Gao, Ruichen Xu, and Yi Liu. GAGA: Gaussianity-aware gaussian approximation for efficient 3d molecular generation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/08de3c1eb4adc7ac3c1949048422d895-Abstract-Conference.html.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9323–9332. PMLR, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9558–9568, 2021. URL https://proceedings.mlr.press/v139/shi21b.html.

Hannes Stark, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Harmonic self-conditioned flow matching for joint multi-ligand docking and binding site design. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 46468–46494. PMLR, 2024. URL https://proceedings.mlr.press/v235/stark24a.html.

Suriyanarayanan Vaikuntanathan and Christopher Jarzynski. Escorted free energy simulations: Improving convergence by reducing dissipation. *Physical Review Letters*, 100:190601, 2008. doi: 10.1103/PhysRevLett.100.190601. URL https://arxiv.org/abs/0804.3055.

Vikram Voleti, Adam Oberman, and Christopher Pal. Score-based denoising diffusion with non-isotropic gaussian noise models. *arXiv preprint arXiv:2210.12254*, 2022. URL https://arxiv.org/abs/2210.12254.

Cong Wang, Hanxin Zhu, Jiayi Luo, Yonglin Tian, Xiaoqian Cheng, Peiyan Tu, Xin Jin, Long Chen, and Zhibo Chen. Physics-informed video generation via mixture-of-experts latent alignment. *arXiv preprint arXiv:2606.04737*, 2026a. URL https://arxiv.org/abs/2606.04737.

Yan Wang, Lihao Wang, Yuning Shen, Yiqun Wang, Huizhuo Yuan, Yue Wu, and Quanquan Gu. Protein conformation generation via force-guided SE(3) diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 56835–56859. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24cv.html.

Zifan Wang, Alice Harting, Matthieu Barreau, Michael M. Zavlanos, and Karl H. Johansson. Source-guided flow matching. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/adedc33539eafdf3d2c14a5e0032db6e-Abstract-Conference.html.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2203.02923.

Zhijian Zhou, Junyi An, Zongkai Liu, Yunfei Shi, Xuan Zhang, Fenglei Cao, Chao Qu, and Yuan Qi. Guiding diffusion models with reinforcement learning for stable molecule generation. *arXiv preprint arXiv:2508.16521*, 2025. URL https://arxiv.org/abs/2508.16521.

APPENDIX

A NOTATION AND EVALUATION TERMINOLOGY

LatentSpring samples initial coordinates from composition, transports them with a learned flow, and applies geometric and physical endpoint corrections. The table below collects the notation for these stages and the supplementary studies.

Inputs and outputs. The condition c fixes atom identities, charge, and spin. In the experiments, charge and multiplicity are fixed at $Q = 0$ and $S = 1$. At state X_t, the network takes t and encoded atom identities and predicts one velocity per atom. Self-conditioning passes a provisional endpoint from the first evaluation to the second. Integrating the corrected field from a random X_0 gives X_{gen}, with no terminal perturbation in the main FM model. Earlier samplers specify their own terminal-noise settings. Atom identities and electronic conditions are unchanged; bonds are inferred afterward for evaluation.

Table 1: Notation for the source, flow, endpoint corrections, and supplementary physical-update studies.

Symbol	Meaning
N, z_i, Q, S, c	Atom count, atomic numbers, total charge, and spin multiplicity. The condition c collects these inputs.
X, x_i, $\mathcal{H}_N$	Atomic coordinate matrix, the position of atom i, and the space where the mean atomic position is zero. Coordinates are in angstroms.
T, a_{ij}, u_i	Auxiliary spatial tree, its edge affinities, and the atom-dependent factors defining those affinities. These are source variables.
ℓ_{ij}, σ_{ij}, h_{ij}	Characteristic separation from covalent radii, Gaussian displacement scale, and the probability density of one edge vector.
$\tau(a)$, $q_0(X \mid c)$	Sum of tree weights and the normalized distribution of starting coordinates after averaging over trees.
X_0, X_1, X_t, X_{gen}, t	Starting, reference, intermediate and returned coordinates; t labels generation progress, not physical time.
V^*, v_θ	Supervised flow velocity and the predicted velocity field.
$v^{(0)}$, $v^{(1)}$, $\widehat{X}_1$	First and second velocity predictions, and the provisional endpoint passed between them during self-conditioning.
v_{base}, H_t	Frozen parent velocity after both self-conditioning passes, and its extrapolated endpoint used to construct physical training targets.
G_ψ, g, H_g	Geometric correction network, its predicted velocity increment, and the endpoint after that correction.
A_ϕ, f, η	Physical correction network, its predicted velocity increment, and the strength applied to that increment.
$\mathbb{E}$, $\|\cdot\|_F^2$	Average over the specified random draws, and the sum of squared entries of a coordinate or force matrix.
θ_{base}, $\theta_{\mathrm{physical}}$, θ_{replay}, θ_{final}	Base weights, weights after physical or replay fine-tuning, and the final generation weights.
A, q_A, σ_A, δ_A	Generated anchor, local Gaussian reference, perturbation width, and force-guided shift.
E_+, F_+	Mirror-symmetrized teacher energy and force, in eV and eV/Å. The force is the negative coordinate gradient of this energy.
$k_B T_{\mathrm{phys}}$, W_A	Thermal energy at the teacher temperature and generalized work used to weight local candidate structures.
Φ, J, Z	Coordinate map, its local volume-change factor, and a target-density normalizing integral.

Evaluation terminology. *Contact support* requires a connected contact graph without excessive overlap. *Graph validity* additionally requires a sanitized inferred chemical graph. *Geometry-and-force yield* applies the bond-length, angle, internal-distance, planarity, and radical screen in Appendix D, together with force RMS at most 5 eV/Å. *Joint graph-and-force yield* uses the same force threshold without the stricter geometry screen. In the paired-update studies, *jointly valid* instead denotes a pair

of outputs that both pass contact and graph checks. The teacher's *same-graph* restriction preserves an anchor's labeled bonds, formal charges, and radical counts.

Energy differences are compared within composition and reported per atom. Force RMS is $[N^{-1}\sum_{i=1}^{N}\|F_i\|_2^2]^{1/2}$. It measures residual atomic forces, whereas energy measures preference under the chosen potential and graph validity measures consistency of inferred connectivity.

B Implementation and analysis of the main experiments

Model and training. The shared EGNN has four message-passing layers, 256 hidden features, and 2,381,566 backbone parameters. All experiments use neutral singlets, so only atom identities vary in the learned conditioning. The main model uses the harmonic-mixture source with edge log-width 0.2 and distance self-conditioning. Initial training uses AdamW for 15,000 updates at learning rate 10^{-4} and batch size 32, with two supervised passes per example and EMA decay capped at 0.999. The complete model adds 4,000 recovery updates and 10,000 geometric-head updates (Appendix C). The 20,000 training structures contain 8–64 atoms; evaluation covers 17–40 atoms, within the implementation limit of 200 (Appendix L).

The 7,106-parameter physical head uses 16-dimensional atom embeddings and a width-64 pair network, bounded by $2p^2$ Å per atom per unit progress. For each fit, 256 native trajectories across 128 training compositions yield 512 late endpoints and require 1,024 mirror-paired eSEN queries. We train on 96 compositions and assess target-prediction error on the other 32. AdamW runs for 20,000 updates at learning rate $3 \cdot 10^{-4}$. Targets use $T_{\mathrm{phys}} = 300$ K, local width at most 0.03 Å, and molecular displacement norm at most 0.1 Å. Strength four was fixed before the five-fit confirmation and additional-baseline evaluation.

Sampling and compute. The main sampler uses 32 midpoint steps, requiring 128 backbone, 64 geometric-head, and 64 physical-head calls per output. Figure 2 evaluates raw coordinates before hydrogen readout, terminal perturbation, or optimization. Initial backbone training matches example-forward counts across methods; LatentSpring adds recovery, head fitting, and force-label preparation. All methods use 128 backbone sampling calls, while total training, FLOPs, and runtime differ. The continuation control matches the extra backbone updates. Archived zero-strength replay tests execute unused head calls for verification; those calls are unnecessary for an unaugmented baseline.

Evaluation and repetitions. The 64-composition panel contains 20 compositions with 17–28 atoms and 44 with 29–40, all absent from the processed training and validation corpora. Each fit generates 1,024 outputs. Figure 2 includes five LatentSpring and GAGA fits and two archived Gaussian FM and EDM fits. The three additional FM/GAGA fits follow the fixed method; the first two informed development (Appendix O.2). The later Gaussian FM and EDM evaluation reuses the panel with settings fixed before generation. Evaluation energies are not used to select checkpoints or outputs.

Component comparison. Figure 3 uses the two archived one-pass Gaussian and harmonic FM parents from Appendix L, without self-conditioning or recovery continuation. Within each fit, both receive the same frozen self-conditioned-FM correction weights at geometric strength one and physical strength four. Corrected sampling uses 128 calls to the backbone and to each head. Source draws and sampler randomness are paired with the uncorrected outputs.

Table 2: Joint graph-and-force yield in the four-setting analysis (percent).

Source	Both corrections	Fit 1	Fit 2
Gaussian	No	5.57	6.93
Harmonic	No	6.93	7.52
Gaussian	Yes	24.51	25.88
Harmonic	Yes	29.30	30.27

We compute paired intervals from 10,000 resamples, preserving the two size strata. With both heads, the harmonic-source joint-yield gain has crossed fit/composition interval $[2.00, 7.13]$; the

geometry-and-force interval is $[-1.46, 4.69]$. Failed physical calculations remain in the denominators. Two-head transfer across Gaussian FM, harmonic FM, EDM, and GAGA adds 8,192 raw outputs and fixed-coordinate GFN2 attempts.

Structured diffusion. In Equation 7, B_T is the centered tree embedding scaled by edge standard deviations, and C_T has rank $N-1$. Let D_T take oriented edge differences divided by those standard deviations. The loss $\|D_T(\widehat{n} - B_T\Xi)\|_F^2/[3(N-1)]$ is the inverse-covariance-weighted noise error on the centered subspace. The second invariant edge channel receives $\log[1+3(C_{T,ii}+C_{T,jj}-2C_{T,ij})]$, adding covariance information without learned parameters or observed bonds.

EDM initializes from $B_T\Xi$. GAGA initializes from a Gaussian with covariance $\alpha_t^2 vP + \sigma_t^2 C_T$ conditional on T, where v is its original training-only coordinate variance and $P = I - \mathbf{1}\mathbf{1}^\top/N$. Both structured backbones use the baseline initializations, 20,000 references, batch schedules, and 30,000 one-pass updates. The FM physical heads remain fixed. This experiment retrains the diffusion backbones for the new source; Appendix E reports the results.

Physical-head-only evaluation and checks. The earlier experiment transfers the same FM physical head to Gaussian FM, harmonic FM, and EDM in both archived fits, adding 12,288 raw outputs and GFN2 attempts without further training or eSEN queries. Released coordinate hashes, independently parsed physical logs, score arrays, and summaries record these results. A 48-output replay verifies that zero-strength correction reproduces native sampling. Table and contrast scripts read the released summaries and per-output arrays directly.

C GEOMETRIC RECOVERY TRAINING AND CORRECTION

Backbone continuation. Each initial FM fit receives 4,000 additional AdamW updates at learning rate 10^{-5} and batch size 32, with gradient-norm limit one and EMA decay capped at 0.999. Odd updates use the original two-pass FM loss. Even updates draw progress uniformly from $[0.60, 0.95]$ and regress both endpoint predictions to the paired OMol25 reference. The continuation control shares these batches, progress values, and update counts but omits added perturbations and local geometry losses.

Recovery updates perturb the late interpolant. Each atom receives Gaussian noise independently with probability 0.25; its standard deviation is chosen uniformly from 0.05, 0.15, and 0.30 Å per molecule. With probability 0.5, a random spatial half of the molecule also receives a shared Gaussian displacement with standard deviation 0.35 Å. We subtract the mean displacement and retain the original reference as the target.

The recovery loss weights endpoint mean squared error by four, local distance error by 0.5, and local angle error by 0.1. For reference distance d_{ij} and radius sum ℓ_{ij}, distinct-atom pairs have weight $w_{ij} = \mathrm{sigmoid}[(1.30 - d_{ij}/\ell_{ij})/0.10]$. Distance errors use distances normalized by ℓ_{ij}; angle errors compare neighbor-direction cosines with weight $w_{ij}w_{ik}$ for distinct i, j, k. Each local loss is normalized by its weight sum within a molecule, then averaged over molecules and both passes. All targets are derived from reference coordinates rather than bond labels.

Geometric-head targets. A fixed hash order selects 4,096 of the 20,000 training rows; 3,783 pass the graph, connectivity, overlap, zero-assigned-radical, and geometry screen. Selection outcomes are recorded, and evaluation compositions are excluded. Inferred graphs are used for screening only; network inputs are coordinates, elements, and progress.

We draw p uniformly from $[0.4, 0.95]$ and perturb the reference with centered Gaussian noise of scale $s(p) = 0.06 + 0.30(1-p)$ Å per coordinate. Fifteen percent of examples retain the original coordinates and receive zero targets. The remaining targets are $(X_{\mathrm{ref}} - \widetilde{H})/(1-p)$, uniformly rescaled to cap the largest atomic norm at $4p^2$ Å per unit progress while preserving zero total velocity. Squared vector errors have weight 0.25 for hydrogen and one for heavy atoms, normalized by the molecular weight sum. Training uses 10,000 AdamW updates at learning rate 10^{-3} and batch size 32, gradient-norm limit one, and EMA decay capped at 0.995. Sampling uses the final EMA.

Geometric-head architecture. The 7,747-parameter head contains a 16-dimensional element embedding, a learned linear neighbor-embedding update, and two width-64 SiLU layers. Contact weights are $\mathrm{sigmoid}[(1.30 - d_{ij}/\ell_{ij})/0.20]$. Each node combines its element embedding with contact-weighted neighbor embeddings. Symmetric pair sums and products are concatenated with distances, coordination summaries, six radial basis functions, the radius sum, and p, p^2.

Three bounded coefficients scale the normalized relative vector and two bounded radial variants. Each pair contributes opposite vectors to its endpoints. Normalizing by three times the maximum weighted degree, floored at one, and multiplying by $4p^2$ bounds each atomic correction by $4p^2$. The final layer starts at zero. Tests check rotation, reflection, permutation, and centering. A parameter-matched variant with directional moments and nonradial vectors did not outperform the radial head during development.

Combined correction and cost. The geometric head first updates the endpoint at strength one. The physical head then uses that endpoint and the current trajectory state at strength four. Sampling fixes both heads and the backbone, using 128 backbone calls and 64 calls to each head. We reuse the original 7,106-parameter physical head, so geometric refinement requires no additional eSEN labels. Per fit, recovery adds 256,000 backbone example-forwards, while geometric-head training adds 320,000 small-network example-forwards and no backbone evaluations.

Development. We compare recovery and geometric-head variants on 24 training-disjoint validation compositions, split equally between 17–28 and 29–40 atoms. Each of two fits generates 16 samples per composition. Table 3 reports the development sequence. We select the radial head and recovery with local geometry supervision, then fix all choices before generating new outputs on the existing 64-composition benchmark. That benchmark had already been used in earlier project work.

Table 3: Geometric-recovery development results, with 768 raw outputs per row. The last column additionally requires GFN2 force RMS at most 5 eV/Å.

Setting	Graph	Geometry	Geom. + force
Initial FM + physical head	32.55	9.11	8.85
Continuation control + physical head	36.07	10.94	10.55
Perturbed recovery + physical head	41.67	10.94	10.68
Recovery with local geometry + physical head	43.62	12.11	11.85
Initial FM + radial geometry + physical head	36.33	11.33	11.33
Initial FM + moment geometry + physical head	36.98	10.94	10.94
Recovery with local geometry + both heads	44.40	14.45	14.32

On the development panel, the combined method adds 5.47 geometry-and-force points over the original model, with fit-specific gains of 4.43 and 6.51. The descriptive composition interval is $[1.82, 9.38]$ and the crossed fit/composition interval is $[1.04, 10.03]$. Incremental comparisons against the individual components have wider intervals containing zero.

Table 4: Geometric and physical corrections on shared recovery-trained parents. Percentages count all 768 attempts per row.

Correction	Geometry	Geom. + force ≤ 2	Geom. + force ≤ 5
None	10.68	0.13	6.77
Physical	12.11	1.56	11.85
Geometric	14.19	1.82	13.54
Geometric + physical	14.45	3.13	14.32

On the 64-composition benchmark, the first two fits improve geometry-and-force yield from 4.88% to 10.25%, compared with 5.76% for matched-update continuation. Each row contains 2,048 attempts. Gains over the original and continued models are 5.37 points ($[3.56, 7.13]$) and 4.49 points ($[2.44, 6.45]$), respectively, with both fits improving. The controls retain their physical head; the complete method additionally uses the geometric head. These comparisons therefore evaluate the full refinement procedure. In a separate 128-output fixed-noise diagnostic, increasing backbone calls

from 128 to 512 changed graph-valid successes from 38 to 39 and left geometry-and-force successes at 13.

Additional fits. We apply the selected recovery and geometric-head procedure to three additional FM parents, retaining their physical heads. Geometry-and-force successes rise from 22, 54, and 51 to 61, 101, and 121 out of 1,024 per fit. The gains are 3.81, 4.59, and 6.84 points, with mean 5.08 and crossed 95% interval $[3.09, 7.52]$. Across five fits, 493 of 5,120 outputs pass geometry and force, compared with 227 for the initial model. Geometry alone accepts 494 outputs, all with distinct inferred connectivities within each composition and fit.

D CHEMICAL GEOMETRY AND ILLUSTRATIVE STRUCTURES

We evaluate a geometric subset of PoseBusters 0.4.4 (Buttenschoen et al., 2024): heavy-atom bond lengths and angles, internal steric distances, aromatic-ring and double-bond planarity, and saturated six-membered-ring nonplanarity. The molecule-configuration tolerances are 25% for bond and angle bounds, 30% for internal distances, 0.25 Å for planar groups, and 0.05 Å for nonplanar rings. Accepted structures must also have zero assigned radical electrons. Conformer-ensemble energy ratios and reference-identity comparisons are excluded.

Table 5 reports this screen and its conjunction with the 5 eV/Å GFN2 force threshold. Coordinates and inferred graphs are unchanged from the original benchmark, and all 5,120 attempts per setting remain in the denominator, including inference failures.

Table 5: Geometric evaluation across five fits, with 5,120 attempts per row. "Geometry" requires graph validity, zero assigned radicals, and the specified local geometry checks.

Method	Graph validity	Geometry	Geometry + force
FM parent	20.12	3.48	1.64
Initial FM + physics	24.86	4.61	4.43
GAGA parent	22.60	4.71	2.40
GAGA + physics	26.25	6.64	6.45
LatentSpring	35.49	9.65	9.63

For the initial FM, physical correction reduces contact-disconnected outputs from 3,092 to 2,702, heavy-atom-disconnected outputs from 2,527 to 2,325, and severe overlaps from 126 to nine. Of the 1,273 graph-valid corrected outputs, 674 fail planarity, 340 fail angles, 237 fail bond lengths, and 296 fail internal distances; these categories overlap. In comparison, 470 of the 512 prepared validation references pass the same screen. Connectivity and local geometry therefore remain substantial sources of error after force correction.

Charge assignment. Among the 494 geometry-passing LatentSpring outputs on the main panel, 466 have assigned formal charges and 378 have charged carbon atoms. The corresponding counts for the 64 panel reference graphs are eight and two. Some functional groups legitimately carry formal charges, but the prevalence of these patterns in generated outputs motivates closer examination of connectivity and charge assignment. We report charge filters as separate descriptive diagnostics rather than incorporate them into the geometric criterion.

Common-element organic panel. We select validation references with 17–28 atoms, elements in CHONF, zero assigned formal charges, and passing geometric checks. Excluding the 24 development and four earlier illustration compositions leaves 13 eligible compositions. A fixed hash order selects the first 12 before generation. They are absent from the 20,000 training rows but belong to the existing validation corpus, defining a selected application-domain panel.

Each of five unchanged methods generates 16 outputs per composition from two fits: 384 attempts per method and 1,920 total. LatentSpring produces 136 geometry-and-force successes, compared with 79 for initial force-corrected FM, 68 for GAGA, 57 for EDM, and 63 for Gaussian FM. The improvement over initial FM is 14.84 points (crossed fit/composition interval $[7.81, 25.52]$).

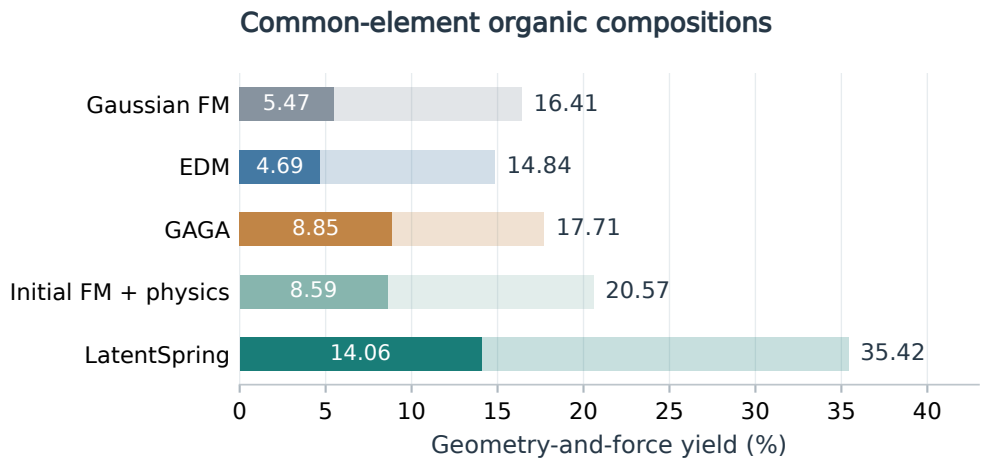


Figure 5: Raw generation on 12 common-element organic compositions. Full bars show geometry-and-force yield; dark portions also require zero assigned atomic charges. Rates use all 384 attempts per method.

Adding a zero-formal-charge filter leaves 54 LatentSpring successes and 33 initial-FM successes, or 14.06% and 8.59% of attempts. Excluding charged carbon gives 20.31% and 13.02%. These filters also exclude some reasonable ionic or charge-separated groups, so they supplement rather than replace the reported criteria.

Illustrated molecule. Figure 1 shows the first full-model output in fit/composition/draw order that passes geometry and force checks with zero assigned atomic charges: fit 1, composition 1, draw 1, using one-based indices. Its formula is $C_7H_{15}N$, its inferred connectivity is `CCC(C)CC=NC`, and its reported GFN2 force RMS is 1.856 eV/Å. It is a raw output from the application panel.

The source is replayed from the original random numbers and rendered with the stored output at the same camera and scale. A PyMOL round-trip check verifies atom identities, bond orders, and coordinate precision. Figure preparation leaves the generated geometry unchanged.

E Combined source and correction transfer to diffusion

Structured diffusion noise. For a sampled tree T, let B_T map independent standard Gaussian edge vectors to centered coordinates using the scales in Equation 2. The covariance is $C_T = B_T B_T^\top$, and the forward process is

$$X_t = \alpha_t X_{\text{clean}} + \sigma_t B_T \Xi, \qquad \Xi \in \mathbb{R}^{(N-1)\times 3}, \quad \Xi_{ij} \sim \mathcal{N}(0,1). \tag{7}$$

The tree remains fixed along each trajectory, and reverse noise uses the same covariance. The denoiser receives invariant pair variances from C_T and predicts coordinate noise $B_T\Xi$, following non-isotropic Gaussian diffusion (Voleti et al., 2022). After training the target backbone for this process, we attach the frozen FM physical head through the endpoint interface. Appendix B gives the covariance channel and GAGA initialization.

EDM and GAGA each use two baseline initialization seeds, the same 20,000 references and batch schedules, and 30,000 updates per fit. Sampling uses the final EMA, 128 backbone evaluations, physical strength four, and no hydrogen readout. Native baselines reuse their original outputs on the same 64 compositions. Each setting contains 2,048 attempts.

This adaptation fails to preserve generation quality. EDM produces no graph-valid outputs. The two GAGA fits produce one and four graph-valid outputs, with only one passing the force threshold. The 36 failed EDM and 21 failed GAGA GFN2 calculations remain in the denominators. The experiment adds 120,000 backbone updates, 3.84 million training-example forwards, and 4,096

Table 6: Original diffusion baselines and combined-design adaptations (percent).

Target	Setting	Graph validity	Joint yield
EDM	Original	22.27	8.15
EDM	Harmonic noise + FM head	0.00	0.00
GAGA	Original	24.22	9.47
GAGA	Harmonic noise + FM head	0.24	0.05

generated trajectories and GFN2 attempts, with no new head fitting or force labels. The release includes every fit, training log, and output array.

F EARLIER SOURCE-STUDY SETUP

Task and training. The source studies adapt a pretrained flow to neutral organic singlets containing C and H, with elements from H/B/C/N/O/F/Si/P/S/Cl/Br/I. Training covers 8–40 atoms and evaluation extends to 64, within the implementation limit of 200. The first two runs share a pretrained checkpoint. Each uses 3,000 distinct OMol25 structures and 3,000 updates, with identical data and initial weights across source variants. The network has 5,904,369 parameters; learning rates are $2 \cdot 10^{-5}$ for the backbone and $3 \cdot 10^{-4}$ for self-conditioning modules. Three further runs reuse the two data blocks with new seeds and evaluate all 1,000-, 3,000-, and 6,000-update checkpoints.

All sources use the same training subset, screened for chemical-graph validity and a spanning scaffold with bounded lengths and degrees. The scaffold is a training-data filter (Appendix G), not a model input.

FlowMol sampler. The coordinate-shaped output D_θ defines velocity through the centered increment $D_\theta - X_t$. In the original midpoint implementation, both field evaluations use the interval midpoint time: the first at the current coordinates and the second at the half-step coordinates. Source-only comparisons add centered Gaussian terminal noise of standard deviation 0.025 Å; learned-correction sampling omits it. These settings apply to the earlier FlowMol studies rather than the main shared-EGNN sampler.

Evaluation panels. The source study uses 12 development compositions, 10 additional-test compositions, a further 24-composition test, and a 64-composition size panel. Exact atom-count checks exclude the latter three panels from both processed OMol25 corpora. Their reference coordinates determine eligibility only. Initial studies use two fine-tuning runs, and the wider source study adds three repetitions. The shared-EGNN models train from scratch; the final correction study uses five paired initializations. Each physical comparison identifies its panel. Appendix H gives selection and source-control details.

Evaluation and uncertainty. The 12- and 10-composition panels use 64 draws per composition; the 24-composition panel uses 32. Contact connectivity and overlap thresholds are 1.25 and 0.6 times the summed covalent radii. Chemical-graph validity additionally requires a connected, sanitized RDKit graph with the requested charge. Rates count all attempts, including failures and validator exceptions. Diversity counts distinct inferred connectivities within composition, and timings cover the full sampling procedure (Appendix G).

Unless specified otherwise, paired 95% bootstrap intervals resample generation draws conditional on the fitted models and compositions. Additional composition-bootstrap intervals describe variation across the evaluation panel. Physical scores use saved raw coordinates, the stated charge and spin, and the specified approximate potential.

G DATA SELECTION AND STRUCTURAL EVALUATION

For each source fine-tuning run, we traverse the processed OMol25 training corpus in a fixed shuffled order, exclude evaluation compositions, and retain the first 3,000 eligible structures. Eligibility

combines the composition and electronic-state requirements in Appendix F with the structural and scaffold checks below.

Training scaffold. Candidate scaffold edges have pair distance, divided by summed covalent radii, strictly between 0.6501 and 1.1999. Kruskal's algorithm constructs a minimum normalized-distance spanning tree. A reference is accepted when the tree is connected and degrees are at most 1 for H/F/Cl/Br/I, 4 for B/C/N/Si, 3 for O, and 6 for P/S. This screening tree is separate from the random source tree.

Bond inference. After contact and overlap checks, we call RDKit 2025.03.2 `DetermineBonds` at the requested charge, using `covFactor=1.25` and `useVdw=True`. We set `allowChargedFragments=True` and `embedChiral=False`, then explicitly sanitize the inferred graph. Acceptance requires successful sanitization, the requested sum of formal charges, and a single connected component. For diversity, we remove stereochemical tags and explicit hydrogens and count distinct canonical SMILES. Rejections and validator errors remain in the attempt denominator.

24-composition confirmation panel. We freeze the Gaussian, harmonic, and two physically updated harmonic models and designate the force-only update as primary before evaluation. Of 256 candidate compositions outside the earlier sets, 121 pass reference-structure checks. Fixed metadata ranks select eight each with 17–20, 21–24, and 25–28 atoms. All 24 are absent from both processed OMol25 corpora; references are used only for eligibility. Four models, two runs, and 32 draws per composition give 6,144 raw outputs.

H Source-development and covariance comparisons

Composition panels. Reference checks retain 31 of 78 metadata-eligible compositions, including 21 used in earlier development. We select 12 for source development and use the remaining 10 for the subsequent test. The latter contain 17–28 atoms and are absent from all 3,902,107 processed training and 39,415 validation rows, as checked by exact atom-count vectors. References determine eligibility only. Both panels use the same two fine-tuning runs; Section M.2 describes the further 24-composition test.

On the development panel, harmonic improves graph validity over isotropic Gaussian by 5.86 points (paired 95% CI [2.67,8.98]; composition interval [0.52,12.17]). Both runs improve, especially run 1. Harmonic also exceeds the shell-tree control by 4.23 points; shell minus harmonic has interval [-7.29,-1.11].

On the 10-composition test, graph validity rises from 52.73% to 61.56%, a gain of 8.83 points (paired CI [5.08,12.50]; composition interval [4.84,12.66]). Contact-geometric validity gains 9.69 points [6.33,12.89]. All graph-valid outputs have distinct connectivities within composition, and no validator errors occur. Sampling time for 640 outputs is nearly identical: harmonic/Gaussian take 175.73/175.43 seconds in run 1 and 178.20/177.74 in run 2.

Covariance-matched control. We approximate the harmonic mixture covariance with 128 sampled trees and average over permutations of identical elements. Relative Frobenius error against an independent 8,192-tree estimate is 1.85–6.25% across the 22 compositions. New single-Gaussian models use this covariance with matched data, optimization, architecture, and sampling budgets, and are compared with the frozen harmonic models.

Harmonic exceeds this Gaussian by 11.65 graph-validity points on development (paired 95% interval [8.59,14.78]; descriptive composition interval [8.07,15.43]) and 15.39 on the additional panel ([11.64,19.06]; [11.95,18.98]). Both runs favor harmonic on both panels. The covariance-matched Gaussian also underperforms the isotropic source. Thus this calibrated second-moment approximation does not account for the harmonic advantage. The comparison adds 2,816 outputs with parameter count, initial self-conditioning weights, references, and optimization fixed.

Table 7: Valid outputs over all attempts for source variants sharing a self-conditioned backbone and 128 sampling calls. The additional panel evaluates checkpoints fixed after development.

Set	Run	Gaussian	Shell tree	Harmonic tree
Development (12)	1	357/768	417/768	436/768
	2	377/768	342/768	388/768
Additional (10)	1	326/640	—	413/640
	2	349/640	—	375/640

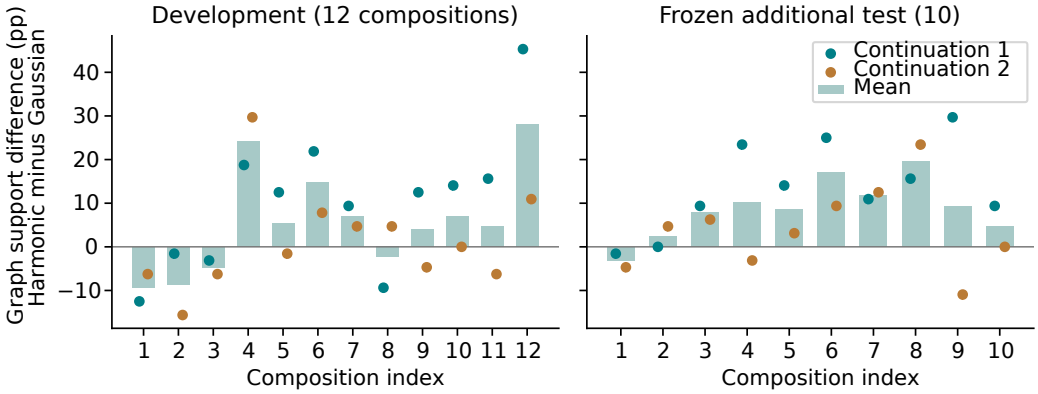


Figure 6: Per-composition harmonic-minus-Gaussian graph validity. Points show individual fine-tuning runs and bars their mean. Both panels use the same fitted models.

I Validity by molecular size

The 64-composition size panel contains 16 compositions per atom-count range. The table reports the pooled Gaussian–harmonic comparison within each range.

I.1 Repeated source training and budget dependence

Three additional paired runs start from the same pretrained FlowMol checkpoint. Runs 3 and 5 use the first 3,000-row block; run 4 uses the second. Each pair changes only the source, sharing initial weights, reference order, and optimization. At 6,000 updates, the 3,000 rows have been visited twice. The study therefore measures stochastic fine-tuning variation on two fixed datasets and one pretrained checkpoint.

We evaluate every new run at 1,000, 3,000, and 6,000 updates on the existing 64-composition panel. The middle checkpoint uses 32 draws per composition, matching the original runs; the other two use 16. All checkpoints are reported.

At 3,000 updates, the five-run mean gain is $+3.89$ points with between-run sample standard deviation 5.27. The composition interval $[2.52, 5.33]$ conditions on those fits. For the three new runs, composition intervals at increasing budgets are $[-4.04, 0.68]$, $[0.36, 3.52]$, and $[1.73, 6.25]$. The source advantage thus depends on training duration and varies substantially across runs. The study adds 36,000 updates and 24,576 outputs, with no physical queries.

J Additional source results

J.1 Effect of the latent spring source

With backbone, initialization, and data fixed, harmonic improves graph validity over isotropic Gaussian by 5.86 points on 12 development compositions, 8.83 on 10 additional compositions, and

Table 8: Covariance-matched Gaussian and harmonic sources at matched training and inference budgets. Entries give valid outputs over all attempts. The Gaussian covariance is estimated from 128 trees.

Set	Run	Covariance Gaussian	Harmonic tree
Development	1	332/768	436/768
	2	313/768	388/768
Additional	1	286/640	413/640
	2	305/640	375/640

Table 9: Raw graph validity by size on 64 compositions, with 1,024 attempts per method and size range across two runs. Rates are percentages; differences are percentage points.

Atoms	Gaussian	Harmonic	Difference
17–28	49.12	55.08	+5.96
29–40	24.02	38.28	+14.26
41–52	16.50	21.29	+4.79
53–64	9.77	12.40	+2.64
All 64 compositions	24.85	31.76	+6.91

8.20 on the 24-composition panel. It also outperforms the tested covariance-matched Gaussian approximation. Appendix H gives the Gaussian, shell-tree, and covariance controls.

J.2 Generalization across molecular sizes

Without further training, we evaluate 64 new compositions, 16 in each of four ranges spanning 17–64 atoms. All are absent from both processed OMol25 corpora and earlier panels. Fixed metadata ranks select eligible references; each fit generates 32 draws per composition.

Pooled graph validity is 24.85% for Gaussian and 31.76% for harmonic. Harmonic is higher in every size range (Appendix I), with overall gain 6.91 points and size-stratified composition 95% interval [4.74,9.03]. The two runs gain 11.18 and 2.64 points. Absolute validity decreases with size, but the pooled source advantage persists beyond the original 17–28-atom range.

Adding three runs gives a five-run 3,000-update mean gain of 3.89 points, positive in four runs, with standard deviation 5.27. For the new runs, the mean advantage is -1.66 at 1,000 updates, $+1.87$ at 3,000, and $+3.94$ at 6,000; all three favor harmonic at the final budget. Appendix I.1 reports all checkpoints.

Table 10: Harmonic-minus-Gaussian graph validity in percentage points. Dashes mark training budgets not run for the original two models.

Fine-tuning run	1,000 updates	3,000 updates	6,000 updates
1	—	$+11.18$	—
2	—	$+2.64$	—
3	-1.37	$+5.66$	$+5.96$
4	-0.10	$+3.32$	$+3.91$
5	-3.52	-3.37	$+1.95$
Mean, new runs 3–5	-1.66	$+1.87$	$+3.94$
Mean, all available runs	-1.66	$+3.89$	$+3.94$

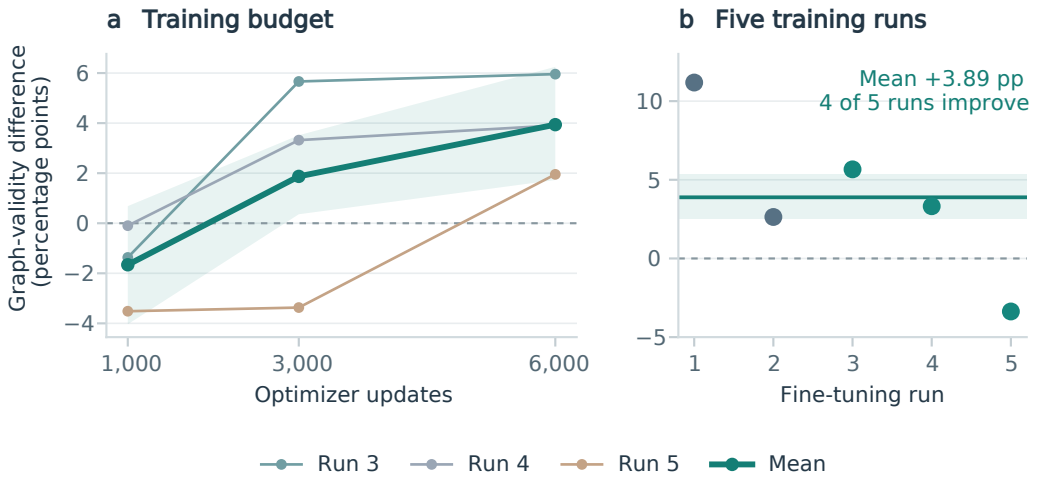


Figure 7: Source gains across training budgets and runs. Thin lines show the three new fits and the thick line their mean. The right panel includes all five runs at 3,000 updates. Shading and the horizontal bar show composition-bootstrap 95% intervals conditional on fitted models.

J.3 Independent backbone and generator objectives

Gaussian FM, harmonic FM, position-conditional EDM, and GAGA (Hoogeboom et al., 2022; Qu et al., 2026) train from scratch on the same EGNN and 20,000 OMol25 structures. Two repetitions share initialization within each comparison, with 30,000 updates at batch size 32. Neither self-conditioning nor physical adaptation is used. The FM models differ only in source, sharing interpolation and atom pairing (Appendix L).

Harmonic improves FM graph validity by 1.34 points (composition 95% interval [0.12,2.52]). EDM and GAGA achieve higher validity in this setting, with different paths, objectives, and samplers. Distance self-conditioning subsequently supplies the FM parent for physical-head comparisons at matched training-forward budgets (Appendices L.1 and O).

A separate 16-composition complete-system comparison gives 38.87% GFN2 joint yield for the inherited 5.90-million-parameter FlowMol backbone and 46.88% with force correction. The 2.38-million-parameter GAGA trained from scratch gives 9.77% at 128 calls and 12.11% at 651. Appendix P details the differing training histories and physical supervision, costs, capacity controls, and fitting-duration study.

Physical weight updates. On 24 new compositions, paired physical training lowers energy and force RMS under both eSEN and independent GFN2 evaluation, with similar graph validity (Appendix M.2). Both fits also favor the paired update over a norm-matched direct update (Appendix M.1). Appendix M.3 compares physical adaptation under both sources.

Training components. At matched training-forward budgets, self-conditioning improves validity by 12.43 points ($[9.70, 15.23]$). Removing the force displacement raises energy on paired graph-valid outputs by 0.01939 eV/atom. Appendix X reports both runs and the update-matched controls.

K Additional generator evaluations

K.1 An independently adapted diffusion baseline

We adapt the official EGNN/EDM backbone (Hoogeboom et al., 2022) to coordinate generation with atom categories fixed as conditions. Positions follow its variance-preserving schedule. Starting from the GEOM-pretrained checkpoint, each run uses the same 3,000 OMol25 references as the

source-model comparison, 30,000 updates, and the original 10^{-4} learning rate. We evaluate the 24 new compositions using 128-call and full 1,001-call samplers, including the final observation step.

The task and adaptation data match, but architectures and pretraining differ: EDM has 2.38 million parameters and one pass per update, versus 5.90 million and two passes for LatentSpring. EDM fine-tuning takes 481.51 and 463.61 seconds, compared with approximately 570–590 seconds for the source models. Generating 768 outputs takes 48.54/47.94 seconds at 128 calls and 377.65/374.12 seconds at 1,001 calls. Thus call counts do not equate runtime. Composition-overlap checks cover the processed OMol25 corpora only. The matched Gaussian–harmonic comparisons provide the controlled source test.

L INDEPENDENT-BACKBONE COMPARISON

We train all methods from scratch on the official EDM EGNN (Hoogeboom et al., 2022): four message-passing layers, 256 hidden features, and 2,381,566 parameters. Two independent initializations are used. Within each repetition, methods share initial weights and minibatches, with atom types fixed throughout.

Data. The shared dataset contains 20,000 OMol25 training structures from 8,323 compositions and 512 internal-validation structures from 299 other compositions. All are neutral organic singlets with 8–64 atoms. Composition hashes assign the split. Selection applies contact-geometric and chemical-graph checks, without the pretrained study's scaffold filter. The 64 evaluation compositions are absent from both selected subsets and both full processed corpora.

Optimization. Each model receives 30,000 AdamW updates at learning rate 10^{-4} and batch size 32, totaling 960,000 training presentations. Batches have a common atom count; sampling count groups proportional to their sizes gives every row equal marginal probability. Gradients are clipped at norm one. Sampling uses the final EMA, whose decay at update s is $\min(0.999, (1+s)/(10+s))$. Validation at 5,000, 15,000, and 30,000 updates monitors training without selecting checkpoints.

Flow matching. Gaussian and harmonic FM share linear interpolation, atom matching, and rotation alignment. Both use one-pass velocity prediction without self-conditioning and 64 midpoint steps, totaling 128 network evaluations. Factorized affinities permit exact tree sampling by drawing $N-2$ independent Prüfer entries with probabilities $u_i / \sum_j u_j$ and decoding the sequence. This gives the same harmonic source law as the main model.

Diffusion. EDM predicts Gaussian coordinate noise under its variance-preserving schedule. Position-conditional GAGA (Qu et al., 2026) uses the published GEOM truncation at 650 of 1,000 steps. Its starting Gaussian has variance $\alpha_t^2 v + \sigma_t^2$, where the training-only estimate $v = 3.0232257866\,\text{Å}^2$ is mean squared centered position per independent coordinate. The truncation is transferred from GEOM rather than selected on OMol25 evaluation results.

Sampling and compute. Each model produces 32 raw structures per evaluation composition using 128 network calls. Full EDM and GAGA schedules are also evaluated at 1,001 and 651 calls, including the observation step. No model in this comparison receives physical or bond labels, output selection, or geometry optimization. Scheduler records identify A100, H200, and A100 MIG allocations. Training presentations and call counts are matched, but runtime depends on the implementation and hardware.

L.1 GEOMETRY FEEDBACK AND PHYSICAL TRANSFER ON THE EGNN

The feedback experiment uses the same 20,000 training rows and two EGNN initializations. A first pass predicts provisional coordinates; their detached pair distances enter an invariant edge channel in the supervised second pass. Training uses 15,000 updates at batch size 32, matching the one-pass models' 960,000 example-forwards. The network contains 2,381,566 parameters, including 1,001 fixed schedule entries. Update counts and runtime differ.

Table 11: Raw graph validity of independently adapted conditional EDM, with 768 attempts per run. The different-backbone harmonic and force-updated models reach 52.73% and 53.45% at 128 calls (Table 17).

Sampler	Run 1	Run 2	Pooled rate
EDM, 128 calls	103	100	13.22%
EDM, 1,001 calls	93	106	12.96%

Table 12: Graph validity on 64 compositions after training from scratch on the shared 2.38-million-parameter EGNN. Each seed generates 2,048 outputs at 128 calls. The final rate pools the per-seed valid-output counts.

Method	Valid, seed 1	Valid, seed 2	Pooled rate (%)
Gaussian FM	182	218	9.77
Harmonic FM	223	232	11.11
EDM	301	216	12.62
GAGA	321	249	13.92

We compare distance feedback with edge-inclusion probabilities from a regularized auxiliary tree distribution. Pooled graph validity on 32 internal-validation compositions selects one feedback type for both seeds. GAGA independently selects its maximum timestep from 350, 500, and 650. Validation chooses distances and timestep 650. We then generate 32 draws per seed for each of 32 previously ungenerated compositions, eight per size range, using 128 network calls and no physical refinement.

Distance feedback exceeds GAGA by 1.86 points (composition 95% interval $[-0.34, 4.44]$), with seed-specific differences $+0.39$ and $+3.32$. Tree feedback is 5.96 points worse than distance feedback ($[-8.11, -3.96]$), and is excluded from the method.

Fixed-recipe physical adaptation. Each selected parent generates 512 candidate anchors on eight training compositions. The same force-target procedure trains physical and replay students: 1,000 two-pass updates for FM or 2,000 one-pass updates for GAGA, at learning rate 10^{-4}. We apply parameter subtraction with coefficient one and evaluate the same 32-composition panel without calibrating this strength on validation.

The unchanged recipe reduces both graph validity and joint yield. FM loses 4.00 joint-yield points ($[-6.49, -2.00]$), and GAGA loses 3.96 ($[-6.25, -1.95]$); eSEN gives the same directions. GAGA retains its fixed noise schedule. These results motivate the separately trained correction head in Appendix O, which uses broader supervision from each parent's trajectory states.

M PHYSICAL FEEDBACK THROUGH PAIRED TRAINING TARGETS

A frozen harmonic generator supplies graph-valid anchors A on training compositions excluded from evaluation. Around each anchor, we sample $X \sim q_A = \mathcal{N}_{\mathcal{H}_N}(A, \sigma_A^2 I)$ in centered coordinate space.

Table 13: Diffusion sampling at 128 calls and at the full reverse schedules. Each initialization generates 2,048 outputs; the final rate pools graph-valid counts across both initializations.

Method	Evaluations	Valid, seed 1	Valid, seed 2	Rate (%)
EDM	128	301	216	12.62
EDM	1,001	308	206	12.55
GAGA	128	321	249	13.92
GAGA	651	273	267	13.18

Table 14: Raw graph validity for every tested feedback configuration, with 1,024 attempts per initialization.

Model	Valid, seed 1	Valid, seed 2	Pooled (%)
Harmonic FM, one pass	107	113	10.74
Harmonic FM, distance feedback	155	165	15.63
Harmonic FM, tree feedback	114	84	9.67
GAGA, selected timestep 650	151	131	13.77

Table 15: Fixed-strength physical adaptation on the EGNN. Joint yield counts graph-valid outputs with GFN2 force RMS at most 5 eV/Å over all 2,048 attempts per model.

Model	Graph validity (%)	Joint yield (%)
FM with distance feedback	15.63	4.79
FM + paired physical update	2.00	0.78
GAGA	13.77	5.86
GAGA + paired physical update	4.79	1.90

The eSEN teacher supplies mirror-averaged energy E_+ and force F_+ (Appendix U). We translate each candidate by $Y = X + \delta_A$, with $\delta_A = \sigma_A^2 F_+(A)/(k_B T_{\text{phys}})$ at $T_{\text{phys}} = 300$ K. Under a locally linear energy, this *force escort* exactly shifts the Gaussian to its Boltzmann-reweighted mean. The width σ_A is fixed before candidate sampling to bound the displacement. We then sample uniformly from candidates that pass geometric checks and preserve the anchor's inferred labeled graph.

Two copies of θ_{base} follow the same fine-tuning schedule and share OMol25 references. The physical student learns from shifted targets, while the replay student learns from the original anchors. Their difference removes the matched replay update:

$$\theta_{\text{final}} = \theta_{\text{base}} + (\theta_{\text{physical}} - \theta_{\text{replay}}). \tag{8}$$

The elementwise parameter arithmetic (Ilharco et al., 2023), with coefficient one, produces a single adapted network. Section M.1 and Appendix U compare it with direct fine-tuning and a norm-matched update.

Work-weighted targets. Complete-work weights account for energy, local Gaussian density, and coordinate-map volume change. Appendices T–V derive the weights and test recovery of specified local distributions. The primary paired-update model instead uses the less expensive uniformly sampled force-shifted targets.

Connectivity and geometry of individual outputs. The first four distinct graph-valid outputs for the first eligible CHON composition in run 1 of the paired-update study have two, two, two, and five independent cycles. The examples were selected in data order without energy ranking or geometry optimization. Only one passes the stricter geometric screen in Appendix D; the others fail angle, planarity, or internal-distance checks. The release contains their coordinates and individual diagnostics. Figure 1 instead illustrates a raw output that passes the stated geometry checks, together with its inferred two-dimensional connectivity.

M.1 Effect of replay subtraction

To distinguish update direction from magnitude, we compare the paired update with direct physical fine-tuning and a norm-matched direct update. The latter is $\theta_{\text{control}} = \theta_{\text{base}} + \alpha(\theta_{\text{physical}} - \theta_{\text{base}})$, where α is the ratio of paired to direct Euclidean parameter-update norms, excluding buffers. The coefficients, computed before evaluation, are 0.2536 and 0.2564.

The norm-matched direct update largely preserves graph validity but raises energy, whereas the paired update lowers it. Paired minus norm-matched direct gives energy change -0.03232 eV/atom ([-0.03981,-0.02485]) and force-RMS change -0.4170 eV/Å ([-0.5529,-0.2859]). Both runs and the all-output comparison favor the paired direction. Graph validity differs by $+0.39$ points ([-1.95,2.73]). The physical benefit therefore depends on update direction, not merely a smaller parameter change.

Table 16: Update controls on 10 compositions, with 1,280 attempts per model. eSEN energy changes are relative to the base model on jointly graph-valid output pairs.

Model	Graph validity (%)	Energy change (eV/atom)
Base	62.11	0
Direct physical update	54.77	$+0.11080$
Same-norm direct update	61.64	$+0.01893$
Paired update	62.03	-0.01554

M.2 New compositions and independent physical evaluation

We evaluate frozen Gaussian, harmonic, force-updated harmonic, and complete-work-updated harmonic models on 24 new compositions. The force-only update was designated primary before the test. Fixed metadata ranks select eight compositions each with 17–20, 21–24, and 25–28 atoms, absent from both processed corpora (Appendix G). The four methods produce 6,144 raw outputs.

Harmonic improves graph validity over Gaussian by 8.20 points (95% CI $[5.08, 11.33]$), with both runs improving. The force update changes harmonic validity by $+0.72$ points ($[-0.98, 2.34]$). All accepted outputs have distinct inferred connectivities within composition.

Physical scores. We evaluate saved coordinates with mirror-averaged eSEN and independent gas-phase GFN2-xTB single-point energies and gradients (Bannwarth et al., 2019). GFN2 is used only for evaluation. Four calculations fail among the 6,144 outputs, all graph-invalid; every graph-valid calculation and reference inversion check succeeds. Force RMS is the root mean squared atomic force magnitude.

The force update lowers energy and force RMS under both potentials on jointly valid pairs, with every composition contributing in both runs. The same directions hold over all eSEN outputs and all paired successful GFN2 calculations. Without physical adaptation, harmonic also lowers aggregate energy relative to Gaussian, mainly in one run.

Complete-work minus force-only energy is -0.00413 eV/atom under eSEN ($[-0.00903, 0.00078]$) and -0.00363 under GFN2 ($[-0.00792, 0.00075]$). Force and graph-validity intervals also include zero. We retain the cheaper force-only teacher because this comparison resolves no additional gain from work weighting.

M.3 Source and physical adaptation

We train paired physical updates for both source families using the same eight training compositions and matched maximum proposal and training budgets. Table 19 compares the four models on 24 compositions; Appendix X.2 details preparation costs and evaluation reuse.

GFN2 joint yield increases by 4.95 points for Gaussian (composition 95% interval [2.86,7.10]) and 3.06 for harmonic ([1.04,5.14]), with both runs improving for each source. On jointly valid pairs, Gaussian adaptation also lowers energy by 0.01465 eV/atom and force RMS by 0.4618 eV/Å.

After adaptation, harmonic exceeds Gaussian by 6.84 graph-validity points ([4.23,9.51]) and 6.97 joint-yield points ([3.26,10.61]). Graph validity improves in both runs, but joint-yield differences are +16.28 and -2.34, showing substantial run variation. eSEN likewise favors adaptation within each source and harmonic in the pooled comparison. Figure 14 in Appendix X shows the threshold dependence.

N Learning a physical correction along the flow

The earlier FlowMol study trains a physical head on predicted endpoints with the backbone fixed. The final EGNN uses the same pair construction with a smaller element vocabulary (Appendix O).

States and targets. The parent generates two trajectories for each of 128 training compositions, with 32 in each of the 8–16, 17–24, 25–32, and 33–40 atom ranges. We record the second field

Table 17: Raw graph validity on 24 new compositions. Each run generates 768 outputs per model at 128 network calls each. Counts are per run; the final rate pools both runs.

Model	Run 1	Run 2	Pooled rate
Gaussian reference	332	352	44.53%
LatentSpring harmonic	435	375	52.73%
With force update	439	382	53.45%
With complete-work update	426	383	52.67%

Table 18: Force-update changes relative to the frozen harmonic model on jointly graph-valid pairs. Negative values indicate lower energy or residual force. Brackets give paired 95% intervals; both fits improve under both potentials.

Potential	Energy change (eV/atom)	Force-RMS change (eV/Å)
eSEN	-0.01308 $[-0.01805, -0.00817]$	-0.34793 $[-0.41498, -0.28478]$
GFN2	-0.01168 $[-0.01616, -0.00724]$	-0.35620 $[-0.42983, -0.28714]$

evaluation at midpoint steps 23 and 29, or times 0.734375 and 0.921875. The resulting 512 states define endpoints $H_t = X_t + (1-t)v_{\text{base}}(X_t, t; c)$ after both self-conditioning passes. These are instantaneous endpoint estimates rather than completed trajectories.

Mirror-averaged, centered eSEN forces at H_t give the displacement in Equation 5, with width at most 0.03 Å and molecular displacement norm at most 0.1 Å. The velocity target is $\delta(H_t)/(1-t)$. We retain all states, including graph-invalid endpoints. Metadata hashes split the parent-training compositions into 96 for fitting and 32 for internal validation, with four states per composition; generator-confirmation compositions are separate.

Equivariant pair network. The small head has 8,178 parameters. Symmetric pair features combine atom embeddings, distances in X_t and H_t, directional alignment, time, atom count, and covalent radii. Two width-64 SiLU layers predict α_{ij} and β_{ij}, each bounded to $[-1, 1]$ by a hyperbolic tangent.

Let $r_{ij} = X_{t,i} - X_{t,j}$ and $\widehat{r}_{ij} = H_{t,i} - H_{t,j}$. Define $\bar{r}_{ij} = r_{ij}/\sqrt{\|r_{ij}\|^2 + \ell_{ij}^2}$ and similarly $\bar{\widehat{r}}_{ij}$, with covalent-radius sum ℓ_{ij}. Contact weights are $w_{ij} = \text{sigmoid}((1.25 - \|\widehat{r}_{ij}\|/\ell_{ij})/0.2)$. With $d_i = \sum_{j \neq i} w_{ij}$ and $d_{\max} = \max(1, \max_i d_i)$, the correction is

$$A_{\phi,i} = v_{\text{cap}}\, t^2 \sum_{j \neq i} \frac{w_{ij}}{2 d_{\max}} \left(\alpha_{ij} \bar{r}_{ij} + \beta_{ij} \bar{\widehat{r}}_{ij} \right), \qquad v_{\text{cap}} = 2 \text{ Å per unit flow time.} \tag{9}$$

Symmetric coefficients and opposite pair vectors give zero total correction and rotation/permutation equivariance (Satorras et al., 2021). The triangle inequality gives $\|A_{\phi,i}\| \leq v_{\text{cap}} t^2$, a bound on instantaneous velocity addition. Its cumulative effect also depends on the parent flow. Zero initialization of the final layer reproduces the parent before fitting.

Two seeds train for 20,000 AdamW updates using mean squared velocity error, learning rate 3×10^{-4}, batch size one, and gradient clipping at one. We also test a 37,540-parameter wider pair network and a 37,586-parameter contextual network with two message-passing rounds and invariant angular summaries. All share the state sequence and targets within each seed.

At 20,000 updates, the contextual head lowers internal-validation target error by approximately 15% relative to the small head and 11% relative to the wide head. The common-panel study tests whether this translates to generation quality. Physical residual correction has also been studied for video flows (Wang et al., 2026a); our supervision uses molecular forces at predicted coordinates.

Sampling and preparation. We integrate $v_{\text{base}} + A_\phi$ for 32 midpoint steps, using 128 backbone and 64 head calls per output without terminal noise or physical queries. A zero-output head matches the parent control's operations. Collecting 512 endpoints uses 256 trajectories and 1,024 mirror-paired eSEN force queries. The shared experiment also evaluates eight translated Gaussian particles per state for a complete-work control, bringing preparation to 9,216 queries; force-target heads do not

Table 19: Source and paired-update comparison under GFN2, with 1,536 raw attempts per model. Force medians use graph-valid, successfully scored outputs. Joint yield requires graph validity and force RMS at most 5 eV/Å, with all attempts in the denominator.

Model	Graph validity (%)	Median force (eV/Å)	Joint yield (%)
Gaussian source	44.53	3.404	35.68
Gaussian + paired update	46.61	3.222	40.63
Harmonic source	52.73	3.287	44.53
Harmonic + paired update	53.45	3.037	47.59

use those particle energies. Median local-work ESS is 1.37 of eight, and the earlier 2,000-update student comparison found no resolved additional generation gain.

O PHYSICAL CORRECTION ON A SHARED BACKBONE

This physical-head-only study precedes geometric recovery. FM and GAGA use the same 2,381,566-parameter EGNN, 20,000 OMol25 rows, head architecture, and physical-supervision budget. Distance-self-conditioned FM receives 15,000 two-pass updates, matching the 960,000 example-forwards of GAGA's 30,000 one-pass updates (Appendix L.1). Each of two independent parents per method receives a separately fitted head.

Endpoint interface. Let p denote schedule progress: flow time for FM and $1 - t/t_{\max}$ for GAGA, with diffusion noise time t and $t_{\max} = 0.65$. The head takes X, predicted endpoint H, atom identities, and p, and predicts the bounded correction from Appendix N. At strength η, the desired endpoint displacement is $\eta(1-p)A_\phi$. The native-output updates are

$$H = X + (1-p)v, \qquad v^{\text{corr}} = v + \eta A_\phi, \tag{10}$$

$$H = \frac{X - \sigma_t \epsilon}{\alpha_t}, \qquad \epsilon^{\text{corr}} = \epsilon - \eta \frac{\alpha_t (1-p)}{\sigma_t} A_\phi. \tag{11}$$

Here v is velocity, ϵ is predicted noise, and α_t, σ_t are signal and noise scales. Both updates induce the stated displacement. All parent tensors, native GAGA transitions, and observation noise remain fixed. With the common 16-element vocabulary, the head has 7,106 parameters and the same hidden pair network as the small FlowMol head.

Supervision and calibration. Each parent generates two trajectories on each of the same 128 training compositions, balanced across four atom-count ranges. Two late evaluations give 512 endpoints: FM progress 0.734375 and 0.921875, and the nearest GAGA progress values $1 - 174/650$ and $1 - 51/650$. All states receive mirror-paired, centered eSEN force labels and target $\delta(H)/(1-p)$. No target reaches the per-atom cap. Each head trains for 20,000 AdamW updates on 96 compositions; the other 32 assess prediction transfer. The four fits lower this error by 29.7–32.1% relative to zero correction.

Eight separate validation compositions select one strength per algorithm across both fits from $\{0, 1/4, 1/2, 1\}$. The criterion is joint yield at force RMS at most 5 eV/Å, with ties favoring smaller strength. Both select $\eta = 1$. We then fix the choice and evaluate 16 new compositions, eight each with 17–28 and 29–40 atoms, drawing 16 outputs per composition and fit. They are absent from the prepared training/validation corpora and earlier evaluation panels.

Physical correction improves FM joint yield by 7.03 points (95% interval $[2.73, 11.91]$), with gains of 6.25 and 7.81 in the two fits. GAGA gains 6.64 points ($[3.71, 9.77]$), also improving in both fits. Thus the method benefits both velocity and noise predictors on the independently trained EGNN.

After matched physical supervision, FM minus GAGA is -2.54 points ($[-6.64, 1.76]$), and the difference between their within-parent gains is $+0.39$ ($[-4.30, 5.47]$). Neither comparison resolves an ordering between the generators. Intervals resample compositions conditional on the two fitted parents and heads, with strengths selected before test generation.

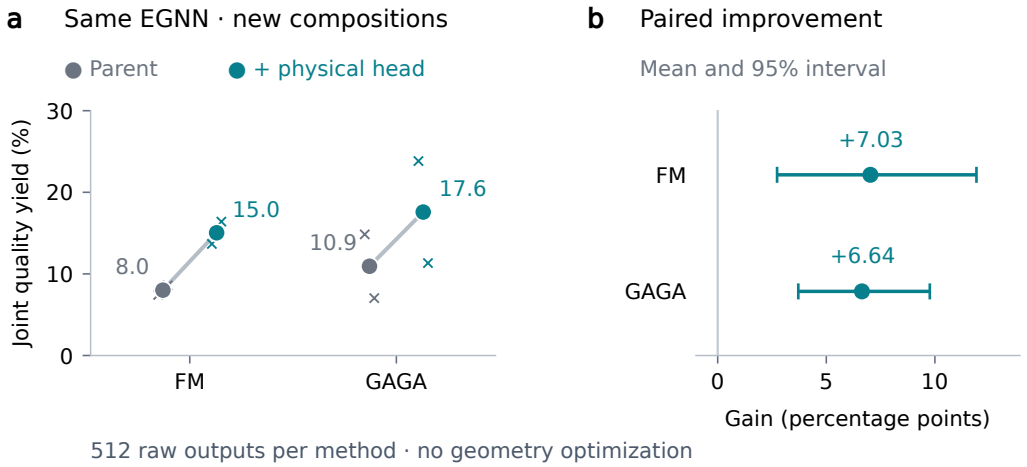


Figure 8: Physical-head training on independent EGNN parents. (a) Pooled joint yields (large circles) and individual fits (crosses) at the 5 eV/Å GFN2 threshold. (b) Within-parent gains with paired composition-bootstrap 95% intervals. Both methods share training compositions, force-query budgets, head architectures, and fitting schedules. Coordinates are unoptimized.

Table 20: Shared-EGNN confirmation on 16 new compositions, with 512 attempts per row. The final columns report joint yield for each parent/head fit. All GFN2 calculations succeed.

Model	Graph (%)	Joint (%)	Fit 1 (%)	Fit 2 (%)
FM, distance feedback	17.58	8.01	7.42	8.59
FM + physical head	20.31	15.04	13.67	16.41
GAGA	18.75	10.94	14.84	7.03
GAGA + physical head	20.31	17.58	23.83	11.33

Cost and verification. Preparation uses 1,024 training trajectories, 4,096 eSEN queries, and 80,000 head updates. Validation and test each add 2,048 outputs and GFN2 calculations. Sampling uses 128 backbone calls plus 64 head calls for FM or 128 for GAGA. Zero-strength controls execute the same head operations. Head-supervision budgets match; FLOPs and runtime are not equated. Replay tests verify native sampling with zero heads and passive observers. Further checks reconstruct targets, measure fitted errors, reevaluate raw graphs, and independently parse xTB gradients. All reported outputs are generated without physical queries or optimization.

O.1 Correction strength and the quality tradeoff

Because both methods selected the largest initial strength, we extend the grid to $\{0, 1/4, 1/2, 1, 2, 4\}$ on the same eight validation compositions and random seeds. Only strengths two and four require new samples. Parents, heads, and labels are unchanged; the applied correction bound scales with η.

A conservative rule limits graph-validity losses relative to strength one to one pooled point and three points per fit. It retains strength one. A second rule, maximizing joint yield alone, selects strength four, with validation yields of 31.64% for FM and 28.91% for GAGA. This rule was introduced after the conservative screen but before new test generation. Both strengths are then evaluated on 32 previously ungenerated compositions, split equally between 17–28 and 29–40 atoms, with 16 draws per composition and fit.

Table 21: Correction-strength comparison, with 1,024 raw outputs per row. Force is the median RMS among graph-valid, successfully scored outputs. Per-output sampling time is measured within the same GPU allocation for each fit, excluding loading and physical evaluation.

Model	Graph (%)	Joint (%)	Force (eV/Å)	Time (s)
FM, $\eta = 1$	20.12	15.14	3.833	0.123
FM, $\eta = 4$	22.17	20.80	3.159	0.122
GAGA, $\eta = 1$	20.80	16.50	3.805	0.143
GAGA, $\eta = 4$	21.00	20.02	3.158	0.143

Increasing strength improves FM joint yield by 5.66 points (composition 95% interval $[2.73, 9.38]$), with fit-specific gains of 7.23 and 4.10. GAGA gains $+3.52$ points ($[0.00, 7.52]$). At strength four, FM minus GAGA is $+0.78$ joint-yield points ($[-4.00, 6.54]$) and $+1.17$ graph-validity points ($[-2.93, 5.86]$). Joint differences of -1.37 and $+2.93$ across fits leave their ordering unresolved. FM samples 14.6% faster in the recorded allocations, using 64 head calls rather than 128, with 128 backbone calls for both.

Over all outputs, stronger correction lowers FM mean GFN2 energy by 0.0622 eV/atom (signed-change interval $[-0.0769, -0.0479]$). Its strength-four energy remains 0.0527 eV/atom above GAGA ($[0.0273, 0.0786]$). These within-composition comparisons include invalid outputs and different inferred connectivities. All graph-valid outputs have distinct canonical connectivities within each 16-draw composition/fit cell, ignoring stereochemistry.

Fragmentation dominates the remaining failures. At strength four, 533 of 1,024 FM outputs and 579 GAGA outputs fail contact connectivity. Only two FM outputs overlap at the specified threshold, both also disconnected; none of the GAGA outputs does. Among graph-valid outputs, 213/227 FM and 205/215 GAGA structures pass the force threshold. The extension adds 1,024 validation outputs and 4,096 confirmation outputs and GFN2 attempts, without new training or eSEN queries. One GAGA strength-one calculation fails and remains in the denominator; the other 4,095 succeed.

O.2 Frozen-method training replication

We repeat the initial physical-head-only models; their later geometric refinement is reported in Appendix C. Architecture, final-checkpoint schedules, strength four, and hydrogen-flow settings are fixed before the new fits. Each FM parent receives 15,000 two-pass updates at batch size 32, and GAGA receives 30,000 one-pass updates. Within a pair, initialization tensors and the batch-schedule prefix match. Backbone example-forward counts match, while optimizer steps, data presentations, and runtime differ. Every model uses its prescribed final checkpoint.

The new metadata pool excludes all compositions from the previous pool. Only 20 previously unused 17–28-atom compositions pass qualification, rather than the planned 32. We therefore select these 20 and the first 44 eligible 29–40-atom compositions by the fixed metadata ranking before generation. The original selection plan and qualification results are archived. All 64 compositions are absent from the processed generator corpora and prepared training/validation rows. Eligibility uses reference geometry and overlap checks, not generated outcomes or energies.

Strength-zero and strength-four trajectories share source draws and sampler randomness. Recorded zero-strength budgets include head calls multiplied by zero. Each corrected structure is evaluated unchanged, after the fixed radial readout, and after the learned hydrogen readout. The hydrogen readout leaves graph-valid parent outputs unchanged. Identical coordinates reuse their physical calculation; failures remain in rate denominators.

The three new pairs form the primary independent-training analysis. A separate five-pair summary includes the two development fits. Composition intervals resample the 64 compositions conditional on fitted models; crossed intervals also resample fit indices. Equal-weight size-stratum averages supplement the pooled 20:44 mixture. Training-variation estimates are based on only three new or five total fits.

Table 22: Joint yield in every fit (percent). Fits 1–2 reuse earlier weights; fits 3–5 are newly trained.

Parent	Module	1	2	3	4	5
FM	None	9.96	7.52	3.61	6.25	8.69
FM	Physical	28.22	26.37	11.72	24.71	29.00
FM	Physical + radial	30.18	27.83	12.70	27.15	32.42
FM	Physical + H flow	30.66	28.03	12.89	27.34	32.71
GAGA	None	12.30	6.64	11.52	8.59	4.69
GAGA	Physical	27.34	24.02	28.61	27.34	19.53
GAGA	Physical + radial	27.93	24.22	29.20	27.83	19.73
GAGA	Physical + H flow	27.93	24.22	29.00	27.83	19.73

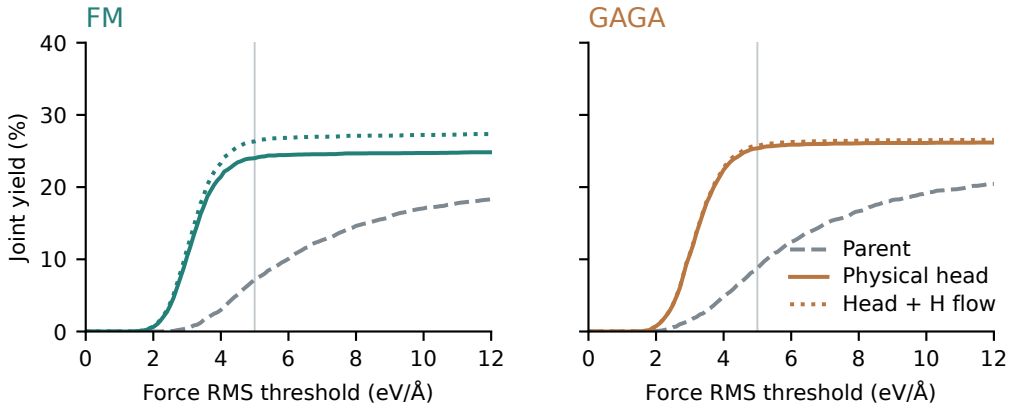


Figure 9: Joint yield versus force threshold, averaged over all five fits. The vertical line marks the prespecified 5 eV/Å criterion; the curves do not select a new threshold.

For FM, physical correction adds 16.80 pooled joint-yield points (crossed interval $[11.50, 21.93]$) and 19.34 equal-stratum points ($[13.85, 24.69]$). GAGA gains 16.62 pooled points ($[12.95, 20.47]$) and 18.46 equal-stratum points ($[14.32, 22.72]$). After hydrogen readout, FM minus GAGA is 0.59 pooled points ($[-8.59, 8.59]$) and 0.95 equal-stratum points ($[-8.93, 9.63]$).

The study adds 20,480 evaluation parent trajectories, 20,480 derived readout outputs, and 22,688 GFN2 attempts. New teachers use 1,536 training trajectories and 6,144 eSEN queries. The three new pairs add 135,000 parent, 120,000 head, and 30,000 hydrogen-model updates. Reused calculations are counted only once.

Six GFN2 calculations initially fail electronic convergence, all on graph-invalid coordinates. Increasing the SCC limit from 250 to 1,000 resolves all six while preserving coordinates, charge, spin, electronic temperature, Hamiltonian, and accuracy. Original failures are archived; completion changes no graph or joint count. All-output energy analyses use the completed scores. Of 1,384 FM pairs valid under both hydrogen readouts, 111 are newly repaired under both. On this descriptive subset, the learned readout lowers energy by 83.42 meV/atom relative to the radial rule.

O.3 Cross-generator weight transfer

The transfer protocol is fixed before generating cross-head outputs and evaluating the three new fit pairs. Fits 1–2 already have target-parent results; fits 3–5 provide the primary independent replication. Each target receives the source family's final 20,000-update head with weights, vocabulary, bounds, and strength unchanged. The endpoint conversion in Equation 6 maps its action to the target sampler. Inputs are coordinates, predicted endpoints, atom identities, and normalized progress, without backbone features.

Table 23: Joint yield of transferred heads in every fit (percent), without hydrogen readout.

Target	Head trained on	1	2	3	4	5
FM	GAGA	29.20	28.81	13.96	26.17	29.79
GAGA	FM	28.42	23.54	27.54	26.95	19.92

In the three new fits, replacing GAGA's own head with the FM-trained head changes joint yield by -0.36 points (composition interval $[-1.66, 0.94]$; crossed $[-2.15, 1.30]$). Replacing FM's head with the GAGA-trained head changes yield by $+1.50$ points ($[0.33, 2.64]$; crossed $[0.00, 3.12]$). Both transfer directions improve over the uncorrected target, while comparisons with target-trained heads are smaller and less certain.

All five pairs and both directions are reported. The study adds 10,240 trajectories and 10,240 successful GFN2 calculations, without head updates or new force labels. A metadata-writing restart reuses checksum-verified coordinates and does not duplicate trajectory counts.

O.4 A conditional flow for hydrogen coordinates

Hydrogen fragmentation can invalidate an otherwise connected heavy-atom skeleton. Among 512 cached training outputs per generator, FM and GAGA produce 346 and 314 connected heavy skeletons, but 60 and 18 outputs have hydrogen-only fragmentation. We therefore train a conditional flow to arrange hydrogens around fixed heavy atoms. Hydrogen decoration is also used by generators such as Quetzal (Cheng et al., 2025); here it is a selective readout of an existing generated structure.

Training. The network takes all coordinates, atomic numbers, and completion time $s \in [0, 1]$, and predicts a velocity for each hydrogen. We perturb reference hydrogens with Gaussian noise of standard deviation 0.4 Å, holding heavy atoms fixed. Half of examples perturb every hydrogen; the other half perturb each independently with probability one half. Minimum squared-distance assignment pairs perturbed coordinates Y_0 with reference coordinates Y_1. The path and loss are

$$Y_s = (1 - s)Y_0 + sY_1, \qquad \mathcal{L}_H = \mathbb{E}\left[\frac{1}{3N_H} \left\| u_\psi(X_{\text{heavy}}, Y_s, s; z) - (Y_1 - Y_0) \right\|_F^2\right]. \tag{12}$$

Here X_{heavy} contains heavy-atom coordinates, z denotes atomic numbers, N_H is the hydrogen count, and u_ψ is the hydrogen velocity. A common translation centers the input. Training uses only coordinates and elements.

The four-layer, width-64 EGNN has 153,790 parameters, including 1,001 fixed schedule entries. Two independently initialized models each receive 10,000 updates at batch size 32 on the shared 20,000 OMol25 references, with learning rate $3 \cdot 10^{-4}$ and EMA decay 0.999. We use the final EMA. These fits require 640,000 example-forwards in total and serve both FM and GAGA.

Readout and controls. The parents use fixed strength-four physical corrections. Readout activates if any hydrogen is farther from every heavy atom than 1.25 times the summed covalent radii;

otherwise coordinates are unchanged. We integrate from $s = 0$ to 1 for four midpoint steps, moving all hydrogens and capping each velocity at 2 Å per unit completion time. Eight small-network calls are required per activated output. Heavy atoms remain fixed, and a final common translation restores the original centroid. No physical evaluator is used.

The radial control moves each detached hydrogen toward its nearest heavy atom to the summed-radius distance, preserving all other relative positions. A second control applies the learned flow only to detached hydrogens. Both start from the same parent coordinates, allowing us to separate simple attachment from joint hydrogen placement.

Selection and confirmation. On 16 development compositions, we compare the radial rule with both learned modes starting at time zero or one half. Moving all hydrogens from zero gives the largest FM joint count, 96/256 versus 95/256 for the radial rule, but not an improvement in both fits. Its lower energy motivates a separate prespecified test: reduced mean energy per atom with retained pooled graph and joint yield. We freeze this hypothesis and model before evaluating 30 new compositions, 15 each with 17–28 and 29–40 atoms. Sixteen draws per composition and fit give 960 attempts per setting. Composition intervals retain both fitted models.

Table 24: Hydrogen readout on 30 new compositions, using fixed strength-four physical heads. Joint yield requires graph validity and GFN2 force RMS at most 5 eV/Å. Energy is the all-output change from each unchanged parent, in meV/atom.

Parent	Hydrogen readout	Graph (%)	Joint (%)	Energy change
FM	None	25.31	24.17	0
FM	Radial rule	27.29	25.94	$+1.89$
FM	Conditional flow	27.40	26.04	-13.32
GAGA	None	23.85	23.54	0
GAGA	Radial rule	24.06	23.75	$+1.05$
GAGA	Conditional flow	24.06	23.75	-2.04

For FM, the learned readout lowers energy by 15.21 meV/atom relative to the radial rule (95% interval $[10.10, 21.71]$), with both fits improving. Graph and joint yield differ by only $+0.10$ points ($[-0.21, 0.42]$). Relative to the unchanged parent, graph validity increases by 2.08 points ($[1.15, 3.13]$) and joint yield by 1.88 ($[1.04, 2.71]$), while every previously graph-valid output is preserved. Moving all rather than only detached hydrogens lowers energy by a further 10.25 meV/atom ($[6.96, 14.72]$) with identical validity counts. GAGA likewise gains 3.10 meV/atom over the radial rule ($[1.04, 6.34]$).

With the shared readout, FM exceeds GAGA in graph validity by 3.33 points ($[0.52, 6.15]$); the joint-yield difference is $+2.29$ ($[-0.63, 5.21]$). Energy averages include all outputs. Among 261 FM pairs valid under both readouts, 18 are newly repaired; on this descriptive subset the learned flow lowers energy by 83.81 meV/atom. These 18 pairs are not used for selection or the primary interval.

Confirmation uses 1,920 parent trajectories, 5,760 derived outputs, and 2,460 new GFN2 calculations: one per parent plus 540 distinct changed geometries. Identical coordinates reuse their scores. Development separately uses 2,560 derived outputs and 155 new GFN2 calculations, in addition to model fitting. All attempts contribute to the reported rates.

P Complete generators on common compositions

We compare complete systems on 16 new neutral organic singlet compositions within GAGA's element vocabulary, eight each with 17–28 and 29–40 atoms. Metadata hashes fix the panel, excluding earlier confirmation panels and both processed generator corpora. Sixteen draws per composition and run give 512 outputs per row of Table 25.

FlowMol variants inherit a 5.90-million-parameter backbone; GAGA uses the 2.38-million-parameter EGNN trained from scratch on 20,000 OMol25 structures. This comparison includes differences in architecture, training, physical supervision, and cost. GAGA uses its validation-selected truncation at 650 with 128-call and full 651-call samplers. The paired-update model retains its two original continuations; head comparisons use two fitting seeds on one shared parent. Intervals resample the 16 compositions conditional on these systems.

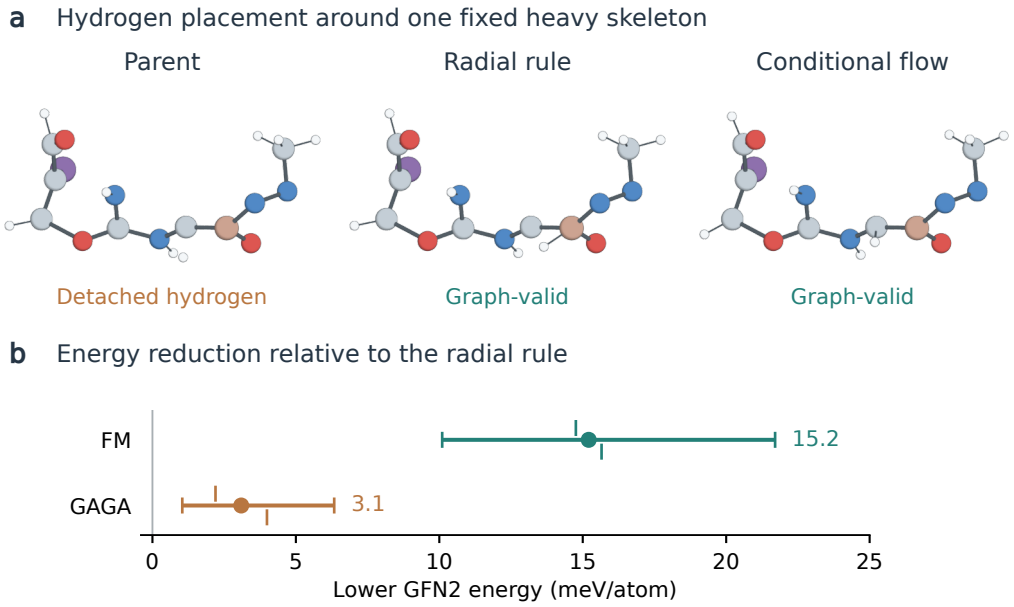


Figure 10: Conditional hydrogen placement. Top: one stored FM output before readout, after radial attachment, and after the learned flow, shown at a common camera and scale. Sticks indicate distance contacts rather than bond orders. A fixed metadata hash selects the example from repaired, jointly graph-valid structures with at most 28 atoms, without energy ranking. Bottom: all-output energy reductions relative to radial attachment, with 95% composition intervals and both fit means. C, H, N, O, B, and I are gray, white, blue, red, tan, and purple.

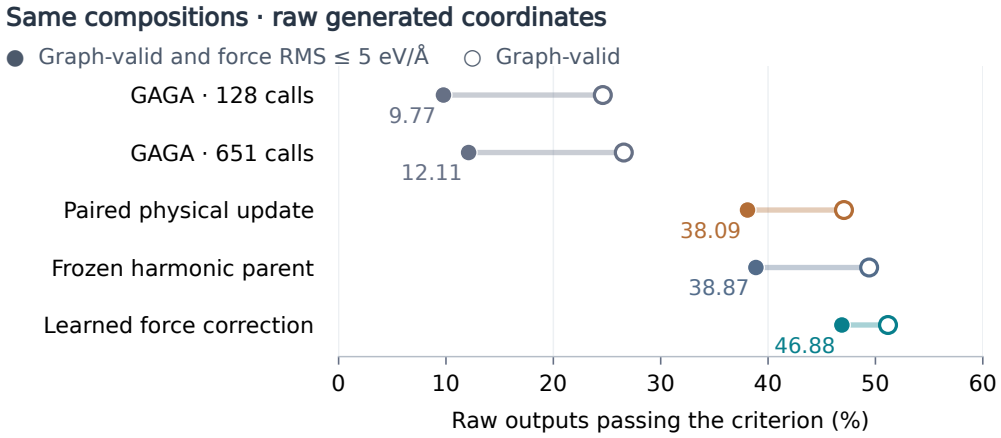


Figure 11: Graph validity and joint yield on the common panel. Open circles show graph validity; filled circles also require GFN2 force RMS at most 5 eV/Å. Connecting segments show the gap between the two criteria. Rates use all 512 attempts. Training histories and inference costs differ across systems, as described in the text.

The prespecified contextual head improves joint yield over its parent by 7.81 points (95% interval $[4.88, 10.94]$), with gains of 7.03 and 8.59 across fits. Its differences from the small and wide 20,000-update heads are -0.20 ($[-1.95, 1.56]$) and $+0.59$ ($[-1.17, 2.34]$). Thus lower internal target error does not yield a resolved advantage over the simpler heads.

The contextual system exceeds GAGA at 128 and 651 calls by 36.91 points ($[26.56, 46.09]$) and 34.57 ($[24.80, 43.95]$). The paired-update system exceeds them by 28.32 ($[20.70, 35.74]$) and 25.98

Table 25: Common-panel results on raw coordinates. Force medians use each model's graph-valid outputs with successful GFN2 calculations; joint yield uses all attempts. "Calls" lists backbone and head evaluations separately.

Model	Calls	Graph (%)	Force (eV/Å)	Joint (%)
GAGA	128	24.61	5.408	9.77
GAGA, complete schedule	651	26.56	5.176	12.11
Paired physical update	128	47.07	3.581	38.09
Paired update, no terminal noise	128	46.88	2.929	40.04
Frozen harmonic parent	128+64	49.41	3.394	38.87
Small correction, 2,000 updates	128+64	50.00	3.077	43.16
Small correction, 20,000 updates	128+64	51.17	2.854	46.88
Wide correction, 20,000 updates	128+64	50.20	2.831	46.09
Contextual correction, 20,000 updates	128+64	50.78	2.864	46.68

($[18.75, 33.40]$). Appendices L.1 and O provide architecture-controlled comparisons: both EGNN generators benefit from equal physical supervision, without a resolved ordering between them.

Terminal noise and fitting duration. The original-model and terminal-noise controls were specified before examining the new GFN2 results. We capture the zero-noise output from the same trajectory before the final 0.025 Å perturbation. Removing this noise adds 1.95 joint-yield points ($[0.39, 3.71]$). The small head exceeds the paired-update model by 8.79 points ($[3.91, 13.67]$) and its zero-noise variant by 6.84 ($[1.95, 11.72]$), although their physical-training sets and parent histories differ.

A later duration diagnostic reuses the saved 2,000-update prefixes of the small-head fits. Targets, initialization, training states, and sampling match the 20,000-update versions. Longer fitting adds 3.71 points ($[1.76, 5.86]$), with both fits improving. This comparison was added after inspecting the confirmation panel.

The small 20,000-update head also lowers all-output GFN2 energy relative to its parent by 0.0951 eV/atom (signed-change 95% interval $[-0.2022, -0.0364]$), with reductions in both runs. The paired averages include invalid outputs and different connectivities.

Graph diagnostics and cost. Requiring zero assigned radical electrons gives joint yields of 46.29% for the small head, 46.09% for the contextual head, and 9.57%/11.52% for GAGA at its two budgets. No graph-valid output duplicates another inferred connectivity within a 16-draw composition/run cell; this measures observed duplication rather than isomer coverage.

The six-system confirmation uses 3,072 trajectories and GFN2 readouts. The parameter-matched control adds 40,000 head updates, following 80,000 updates in the earlier capacity study. Supplemental original-model and short-fit comparisons add 1,024 trajectories, 512 shared-trajectory noise variants, and 1,536 GFN2 readouts. All reuse the physical teacher and evaluate unoptimized coordinates.

P.1 Learning from physically corrected references

We compare fine-tuning on original and physically corrected OMol25 references to isolate the contribution of corrected labels at a fixed additional training budget.

Reference pairs. Fixed metadata hashes select 128 compositions from the existing 3,000-row block, with 32 each in the 8–16, 17–24, 25–32, and 33–40 atom ranges. Each contributes one reference. A bounded Cartesian L-BFGS search under inversion-averaged eSEN moves all coordinates while preserving geometric validity and the inferred labeled graph. The budget is 96 steps and 160 energy/force evaluations, with a 0.05 Å per-atom step cap and stopping threshold of 0.1 eV/Å for the maximum atomic force.

We retain the last accepted point when a search stalls or exhausts its budget, keeping all 128 compositions in both sets. Of these searches, 110 meet the force threshold and 124 reduce force RMS.

Preparation uses 6,576 eSEN evaluations including inversion pairs. Independent GFN2 calculations score all 256 original and corrected references.

Table 26: The 128 training references before and after physical correction. GFN2 independently evaluates the original and corrected labels.

Metric	Original references	Corrected references
Median eSEN force RMS (eV/Å)	2.005	0.050
Median GFN2 force RMS (eV/Å)	1.756	0.397
GFN2 force RMS ≤ 1 eV/Å (%)	22.66	89.84

Adaptation and evaluation. Two runs start from the same 5.90-million-parameter harmonic checkpoint. Within each run, original- and corrected-reference students share reference indices, source draws, rotation and matching randomness, optimizer settings, and 0.005 Å centered endpoint smoothing. Each receives 2,000 batch-size-one updates with uniform composition frequency. Training is direct FM regression to the selected empirical reference distribution, without parameter subtraction.

The confirmation panel contains 16 previously ungenerated compositions, eight per size range, absent from both processed corpora and earlier panels. Each model generates 16 draws per composition and run using 128 backbone calls and zero terminal noise; the original parent is regenerated under the same settings. Checkpoints are fixed before evaluation, and outputs remain unoptimized.

Table 27: Raw generation on the reference-target panel, with 512 attempts per model. Joint yield counts graph-valid outputs with GFN2 force RMS at most 5 eV/Å over all attempts. Force medians use each model's graph-valid outputs with successful calculations.

Model	Graph validity (%)	Median force (eV/Å)	Joint yield (%)
Unadapted parent	38.09	3.083	31.64
Original-reference fine-tuning	30.47	2.762	27.34
Corrected-reference fine-tuning	39.45	2.558	33.01

Corrected-reference training exceeds matched original-reference training by 5.66 joint-yield points (composition 95% interval $[2.54, 8.59]$), with gains of $+8.20$ and $+3.13$. Relative to the unadapted parent, the difference is $+1.37$ points ($[-3.52, 6.05]$), with changes of -1.95 and $+4.69$. Corrected labels therefore improve this fine-tuning procedure, while a net advantage over the parent remains unresolved. Intervals condition on the two fits. Beyond label preparation and 256 GFN2 reference checks, the study adds 8,000 updates, 1,536 raw outputs, and 1,536 GFN2 readouts.

Physical scores of frozen source models. We evaluate every output and reference on the 10-composition additional panel with the same conservative eSEN checkpoint, preserving charge and spin. Inversion averages are $E_+(X) = [E(X) + E(-X)]/2$ and $F_+(X) = [F(X) - F(-X)]/2$. Differences are paired within composition. Subtracting each composition's reference energy per atom leaves the paired model difference unchanged. Evaluation requires 5,160 energy/force queries, including both reference inversions in each run.

We average per-structure force RMS over outputs. Jointly valid comparisons contain 3–35 accepted pairs per composition/run cell; all-attempt comparisons also include invalid structures. Most of the pooled improvement comes from run 1. All-attempt energy changes are -0.08049 and -0.00070 eV/atom in runs 1 and 2, with force-RMS changes -1.21996 and $+0.29302$ eV/Å.

Original endpoint checkpoint. The original 30,000-step OMol25 checkpoint generates coordinates from fixed composition and neutral/no-bond inputs using a bootstrap step followed by 127 Euler updates. At 128 calls, retaining predicted categorical history gives 106/640 and 113/640 graph-valid outputs in two sampling streams. Fixing history categories and centering the bootstrap gives 116/640 and 120/640. Both streams share the checkpoint, and the second configuration changes both history and centering. LatentSpring source models additionally receive shared task adaptation and 3,000 specialized updates; their matched Gaussian and covariance-Gaussian variants isolate source choice.

Table 28: Harmonic-minus-Gaussian eSEN changes on the 10-composition panel. Negative values favor harmonic. Rows distinguish all attempts from jointly graph-valid pairs; brackets give paired 95% intervals.

Metric	Outputs	Difference	Interval
Energy (eV/atom)	All attempts	-0.0406	$[-0.0579, -0.0232]$
	Both outputs valid	-0.0340	$[-0.0692, -0.0003]$
Force RMS (eV/Å)	All attempts	-0.4635	$[-0.6782, -0.2512]$
	Both outputs valid	-0.5389	$[-0.8950, -0.2089]$

Q Normalization of the harmonic mixture

For a tree T, let $B_T \in \mathbb{R}^{(N-1)\times N}$ be its oriented edge-incidence matrix, so that $B_T X$ contains the edge displacements. Let $U \in \mathbb{R}^{N\times(N-1)}$ have orthonormal columns spanning the centered subspace in one spatial dimension. We compute the Jacobian from independent edge vectors to centered positions and then average the resulting density over trees.

Since $B_T\mathbf{1} = 0$, $(B_T U)(B_T U)^\top = B_T B_T^\top$. Cauchy–Binet expresses $\det(B_T B_T^\top)$ as the sum of squares of the N maximal minors. Each minor of a tree incidence matrix is ± 1, giving determinant N and absolute Jacobian $\sqrt{N}$ per spatial dimension. The three-dimensional density is therefore $q(X \mid T, c) = N^{3/2} \prod_{e\in T} h_e(B_T X)$, with each factor evaluated at its edge displacement. It integrates to one. Averaging with the tree law in Equation (1) gives the normalized mixture in Equation (3).

R Source invariance under symmetry coupling

The alignment used to pair source and reference structures depends on both endpoints. A shared random symmetry restores the source marginal after this alignment.

Let G be the compact group of proper rotations and permutations of identical conditioning labels, and let $X \sim q_0$ with q_0 invariant under G. Alignment selects $g(X, Y) \in G$ using the reference Y. Draw an independent Haar-uniform $H \in G$. Conditional on the orbit of X and on g, Haar invariance gives HgX the same uniform orbit distribution as HX. Alignment preserves the orbit, and averaging over source orbits recovers q_0. The paired endpoints may remain correlated; the result establishes the marginal required for the conditional-displacement training objective.

S Covariance controls

The shell-tree control samples an edge length $R = \ell \exp(sZ - s^2/2)$, with $Z \sim N(0, 1)$ and $s = 0.2$, and an independent uniform direction. Since $\mathbb{E}R^2 = \ell^2 \exp(s^2)$ and the directional second moment is $I/3$, the edge covariance is $vI = \ell^2 \exp(s^2) I/3$.

Let A_T map edge vectors to centered atom coordinates. Edge independence gives $C_T = A_T \operatorname{diag}(v_e) A_T^\top$ for the atom-position covariance along one axis. Gaussian edge vectors with covariance $v_e I$ produce the same C_T. All component means are zero, so the shell-tree and harmonic mixtures also share covariance $\mathbb{E}_T C_T$.

The single-Gaussian control approximates this mixture covariance by averaging a finite number of tree covariances, seeded by composition, and then averaging over permutations within each element. This construction preserves positive definiteness on the centered subspace and invariance to identical-atom relabeling. The Gaussian is normalized exactly for the resulting estimated covariance.

T Jarzynski identity for escorted transport

The work correction follows from a change-of-variables identity. Jarzynski's relation links exponential work averages to free-energy differences (Jarzynski, 1997; Hummer & Szabo, 2001); escorted

transport includes an imposed coordinate map and its volume change (Vaikuntanathan & Jarzynski, 2008).

Fix a composition and use intrinsic volume measure on its centered coordinate space. Let x denote independent coordinates and $\beta = 1/(k_B T_{\mathrm{phys}})$. Consider $\rho_1(x) = \mathbf{1}_\Omega(x) \exp[-\beta E(x)]$ with finite, nonzero $Z_1 = \int_{\mathcal{H}_N} \rho_1(x)\, dx$. The region Ω specifies the allowed configurations and can provide the confinement needed for a finite normalizer. The target density is ρ_1/Z_1.

Let $q_0 > 0$ be the normalized source in Equation (3), and let $\Phi : \mathcal{H}_N \to \mathcal{H}_N$ be differentiable and invertible with nonzero intrinsic Jacobian. For $\Phi(x_0) \in \Omega$, define

$$W_\Phi(x_0) = E(\Phi(x_0)) + \beta^{-1} \log q_0(x_0) - \beta^{-1} \log |\det D\Phi(x_0)|, \tag{13}$$

Set $W_\Phi = +\infty$ outside the allowed region. The source-density and Jacobian terms account for the proposal distribution and local volume change. Changing variables gives

$$\mathbb{E}_{q_0}[e^{-\beta W_\Phi}] = \int_{\mathcal{H}_N} \rho_1(\Phi(x_0)) |\det D\Phi(x_0)|\, dx_0 \tag{14}$$

$$= Z_1. \tag{15}$$

The auxiliary source energy $E_0 = -\beta^{-1} \log q_0$ has $Z_0 = 1$, yielding the Jarzynski-type normalizer identity. For an observable f, self-normalized importance sampling estimates its target mean by $\sum_j w_j f(\Phi(x_j))/\sum_j w_j$, with $w_j = e^{-\beta W_\Phi(x_j)}$. Its finite-sample variance depends on proposal coverage of the target.

For a sufficiently regular exact ODE flow, the intrinsic log determinant is the integrated velocity divergence. The implemented local-work studies instead use Gaussian translations (Appendix U) and periodic rotor maps with analytic Jacobians (Appendix V). They evaluate the densities of these explicit local maps, not an exact likelihood of the discretized neural generator.

U LOCAL FORCE-ESCORTED WORK DISTILLATION

The physical teacher constructs nearby structures and assigns their sampling weights as FM targets. We compare uniform sampling of force-shifted candidates with energy-weighted training (Dern et al., 2025; Midgley et al., 2023). A normalized Gaussian around each anchor makes the complete escorted-work weights tractable (Vaikuntanathan & Jarzynski, 2008).

Mirror-symmetric energy and force. Given $E(X)$ and $F(X) = -\nabla_X E(X)$, we define $E_+(X) = [E(X) + E(-X)]/2$ and $F_+(X) = [F(X) - F(-X)]/2$. This construction gives $F_+ = -\nabla E_+$, equal energies for inverted structures, and opposite forces. Subtracting the mean atomic force keeps subsequent displacements in the centered subspace.

Local reference and target. For a graph-valid anchor A with inferred labeled graph G_A, define $q_A(X) = \mathcal{N}_{\mathcal{H}_N}(A, \sigma_A^2 I)$. At 300 K, $kT = k_B T_{\mathrm{phys}} = 0.0258519998$ eV. Set

$$\sigma_A = \min\left(0.03\,\text{Å}, \sqrt{\frac{0.10\,\text{Å}\, kT}{\|F_+(A)\|}}\right), \qquad \delta_A = \frac{\sigma_A^2}{kT} F_+(A). \tag{16}$$

The Frobenius force norm determines a width at most 0.03 Å and displacement norm at most 0.10 Å. At zero force, the maximum width is used. These quantities are fixed before drawing $X \sim q_A$. Translation to $Y = X + \delta_A$ has determinant one and preserves centering.

The unnormalized target is

$$\rho_A(Y) = q_A(Y) \exp\left[-\frac{E_+(Y) - E_+(A)}{kT}\right] \mathbf{1}_{G_A}(Y). \tag{17}$$

The Gaussian restrains the neighborhood around A, the energy factor favors lower-energy configurations, and the indicator preserves inferred labeled bond orders, charges, and radicals while enforcing connectivity and nonoverlap. The target is therefore a restrained, same-graph local distribution.

Force-shift derivation. Linearizing energy gives $E_+(Y) - E_+(A) \simeq -\langle F_+(A), Y - A\rangle_F$. Completing the square in the Gaussian times this Boltzmann factor yields

$$q_A(Y) \exp\left[\frac{\langle F_+(A), Y - A\rangle_F}{kT}\right] \propto \mathcal{N}_{\mathcal{H}_N}\left(Y; A + \frac{\sigma_A^2}{kT} F_+(A), \sigma_A^2 I\right). \tag{18}$$

Thus force translation exactly samples the linearly reweighted Gaussian before graph restriction. Equivalently, the Gaussian has spring stiffness kT/σ_A^2 and its mean shifts by force divided by stiffness. Applying the same-graph screen truncates this distribution.

Complete work. To correct the linear approximation, we evaluate the nonlinear energy. With $H_0 = -kT \log q_A$ and $H_1 = H_0 + E_+ - E_+(A)$ on the allowed support, the work is

$$W_A(X) = E_+(Y) - E_+(A) + kT[\log q_A(X) - \log q_A(Y)] \tag{19}$$

and is $+\infty$ outside it. The density ratio accounts for translation from X to Y, giving $\mathbb{E}_{q_A}[\exp(-W_A/kT)] = \int \rho_A(Y)\, dY$. Analytic tests recover the constant-force identity and the quadratic-energy normalizer and weighted mean. Under Equation (18), every allowed candidate has $W_A = -\sigma_A^2 \|F_+(A)\|_F^2/(2kT)$, so uniform weights recover the linear approximation.

Training pools. Each run selects eight 17–28-atom training compositions by metadata hash and generates 32 anchors per composition. Evaluation compositions are excluded. Invalid anchors remain in generation counts but not teacher pools. Each valid anchor produces eight shifted candidates. Complete-work training uses self-normalized $\exp(-W_A/kT)$ weights, with zero weight outside the support; the unweighted teacher samples uniformly from the same admissible candidates. Replay uses original retained anchors. Anchors without an admissible candidate are recorded and excluded consistently from all students.

Each student receives 500 reference FM updates alternating with 500 generated-target updates, sharing initialization, anchor draws, optimization, and sampling budgets. Replay needs no physical labels; uniform escort requires anchor forces; complete work also requires candidate energies. Actual shared-experiment queries and standalone method requirements are reported separately. Sampling uses the final network, with a fixed temperature feature; 300 K determines the local teacher weights.

Self-normalized weighting has finite-particle bias. We report ESS as $(\sum_j w_j)^2 / \sum_j w_j^2$, ranging from one for a dominant candidate to eight for equal weights.

Direct fine-tuning. Each run trains three students for 1,000 updates. Of 512 anchors, 351 pass graph checks and 349 admit at least one shifted candidate. Composition-level mean ESS is 1.27–1.87 of eight, indicating concentrated weights and limited local-target resolution.

On jointly valid pairs, complete-work training lowers energy relative to uniform escort by 0.00758 eV/atom (signed-change 95% interval $[-0.01424, -0.00113]$) and force RMS by 0.16174 eV/Å ($[-0.27660, -0.05590]$). Both runs and the all-output comparison favor complete work.

Nevertheless, all three direct students underperform the frozen parent. Complete work gives 55.16% graph validity versus 62.11%, and raises jointly valid paired energy by 0.10587 eV/atom ($[0.09009, 0.12168]$). Replay also degrades performance, motivating subtraction of its matched update.

Teacher preparation uses 6,232 queries: 702 at anchors and 5,530 at allowed candidates. Uniform escort can omit candidate energies; replay requires no physical queries. Evaluation of the four models adds 10,240 queries, separate from sampling.

Paired physical update. For both uniform-escort and complete-work students, we subtract the replay update through parameter arithmetic (Ilharco et al., 2023):

$$\theta_{\text{final}} = \theta_{\text{base}} + (\theta_{\text{physical}} - \theta_{\text{replay}}), \tag{20}$$

The coefficient is one, and buffers and the source distribution remain fixed.

Both paired variants lower energy and force RMS in both runs. Complete work reduces jointly valid force RMS by 0.38403 eV/Å ($[-0.48124, -0.28932]$), also improving all-output energy and force. Its graph-validity change is -0.55 points ($[-2.58, 1.56]$).

Table 29: Graph-valid outputs after direct fine-tuning, out of 640 attempts per run. All students share the additional training budget; the frozen parent receives no updates.

Run	Frozen	Replay	Unweighted escort	Complete work
1	418	340	330	344
2	377	345	371	362

Table 30: Paired-update results following direct fine-tuning. Energy changes are relative to the frozen model on jointly valid pairs, with paired 95% intervals. Counts report the two generation streams.

Model	Graph-valid outputs (two streams)	Energy change (eV/atom)
Frozen	418/640; 377/640	0
Paired force update	407/640; 387/640	-0.01554 $[-0.02084, -0.01031]$
Paired work update	406/640; 382/640	-0.01604 $[-0.02188, -0.01013]$

Differences between the two paired teachers are small: complete work minus uniform escort gives -0.00243 eV/atom on jointly valid pairs ($[-0.00756, 0.00264]$) and $+0.00022$ over all attempts ($[-0.00557, 0.00649]$). Force and validity intervals also include zero. Uniform escort requires 702 anchor queries versus 6,232 for complete work; both sample with 128 neural calls. We therefore select force-only targets for the new-composition test in Section M.2.

Evaluation adds 2,560 outputs and 5,120 physical queries. Parameter combination requires no optimizer updates.

V Molecular rotor work protocol

We test the work correction on three controlled methyl rotations: `C=C(N)N(C)C(N)=O`, `CCOC(C)(C)N`, and `CC(=O)NCC=O`. Among the first five rows of an existing training subset, we select the first three distinct compositions whose rotor preserves the inferred labeled graph on a 64-angle grid. Selection uses geometry alone. Only methyl hydrogens rotate around the carbon–anchor axis; other internal coordinates remain fixed. We remove translation and define densities with respect to $d\phi$ on $[-\pi, \pi)$.

Reference and map. The circular reference is $q_0(\phi) = \exp(\cos\phi)/(2\pi I_0(1))$, with $2\pi I_0(1) = \int_{-\pi}^{\pi} \exp(\cos u)\, du$. Its Hamiltonian is $H_0 = -kT \log q_0$, and the target is defined by $H_1 = H_0 + U - U(0)$. For $a \in \{-0.5, 0, 0.5\}$, the map $\psi_a(\phi) = \phi + a \sin\phi$ has Jacobian $J_a(\phi) = 1 + a\cos\phi$. Because $|a| < 1$, it is smooth and invertible on the circle. Complete work is

$$W_a(\phi) = U(\psi_a(\phi)) - U(0) + kT\left[\log q_0(\phi) - \log q_0(\psi_a(\phi)) - \log J_a(\phi)\right]. \tag{21}$$

The reference-density and Jacobian terms correct changes in angular sampling and interval volume. Changing variables gives

$$\mathbb{E}_{q_0} e^{-W_a/kT} = \int_{-\pi}^{\pi} q_0(y) e^{-[U(y)-U(0)]/kT}\, dy = Z_1/Z_0, \qquad Z_0 = 1. \tag{22}$$

Numerical reference. All 1,536 GFN2 single-point calculations succeed. Each rotor energy curve uses a periodic cubic interpolation of 512 values. At the added midpoint samples, the 256-knot interpolation differs by at most 10^{-8} eV. Doubling quadrature resolution from 32,768 to 65,536 changes free energy by less than 10^{-10} eV.

Estimators. Total variation uses 64 equal angular bins with quadrature-based target probabilities. The mean unnormalized weight is unbiased for the normalizer, whereas self-normalized observable estimates have finite-sample bias. Free-energy RMSE for $-kT \log \widehat{Z}$ includes its finite-sample bias. At 2,048 draws, complete-work ESS ranges from 788 to 1,892 across six nonidentity cases. Incomplete weights can have higher ESS while targeting an incorrect distribution, so distribution error is evaluated separately.

V.1 Complete-work correction on molecular rotors

For each of the three rotors above, a 512-point GFN2 interpolation defines $U(\phi)$. At 300 K, the target is $\rho(\phi) \propto q_0(\phi) \exp[-(U(\phi) - U(0))/kT]$, where $kT = k_B T_{\text{phys}}$ and q_0 is the concentration-one von Mises reference. The target retains this angular restraint while reweighting by energy.

For each of the three maps, we compare complete-work, energy-only, no-Jacobian, and unweighted estimates. Complete work accounts for both reference-density change and interval volume. Each setting uses 128 repetitions at sample counts from 8 to 2,048 (Appendix V).

At 2,048 samples, complete work gives lower mean total-variation error than either incomplete weighting in all six nonidentity cases. Mean errors are 0.0636, 0.1532, and 0.1134 for complete work, energy-only, and no-Jacobian weighting, respectively: a 58.5% reduction relative to energy-only weights. Free-energy RMSE ranges from 0.158 to 0.734 meV. Independent density-ratio quadrature recovers the normalizer in all nine molecule/map cases. For the identity map, all three weighted methods coincide. These tests establish recovery of the specified restrained rotor distributions.

W Physical quality of the matched generators

We score the 16,384 saved 128-call outputs from Section J.3, without retraining or changing coordinates. The evaluation protocol is fixed before scoring. eSEN evaluates every output and its reflection, requiring 33,024 calls. GFN2 evaluates graph-valid outputs and both reference inversions; all 2,198 calculations succeed, including 256 reference checks.

Joint yield counts outputs passing graph validity and a force threshold over all attempts. Figure 13 varies the threshold to show sensitivity to this choice. A force cutoff measures residual-force quality, not stationarity. These models share the architecture and training budget in Appendix L and receive no physical updates.

Table 31: GFN2 scores of matched generators. Energy and force are medians over each method's graph-valid outputs; yield uses all 4,096 attempts. Energy excess is relative to one same-composition OMol25 reference, whose connectivity may differ from the generated graph.

Method	Energy excess (eV/atom)	Force RMS (eV/Å)	Yield at 5 eV/Å (%)
Gaussian FM	0.6820	5.2911	4.27
Harmonic FM	0.7114	5.8183	4.03
EDM	0.6799	5.8768	4.69
GAGA	0.6604	5.2785	6.15

GAGA has the stronger pooled physical scores in this matched comparison. The harmonic source improves graph validity without improving physical scores among valid outputs. The physical-adaptation studies separately test whether local force targets address this gap.

X Training-component controls

X.1 Self-conditioning

We remove self-conditioning while retaining the harmonic model's initialization, 3,000 references, source, and pairing procedure. One single-pass model receives 3,000 updates; another receives 6,000 to match the self-conditioned training-forward count. Both use 64 midpoint steps, so all variants sample with 128 network calls.

Self-conditioning adds 7.23 graph-validity points at matched updates (95% interval [4.69,9.77]) and 12.43 at matched forward counts ([9.70,15.23]), with both runs improving. Additional single-pass training therefore does not recover the gain in these runs. Forward counts are matched rather than runtime. Section M.1 gives the update-direction control.

Force displacement. We remove the force shift while retaining Gaussian perturbations, candidate indices, force-dependent widths, selection, training schedule, and paired parameter update. Relative to

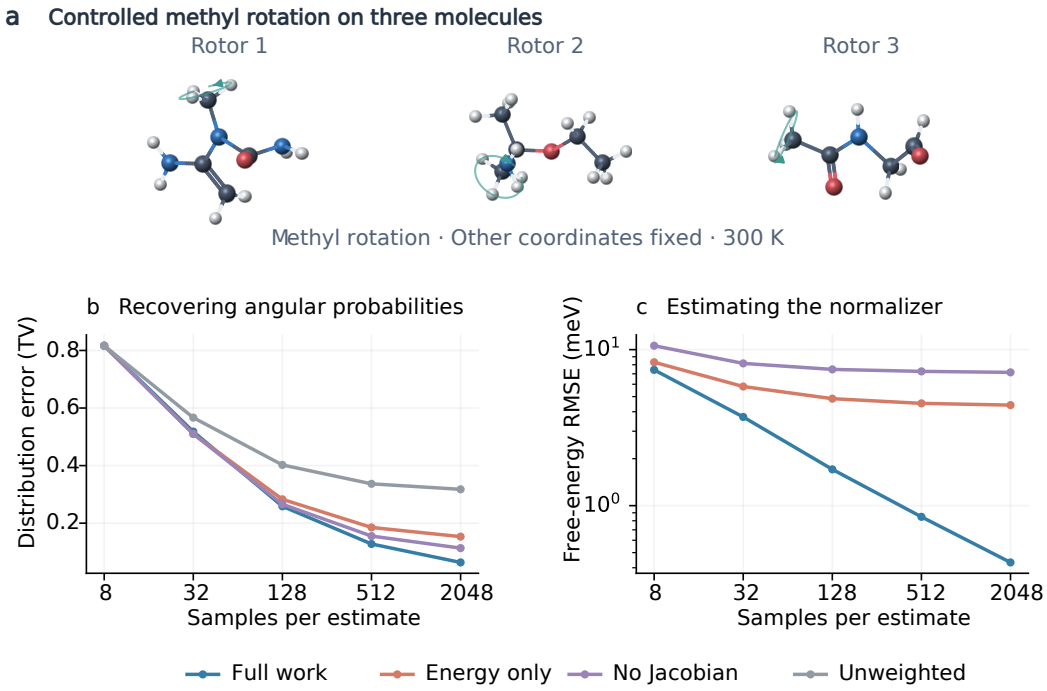


Figure 12: Work correction for restrained molecular rotors at 300 K. (a) Reference structures and methyl rotations. (b,c) Angular-distribution and normalizer errors averaged over three molecules and two nonidentity maps, with 128 repetitions per case. Unweighted samples have no free-energy estimator; all targets retain their angular reference restraint.

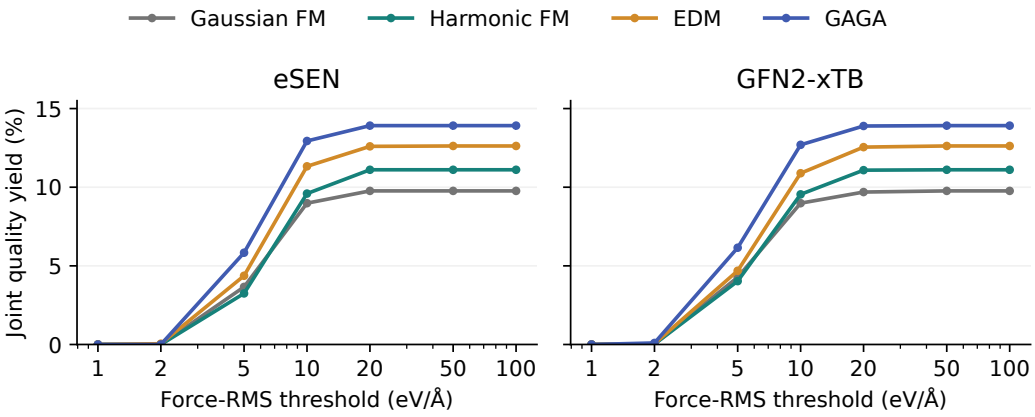


Figure 13: Joint yield versus force threshold, pooling two initializations and 64 compositions. All attempts contribute to the denominator. GAGA has the highest pooled yield at most nontrivial thresholds under both potentials.

this control, shifted targets lower jointly valid paired energy by 0.01939 eV/atom (signed-change 95% interval [-0.02417,-0.01455]) and force RMS by 0.3926 eV/Å ([-0.4696,-0.3263]). Both runs improve. Graph validity changes by $+0.39$ points ([-1.33,2.19]). The comparison isolates the displacement, with other force-dependent target-construction choices retained.

Table 32: Self-conditioning on 24 compositions, with 768 attempts per run. The valid-output column reports runs 1 and 2; the self-conditioned model exceeds both single-pass training budgets.

Model	Updates	Training passes	Valid outputs	Rate (%)
Single pass	3,000	3,000	363;336	45.51
Single pass	6,000	6,000	277;342	40.30
Self-conditioned	3,000	6,000	435;375	52.73

X.2 Matched experiment on sources and physical adaptation

We complete the four-setting comparison by constructing force targets from Gaussian-source outputs. Each run generates 32 anchors on each of the same eight training compositions, records all 256 attempts, and discards invalid anchors without replacement. The eight-candidate escort, temperature, perturbation bounds, source seeds, and 1,000-update schedule match the harmonic procedure. Physical and replay students share the Gaussian base, references, source randomness, times, and teacher draws; their parameter difference is added with coefficient one.

After support checks, the Gaussian runs retain 160 and 138 anchors. Preparation uses 5,276 eSEN queries, including complete-work diagnostics; the students sample uniformly from admissible shifted candidates. Maximum preparation budgets match, while realized anchor and query counts vary by source. Four new 1,000-update student fits yield two adapted Gaussian models. Harmonic and unadapted results reuse the existing checkpoints and outputs.

The new models generate 1,536 raw outputs on the existing 24-composition panel at 128 calls each. Evaluation adds 3,072 eSEN queries and 1,632 GFN2 attempts including reference inversions. One graph-invalid calculation fails; all graph-valid and reference calculations succeed. Checks reconstruct proposals and graph restrictions, replay target draws and parameter arithmetic, and reparse scores. The comparison is a follow-up on the previously evaluated panel.

For Gaussian, adaptation changes jointly valid paired energy by -0.01596 eV/atom under eSEN (95% interval [-0.02182,-0.01021]) and -0.01465 under GFN2 ([-0.02041,-0.00929]). Force-RMS changes are -0.3795 eV/Å ([-0.4618,-0.3045]) and -0.4618 ([-0.6195,-0.3370]).

At 5 eV/Å, GFN2 joint-yield gains are 4.56 and 5.34 points for Gaussian and 3.13 and 2.99 for harmonic. After adaptation, harmonic minus Gaussian is 16.28 in run 1 and -2.34 in run 2. These results show within-source adaptation benefits and a positive pooled source difference, with substantial run variation; a positive source-by-update interaction is not established.

Y Curvature-aware local work sampling

Force translation changes the proposal mean but leaves its shape fixed. To account for stiff and soft molecular directions, we fit local curvature and adjust the proposal covariance. The target distribution and maximum candidate-query budget remain unchanged.

Curvature fit. Let $Q_N \in \mathbb{R}^{3N\times d}$ have orthonormal columns spanning centered coordinate space, with $d = 3(N-1)$. A displacement $u \in \mathbb{R}^d$ in angstroms corresponds to $Q_N u$ in atomic coordinates, and f_A is the anchor force in this basis. Eight independent pilot displacements form the columns of D; their negative force changes form G. We fit a symmetric B_A to $G \approx B_A D$:

$$\|BD - G\|_F^2 + \lambda\|B - \kappa I\|_F^2, \qquad \kappa = \max\left(0, \frac{\langle D, G\rangle_F}{\|D\|_F^2}\right), \qquad \lambda = 0.01\,\|D\|_F^2/K. \tag{23}$$

Here K counts admissible pilot columns. With none, or with exactly zero displacements, we set $B_A = 0$ and recover force translation. Otherwise, $\|D\|_F^2$ is floored at 10^{-20} in scalar and ridge calculations. The nonnegative stiffness κ provides an isotropic shrinkage target. We clip negative fitted eigenvalues to zero; an isotropic control uses $B_A = \kappa I$. The map is fixed before drawing production candidates.

Proposal transformation. The fitted energy is $E_+(A + Q_N u) - E_+(A) \approx -f_A^\top u + u^\top B_A u/2$. Reweighting the original Gaussian by this quadratic Boltzmann factor changes both its mean and

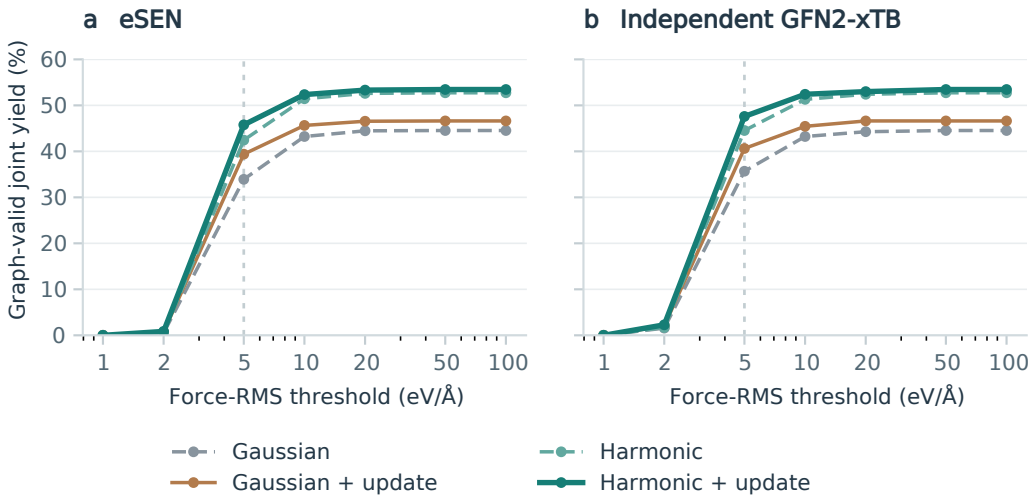


Figure 14: Joint yield versus force-RMS threshold for both sources, with and without paired physical updates. Curves use all raw attempts and the same threshold grid for both potentials.

covariance. With dimensionless precision $P_A = I + \sigma_A^2 B_A/(k_B T_{\mathrm{phys}})$, define

$$v = P_A^{-1/2} u + b_A, \qquad b_A = P_A^{-1} \frac{\sigma_A^2 f_A}{k_B T_{\mathrm{phys}}}, \qquad \log J_A = -\tfrac{1}{2} \log \det P_A. \tag{24}$$

The factor $P_A^{-1/2}$ contracts stiff directions, b_A adjusts the mean shift, and J_A is the intrinsic volume factor. Reference width and temperature stay fixed. For $X = A + Q_N u$ and $Y = A + Q_N v$, the work is

$$W_A = E_+(Y) - E_+(A) + k_B T_{\mathrm{phys}}[\log q_A(X) - \log q_A(Y) - \log J_A]. \tag{25}$$

Candidates outside the original same-graph support receive zero weight. This applies targeted and escorted reweighting to the local proposal (Jarzynski, 2002; Vaikuntanathan & Jarzynski, 2008).

Weight variation. Writing the true energy as the fitted quadratic plus remainder $R_A(v)$ gives $W_A = -k_B T_{\mathrm{phys}} \log Z_{\mathrm{quad}} + R_A(v)$, where Z_{quad} integrates the reference Gaussian times the quadratic Boltzmann factor. Only the remainder varies across allowed candidates. An exact quadratic therefore gives equal weights. Tests verify this identity for anisotropic quadratic energy and check rotation equivariance of the fitted map.

Equal-query comparison. In each run, we select the first four eligible anchors from each of eight training compositions, without energy or ESS ranking. Translation uses 24 fresh production particles; curvature methods use at most eight pilot queries and 16 fresh production particles. Common anchor queries and reflection evaluations are counted separately. All production energies are evaluated, including graph-invalid candidates, and pilot samples are not reused in production.

Table 33: Complete-work ESS at a maximum of 24 particle queries per anchor, with 32 anchors per run. Curvature methods use eight pilot and 16 production queries; translation uses 24 production queries.

Escort	Production particles	ESS, run 1	ESS, run 2
Force translation	24	2.22	1.86
Isotropic curvature	16	3.55	3.94
Directional curvature	16	6.69	6.70

Directional curvature increases ESS on 61 of 64 anchors, with mean gain 4.66 effective samples and composition 95% interval $[3.81, 5.46]$. Same-graph retention is 99.51%, versus 99.54% for

translation. Independent Gaussian density calculations reproduce the weights. Fresh full pools for all 349 eligible anchors give translation/curvature ESS of 2.32/6.71 in run 1 and 2.33/6.37 in run 2.

Effect on generation. We train students using uniform sampling of 24 translated candidates, complete-work weighting of those same 24, or complete-work weighting of 16 curvature-adjusted candidates. All use the original 1,000-update schedule, reference/anchor selection, replay model, and coefficient-one parameter subtraction. All 349 anchors remain eligible. Work-weighted methods have equal maximum query budgets; uniform targets can omit candidate-energy queries. These larger pools form a follow-up, while the primary model retains its preselected eight-candidate force teacher.

Table 34: Generation after training with three local teachers: 1,536 attempts per method on 24 compositions. Energy changes compare weighted and uniform-force students on jointly valid pairs. All methods use 128 neural calls and raw output coordinates.

Teacher	Graph validity (%)	eSEN change (eV/atom)	GFN2 change (eV/atom)
Uniform force, 24 candidates	54.30	0	0
Complete work, 24 candidates	53.39	$+0.00959$	$+0.00845$
Curvature work, 16 candidates	53.32	$+0.00793$	$+0.00783$

Curvature minus translation work changes energy by -0.00184 eV/atom under eSEN (95% interval $[-0.00654, 0.00289]$) and -0.00068 under GFN2 ($[-0.00484, 0.00341]$), with opposite signs across runs. Better teacher ESS does not translate into a resolved generation gain. Uniform force targets give lower energy and force RMS than both weighted alternatives. Relative to the frozen parent, the 24-candidate uniform model lowers energy by 0.02088 eV/atom under eSEN and 0.01918 under GFN2. These results favor retaining the inexpensive force-only recipe.