CURVATURE-AWARE FLOW MATCHING FOR MOLECULAR STRUCTURE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models produce three-dimensional molecular structures with high geometric accuracy, yet they are trained on distributions of atomic coordinates without explicit access to the underlying physics. The potential energy surface (PES) describes how the energy changes around a molecular geometry, and the root mean square deviation (RMSD) between a generated structure and its reference captures this local landscape only indirectly. Supervising the PES during training could close this gap, but conformers and transition states (TSs) are stationary points at which the forces vanish, so first-order quantities carry little information about them. What remains is the local curvature, encoded in the Hessian of the PES. Quantum-chemical Hessians are too expensive to be available at the scale required for training, but recently proposed machine-learned predictors make curvature inexpensive enough to supervise directly. We introduce CurvFM, a flow matching (FM) formulation whose training objective constrains the curvature at the generated structure. A frozen Hessian prediction model is evaluated at the generated endpoint and at the reference structure, which removes the need for quantum-chemical Hessians and keeps the residual cheap enough to impose at every training step. Across TS and conformer generation, CurvFM substantially improves physical fidelity at unchanged global structural accuracy, reducing the median deviation in maximum force by 63% on GEOM-QM9 and raising the fraction of valid TSs by up to 10% across three reaction datasets.

1 INTRODUCTION

Generative models have become the dominant approach for producing three-dimensional molecular structures of small organic molecules, covering conformer ensembles as well as reaction transition states (TSs) (Jing et al., 2022; Hassan et al., 2024; Kim et al., 2024; Duan et al., 2025). They are trained on coordinate-based objectives and evaluated with geometric metrics such as the root mean square deviation (RMSD), which measures how far a generated structure sits from its reference. Agreement in this sense is only loosely related to agreement in energy, even though the geometry determines the energy exactly. A generated molecule can sit close to the reference under RMSD and still contain a steric clash or a distorted bonding environment that raises its energy considerably (Galustian et al., 2025; Darouich et al., 2026b; Zhao et al., 2025). Such errors are costly because barrier heights and Boltzmann-weighted ensemble properties depend exponentially on the energy, so even moderate energetic errors translate into large errors in the predicted property.

The relation between geometry and energy is given by the potential energy surface (PES). What makes a generated structure physically meaningful is where it sits on this surface, not how close it lies to a reference geometry. Supervising a model on both, the geometry and its position on the PES, is therefore the natural way to close this gap. Constraining a generative model against a known physical law during training is an established idea, developed most systematically for partial differential equations (PDEs), where the governing equation is available in closed form and its residual can be evaluated on the generated sample at every step (Bastek et al., 2025; Baldan et al., 2026). For molecules no such closed-form operator exists, and evaluating the PES at quantum-chemical accuracy inside a training loop is prohibitively expensive. Physical information has therefore been introduced through surrogates, in the form of physics-based potentials or learned energy models that steer the

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

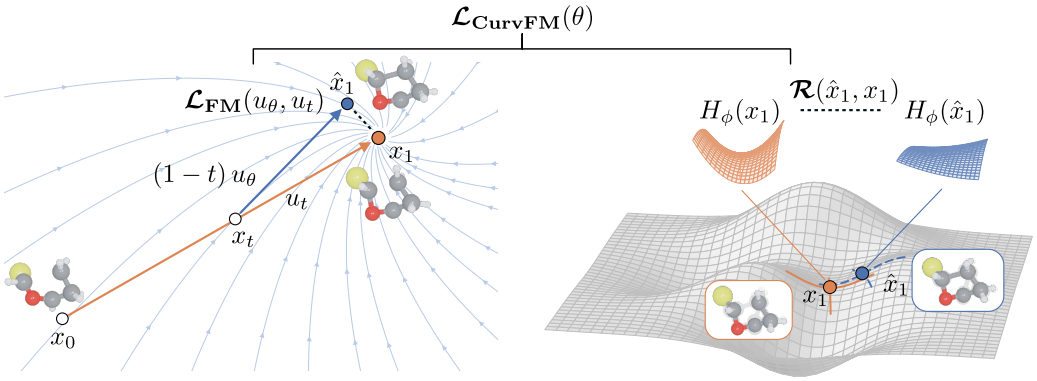

Figure 1: **Conceptual illustration of CurvFM**. **(i)** At an interpolated sample x_t, the model predicts a velocity $u_\theta(x_t, t)$ that is regressed against the conditional velocity field u_t pointing to the target structure x_1. Integrating the velocity field to $t = 1$ gives an estimate $\hat{x}_1$ of the clean endpoint. **(ii)** Both $\hat{x}_1$ and x_1 are passed through a frozen Hessian predictor H_ϕ, and the residual $\mathcal{R}(\hat{x}_1, x_1)$ compares the curvature of the PES at the two geometries. Penalizing their disagreement encourages the generated structure to lie at the right place on the energy landscape, which the FM loss alone does not account for.

trajectory at inference (Wohlwend et al., 2025; Xu et al., 2025), and of force-field gradients or rewards used to finetune a pretrained model (Berlaga et al., 2025; Zhou et al., 2025).

All of these signals are of zeroth or first order in the energy, which limits what they can express. The structures most commonly targeted by molecular generative models, such as conformers and TSs, are stationary points of the PES. A conformer is a local minimum and a TS is a first-order saddle point, so the forces vanish at either and no first-order signal separates them. What distinguishes the two is the local curvature, encoded in the Hessian of the energy with respect to the atomic positions, which is therefore the lowest-order quantity that characterizes these structures at all. Supervising a generative model on curvature has until recently been impractical. Machine-learned interatomic potentials (MLIPs) give access to the PES at a fraction of the quantum-chemical cost. Obtaining a Hessian from one, however, requires differentiating the model twice. Inside a training loop the residual must in addition be differentiated to update the generator, which adds considerable overhead to every step. Models that predict the Hessian directly rather than by differentiation avoid this entirely, returning the full matrix from a single forward pass (Cui et al., 2025; Burger et al., 2025). Curvature thereby becomes affordable as a training signal.

We introduce CurvFM, a **curv**ature-aware **f**low **m**atching (FM) formulation that supervises the curvature of the PES at the structures the model generates. During training, the predicted velocity is used to form an estimate of the clean endpoint, which is passed through a frozen Hessian prediction model and compared against the prediction of the same model at the ground-truth endpoint. Evaluating both sides with the same surrogate cancels its systematic error and removes the need for reference Hessians, so the residual can be applied to datasets for which no such labels exist. Across TS and conformer generation, CurvFM leaves global agreement with the reference geometry unchanged while placing the generated structures closer to the stationary points they are meant to occupy. The median deviation in maximum force decreases by 63% on GEOM-QM9, and the fraction of valid TS rises up to 10% across three reaction datasets. Our main contributions are as follows.

- We introduce CurvFM, a second-order supervision scheme for generative models that constrains the curvature of the PES at the generated sample using a frozen Hessian predictor.
- We show that a surrogate-based residual must be evaluated with the same model on both sides, so that its systematic error cancels and no quantum-chemical labels are required.
- Across four datasets and two architectures, CurvFM improves physical fidelity at unchanged global structural accuracy, and the resulting structures require fewer steps to relax to a true stationary point.

2 Related work

Physics-constrained generative modeling. Physical constraints have been incorporated into generative models most extensively for PDEs, either at inference, by projecting or guiding samples toward the constraint set during integration (Huang et al., 2024; Jacobsen et al., 2025; Utkarsh et al., 2025), or during training, by adding a residual term to the generative objective (Bastek et al., 2025; Baldan et al., 2026). For molecules, where no such residual can be evaluated in closed form, the physical signal comes from a surrogate. Most operate at inference, steering the sampling trajectory with physics-based potentials or with a learned energy model to remove clashes and other unphysical artifacts (Wohlwend et al., 2025; Xu et al., 2025; Mark et al., 2025). Others act after training, finetuning a pretrained model against energy and force rewards derived from a machine-learned force field (Zhou et al., 2025; Li et al., 2026) or steering its vector field with force-field gradients via adjoint matching (Berlaga et al., 2025).

Molecular structure generation. Conformer and TS generation are both instances of generating 3D geometry conditioned on a graph-level specification. For conformers, the condition is a single molecular graph and the target is a set of geometries compatible with it. Distance-geometry and rule-based generators such as RDKit (Riniker & Landrum, 2015) remain widely used, while diffusion and flow matching models have emerged as a data-driven alternative (Luo et al., 2021; Shi et al., 2021; Xu et al., 2022; Jing et al., 2022; Liu et al., 2025; Cao et al., 2025; Xu et al., 2025; Darouich et al., 2026c). For TSs, the condition is a reactant and product pair and the target is the saddle-point geometry separating them. Locating such a point classically means running an iterative or path-based search (Jónsson et al., 1998; Henkelman et al., 2000; Banerjee et al., 1985; Henkelman & Jónsson, 1999), whose cost is dominated by the electronic-structure calls it requires. Replacing these calls with a MLIP reduces the cost by orders of magnitude while retaining the guarantees of an explicit optimization (Yuan et al., 2024; Wander et al., 2025). The search remains iterative, however, and its outcome depends on how well the potential describes the reactive region of the surface, which is where reference data is typically scarcest. Diffusion and flow models instead sidestep the search altogether by generating the saddle-point geometry directly (Duan et al., 2023; Kim et al., 2024; Duan et al., 2025; Galustian et al., 2025; Darouich et al., 2026b;a).

3 Background

Physics-based flow matching. Flow matching learns a time-dependent velocity field u_θ that transports a tractable source distribution p_0 to the data distribution p_1 (Lipman et al., 2023). Training is done in a simulation-free fashion by drawing a noise sample x_0 and a data sample x_1 and interpolating between them. Under the optimal-transport formulation (Tong et al., 2024) the interpolant is the straight line $x_t = (1-t)\,x_0 + t\,x_1$, along which the conditional velocity is constant, $u_t = x_1 - x_0$. The network is then regressed against this target,

$$\mathcal{L}_{\mathrm{FM}}(u_\theta, u_t) = \mathbb{E}_{t,x_t} \left\| u_\theta(x_t, t) - u_t \right\|_2 , \tag{1}$$

and sampling integrates the learned field from $t = 0$ to $t = 1$ with a numerical ODE solver.

Physical constraints can be added to this objective through a residual $\mathcal{R}$ that penalizes violations of a known physical law. Such a residual cannot be evaluated on x_t, which interpolates between noise and data and carries no physical meaning, so it is imposed instead on an estimate $\hat{x}_1$ of the clean endpoint obtained from the predicted velocity, giving $\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \|\mathcal{R}(\hat{x}_1)\|_2$ (Bastek et al., 2025; Baldan et al., 2026). This substitution introduces Jensen's gap (Zhang et al., 2025), since for a nonlinear residual $\mathbb{E}[\mathcal{R}(x_1) \mid x_t] \neq \mathcal{R}(\mathbb{E}[x_1 \mid x_t])$. The quantity minimized at intermediate noise levels is therefore not necessarily the residual of interest. PBFM (Baldan et al., 2026) reduces this discrepancy by unrolling the ODE over n steps of size $(1-t)/n$ before evaluating $\mathcal{R}$, and by weighting the residual loss with t^p to down-weight the imprecise endpoint estimates obtained near $t = 0$.

The potential energy surface and its curvature. The PES relates the energy of a molecular system to its geometric configuration and underlies the theoretical description of chemical reactivity (Lewars, 2024). Quantum-chemical methods such as density functional theory (DFT) provide approximate PESs, but exploring one exhaustively is infeasible for all but the smallest systems, since a nonlinear molecule with N atoms spans $3N-6$ internal degrees of freedom. In practice, one is rarely interested

in the surface as a whole, but in the few points that carry chemical meaning and in the local behavior of the surface around them (Dewyer et al., 2018). Conformers and TSs are two such points, both stationary and therefore characterized by the local curvature of the surface. That curvature is encoded in the Hessian $H \in \mathbb{R}^{3N\times 3N}$, the second derivative of the energy with respect to the atomic positions. Its eigenvectors are the normal modes of the system and the corresponding eigenvalues give the curvature along them. Six eigenvalues vanish, corresponding to the global translations and rotations that leave the energy unchanged. A conformer is a local minimum, so the remaining $3N - 6$ eigenvalues are positive, whereas a TS is a first-order saddle point, so exactly one is negative and its eigenvector spans the reaction coordinate. Obtaining Hessians at the quantum-chemical level is expensive, which motivates the use of surrogate models. MLIPs replace the quantum-chemical PES with a surrogate fitted to energy and force labels, and are differentiable, so Hessians can in principle be obtained by automatic differentiation and used in second-order TS optimization schemes (Yuan et al., 2024). Curvature is, however, not recovered reliably from first-order labels alone (Zhao et al., 2025; Cărare et al., 2025; Bheemaguli et al., 2026), which has motivated Hessian supervision during MLIP training (Rodriguez et al., 2025; Cui et al., 2025; Rodriguez et al., 2026; Yin et al., 2026). A more direct route bypasses differentiation altogether and predicts the Hessian as a model output (Wu et al., 2026), as in Hessian Interatomic Potentials (HIP) (Burger et al., 2025), which read the full matrix out of an SE(3)-equivariant backbone.

4 Curvature-Aware Flow Matching

Purely geometric supervision leaves an important aspect of molecular structure unconstrained. We characterize this gap below and then introduce CurvFM, which closes it by supervising the model on curvature information.

4.1 The limits of geometric supervision

Coordinate-based objectives and the RMSD-type metrics used to evaluate them treat all displacement directions alike, but the PES does not. A displacement along a soft torsion changes the energy negligibly, whereas one of the same magnitude along a bond stretch changes it by orders of magnitude more. Averaging squared displacements over all $3N$ coordinates therefore obscures exactly the directions in which the energy varies most, and does so increasingly as the molecule grows. The consequences are documented across both tasks. TS energies are highly sensitive to small structural perturbations (Duan et al., 2025), and generated conformers must generally be relaxed before ensemble properties can be reliably computed (Jing et al., 2022). A structure can therefore lie close to the reference and still be far from the stationary point it is meant to represent (Zhao et al., 2025; Galustian et al., 2025; Darouich et al., 2026b;a; De Landsheere et al., 2026).

4.2 Curvature as a training signal

Closing the gap between geometric agreement and physical validity requires a training signal that refers to the PES itself, rather than to the reference geometry alone. Evaluating the local shape of the surface at the structures the model generates makes their position on the PES accessible during training. Since conformers and TSs are both stationary points, at which the forces vanish, this information must come from the curvature, and the Hessian is the lowest-order object in which a minimum and a first-order saddle point remain distinguishable.

What limits the use of this signal is the cost of obtaining it during training. Quantum-chemical Hessians from DFT are far too expensive to evaluate for every sample in every batch. MLIPs provide a cheaper surrogate, but obtaining a Hessian from one requires differentiating the model twice with respect to the atomic positions. This costs $3N$ backward passes per structure and produces an activation footprint that grows with the square of the system size. In a generative training loop the burden compounds further, since the residual must itself be differentiated with respect to the generator parameters, which places the Hessian evaluation on the backward path of every update and exposes third derivatives of the surrogate. Models that predict the Hessian directly avoid both difficulties, obtaining the full matrix in a single forward pass (Burger et al., 2025). Curvature then costs no more than an energy evaluation, and the backward path through the frozen surrogate remains first order, so the residual can be imposed at every training step at marginal cost in memory and runtime.

4.3 RESIDUAL FORMULATION

Curvature enters the residual $\mathcal{R}$ through a direct Hessian predictor $H_\phi : \mathbb{R}^{3N\times 3} \to \mathbb{R}^{3N\times 3N}$, which remains frozen throughout training and is queried at the structures the generative model produces. The residual compares the curvature at the generated endpoint $\hat{x}_1$ against the curvature at the corresponding reference structure x_1,

$$\mathcal{R}(\hat{x}_1, x_1) = \mathcal{D}\big(H_\phi(\hat{x}_1),\, H_\phi(x_1)\big), \tag{2}$$

where $\hat{x}_1$ is an estimate of the clean endpoint obtained during training and $\mathcal{D}$ is a discrepancy between two curvature matrices. The choice of reference curvature, either the surrogate prediction or a quantum-chemical calculation at x_1, has a substantial effect on the resulting training signal. Using the surrogate on both sides means the two predictions carry largely the same systematic error, which cancels in the comparison, as we verify in Appendix B.2. The residual then reflects the quality of $\hat{x}_1$ rather than the inaccuracies of the surrogate itself. Avoiding quantum-chemical reference Hessians also makes the residual applicable to datasets for which no such labels exist. The reference term depends solely on the data, so in practice $H_\phi(x_1)$ is precomputed once for the training set and a training step requires only a single surrogate evaluation.

The endpoint $\hat{x}_1$ at which the residual is evaluated is obtained from the velocity predicted by the FM model. At time t, an intermediate sample x_t is drawn along the transport path and the model predicts a velocity $u_\theta(x_t, t)$. Integrating this velocity to $t = 1$ yields an estimate of the clean endpoint,

$$\hat{x}_1 = x_t + (1 - t)\, u_\theta(x_t, t), \tag{3}$$

which we refer to as the single-step case. Unrolling the ODE over n steps of size $(1 - t)/n$ brings $\hat{x}_1$ closer to the true endpoint at the cost of an additional model call per step.

The choice of $\mathcal{D}$ determines which aspect of the curvature is constrained. The most direct instantiation compares the two Hessian matrices entrywise,

$$\mathcal{D}_{\mathrm{MAE}} = \frac{1}{(3N)^2} \sum_{i,j} \Big| \big[H_\phi(\hat{x}_1)\big]_{ij} - \big[H_\phi(x_1)\big]_{ij} \Big|, \tag{4}$$

which supervises the full curvature at the generated geometry without singling out any part of the spectrum. An alternative restricts the comparison to the modes that define the stationary point. Writing V and Λ for the k lowest eigenvectors and eigenvalues of $H_\phi(x_1)$, this variant penalizes the deviation of the predicted Hessian projected onto that subspace,

$$\mathcal{D}_{\mathrm{sub}} = \frac{1}{k^2} \sum_{i,j} \Big| \big[V^\top H_\phi(\hat{x}_1) V - \Lambda\big]_{ij} \Big|, \tag{5}$$

following the projection used for curvature supervision of interatomic potentials (Burger et al., 2025). Diagonalizing the reference side rather than the prediction keeps the eigendecomposition outside the computational graph, which avoids the numerical instabilities that arise when differentiating through eigenvectors of a matrix with near-degenerate eigenvalues. Both comparisons are entrywise and therefore require that the two Hessians refer to a common frame, since the Hessian is equivariant and transforms blockwise as $H_{ij} \mapsto R H_{ij} R^\top$ under a rotation R. Aligning x_0 to x_1 before constructing the transport path is sufficient, as the entire path and the equivariant surrogate then share the frame of x_1. Unless stated otherwise we use $\mathcal{D}_{\mathrm{MAE}}$, and analyze $\mathcal{D}_{\mathrm{sub}}$ and the combination of both in Sec. B.

The full objective weights the residual by t^p to down-weight the imprecise endpoint estimates obtained near $t = 0$,

$$\mathcal{L}_{\text{CurvFM}}(\theta) = \mathcal{L}_{\mathrm{FM}}(u_\theta, u_t) + \lambda\, t^p\, \mathcal{R}(\hat{x}_1, x_1), \tag{6}$$

and gradients propagate to θ through $\hat{x}_1$ only. The two terms act on complementary aspects of the same sample, with the velocity loss constraining the geometry and the residual constraining the curvature at the geometry the model produces. Figure 1 illustrates the method conceptually, and Algorithm 1 summarizes a single training step for the $\mathcal{D}_{\mathrm{MAE}}$ residual.

5 EXPERIMENTS

We evaluate CurvFM on two complementary molecular structure generation tasks, TS generation and molecular conformer generation. Across both settings, we investigate whether curvature supervision improves the physical fidelity of generated structures while preserving their geometric accuracy, and analyze which aspects of the local energy landscape account for the observed gains.

Algorithm 1 One training step of CurvFM using the $\mathcal{D}_{\mathrm{MAE}}$ residual.

Require: Flow model u_θ, frozen Hessian model H_ϕ, source sample x_0, target sample x_1, residual weight λ, exponent p

$x_0 \leftarrow \mathrm{align}(x_0, x_1)$ {common frame for the whole path}
$t \sim \mathcal{U}(0, 1)$
$x_t \leftarrow (1-t)\, x_0 + t\, x_1$
$v \leftarrow u_\theta(x_t, t)$
$\mathcal{L}_{\mathrm{FM}} \leftarrow \|v - (x_1 - x_0)\|_2^2$
$\hat{x}_1 \leftarrow x_t + (1-t)\, v$ {single-step endpoint estimate}
$H_{\mathrm{pred}} \leftarrow H_\phi(\hat{x}_1)$
$H_{\mathrm{ref}} \leftarrow \mathrm{sg}[H_\phi(x_1)]$ {target side, no gradient}
$\mathcal{R} \leftarrow \frac{1}{(3N)^2} \sum_{i,j} \left| [H_{\mathrm{pred}}]_{ij} - [H_{\mathrm{ref}}]_{ij} \right|$
$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{FM}} + \lambda\, t^p\, \mathcal{R}$
return $\mathcal{L}$

5.1 Transition state generation

Datasets and Metrics. As a first task, we consider TS generation and evaluate CurvFM on three datasets of increasing chemical diversity. Transition1x contains 10,073 gas-phase reactions at the ωB97x/6-31G(d) level involving H, C, N, and O with up to seven heavy atoms (Schreiner et al., 2022). Building on its reactions, Transition1x-2p3p4p applies elemental swaps up to the 4p group to give 14,000 reactions, and Transition1x-TMC comprises 36,446 reactions covering ten catalytically relevant TMs with up to 26 atoms, both at the GFN2-xTB level (Darouich et al., 2026a). Following Darouich et al. (2026a), each dataset is split randomly into training, validation, and test sets with an 80/10/10 ratio. As Hessian surrogate we use HIP (Burger et al., 2025). For Transition1x-2p3p4p and Transition1x-TMC we compute reference Hessians for the reactant, TS, and product structures at the level of theory of the respective dataset, and finetune the public checkpoint on them to cover the elements they contain. Further details are given in Appendix A.1. Structural accuracy is measured by the RMSD for global similarity and the distance matrix absolute error (DMAE) for finer structural differences. Since the central claim of this work concerns physical rather than geometric fidelity, we evaluate the generated structures directly on the PES, computing energies and forces at the corresponding level of theory. We then assess the curvature that CurvFM is trained on by comparing the reference Hessian of the generated structure against that of the true TS, computed at the same level of theory rather than with the surrogate.

Results. To isolate the effect of curvature supervision, we compare against the identical model trained with the structural FM loss alone under the same training budget. Table 1 reports both variants within the React-OT (Duan et al., 2025) framework across all three datasets. The global structural metrics remain essentially unchanged, consistent with their weak coupling to energetic accuracy discussed above. Resolving the error by internal coordinate shows where the geometry does change (Table 5). On Transition1x-2p3p4p and Transition1x-TMC, the median bond and angle errors decrease by up to 17 and 8%, while the softer torsions and non-bonded distances remain unchanged. The residual thus acts on the degrees of freedom along which the energy varies most steeply, and the energetic metrics improve accordingly. Both the median barrier error and the maximum force decrease by about 21% on these two datasets, reducing the median maximum force from 33.50 to 26.43 and from 46.74 to 36.58 kcal/mol/Å. On Transition1x, where the baseline is already strong, the gain is roughly 3%. These gains are obtained at an increase in training time of 30% and in peak memory of 12%, while sampling remains unchanged (Appendix B.1).

Table 2 evaluates the curvature directly, comparing the Hessians of the generated and the reference structures, both computed with quantum chemistry at the level of theory of the respective dataset. The fraction of valid TSs, those with exactly one negative eigenvalue, increases on all three datasets, from 70 to 74% on Transition1x, from 57 to 67% on Transition1x-2p3p4p, and from 32 to 41% on Transition1x-TMC. On Transition1x-2p3p4p and Transition1x-TMC these gains are reflected in a more accurate curvature overall, with both error measures improving and the median eigenvalue error decreasing by 21 and 19% respectively, whereas on Transition1x the full Hessian error remains unchanged.

Table 1: Structural and energetic metrics for TS generation. Best in bold, second-best underlined.

		RMSD (Å) ↓		D-MAE (Å) ↓		ΔE (kcal/mol) ↓		$F_{\max}$ (kcal/mol/Å) ↓	
Dataset	Model	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Transition1x	React-OT	0.188	0.108	0.068	0.042	**3.89**	1.35	33.99	26.07
Transition1x	React-OT$_{\text{CurvFM}}$	0.188	0.108	0.068	0.042	4.06	**1.31**	**33.50**	**25.33**
Transition1x-2p3p4p	React-OT	**0.196**	0.134	0.083	0.060	5.85	2.53	40.96	33.50
Transition1x-2p3p4p	React-OT$_{\text{CurvFM}}$	0.197	**0.133**	**0.082**	**0.058**	**5.19**	**2.00**	**32.69**	**26.43**
Transition1x-TMC	React-OT	0.294	**0.231**	0.125	0.101	9.92	6.40	55.78	46.74
Transition1x-TMC	React-OT$_{\text{CurvFM}}$	0.294	0.232	**0.124**	**0.100**	**8.27**	**4.91**	**43.08**	**36.58**

To understand where these gains originate, we resolve the Hessian into its eigenmodes, each describing a collective displacement of the atoms together with the curvature of the energy along it. The reactive mode is the single direction of negative curvature along which the reaction proceeds, while the remaining vibrational modes range from soft bends and torsions to stiff bond stretches. As shown in Table 8, the structural objective alone already recovers the reactive mode well, with a median overlap of 95.9% on Transition1x-2p3p4p between the predicted direction and its reference, measured as the absolute cosine similarity between the two eigenvectors. Curvature supervision raises this only to 96.6%. The residual therefore does not primarily sharpen the reaction coordinate, but corrects the curvature of the surrounding energy landscape.

This raises the question of whether the gain requires curvature at all, or whether any physical signal evaluated at the endpoint would suffice. Forces are the natural first-order alternative. They are cheaper to obtain, but they vanish at stationary points and so carry little information near the structures being generated. Replacing the Hessian with the forces at the same two endpoints, with the loss weight and time weighting retuned, recovers only about a third of the gain. Despite supervising the maximum force directly, the force residual reduces it by less than half as much as the curvature residual does (Table 13).

Given that second-order information is needed, the remaining question is which part of the spectrum carries it. Table 12 compares the full residual of Eq. 4 against the subspace variant of Eq. 5, which constrains only the projection onto the $k = 8$ lowest reference eigenvectors and eigenvalues, and against the combination of both. The subspace variant stays close to the purely structural baseline, reducing the median eigenvalue error by 6% and leaving the valid TS rate essentially unchanged, in contrast to the 21% and ten point improvements obtained with the full residual. The gap is widest among the stiff modes, where the median relative error decreases by 7% under the subspace variant and by 27% under the full residual, since a residual confined to the low-lying part of the spectrum leaves the stiff curvature, which dominates the Hessian, unconstrained. Adding the subspace term to the full residual degrades every metric slightly, so the targeted supervision does not merely fail to help but competes with the denser signal it is combined with.

Table 2: Hessian metrics for TS generation. Hessians of both the generated and the ground-truth structures are computed with quantum chemistry at the level of theory of the respective dataset. Best in bold, second-best underlined.

		Hessian MAE (kcal/mol/Å^2) ↓		Eigenvalue MAE (kcal/mol/Å^2) ↓		Valid TS (%) ↑
Dataset	Model	Mean	Median	Mean	Median	Mean
Transition1x	React-OT	**13.43**	7.95	**25.74**	17.69	70.00
Transition1x	React-OT$_{\text{CurvFM}}$	13.52	**7.94**	25.99	**16.93**	**74.00**
Transition1x-2p3p4p	React-OT	12.13	8.40	30.89	22.95	57.00
Transition1x-2p3p4p	React-OT$_{\text{CurvFM}}$	**11.39**	**7.52**	**25.31**	**18.08**	**67.00**
Transition1x-TMC	React-OT	11.13	9.14	33.24	27.57	32.00
Transition1x-TMC	React-OT$_{\text{CurvFM}}$	**10.32**	**8.32**	**26.76**	**22.43**	**41.00**

The source of the reference curvature matters as much as its formulation. A quantum-chemical target might appear preferable to a surrogate one, since it removes an approximation from the comparison, but it degrades the model below the purely structural baseline (Table 6). The reason is an irreducible floor of the surrogate model. Comparing the surrogate prediction at $\hat{x}_1$ against a quantum-chemical Hessian at x_1 leaves a residual that does not vanish even when the two geometries coincide. Once

the endpoint estimate becomes accurate enough that its own contribution falls below this floor, the residual stops varying with $\hat{x}_1$ and the gradient carries no usable information. Figure 2 shows this directly. The RMSD of the endpoint estimate decreases with t for both variants, but the residual evaluated against a quantum-chemical target stops decreasing once it reaches the surrogate accuracy. The surrogate residual continues below it, since most of its error is shared between the two geometries and cancels (Appendix B.2).

Training against a surrogate raises the question of whether the model inherits its bias. If it did, the gains would appear larger when evaluated with the surrogate than with quantum chemistry. Repeating the Hessian evaluation with HIP instead gives nearly the same picture (Table 7), so the gains reflect improved curvature rather than agreement with the surrogate used during training.

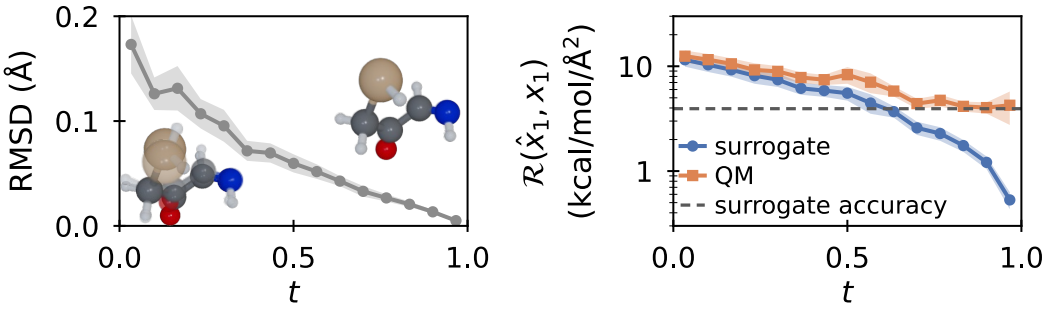

Figure 2: **A quantum-chemical reference bounds the resolvable residual.** Quality of the one-shot prediction $\hat{x}_1$ on Transition1x-2p3p4p as a function of the FM time t. Markers joined by lines are binned means and shaded regions are confidence intervals of the mean. **Left:** Kabsch-aligned RMSD between $\hat{x}_1$ and the reference TS x_1, which depends only on the predicted geometry and is therefore identical for both variants. **Right:** the residual $\mathcal{R}(\hat{x}_1, x_1)$ for the same predictions, evaluated against the surrogate and against the QM Hessian of the reference structure. The dashed line marks the error of the surrogate itself at the reference geometry.

5.2 Conformer Generation

Datasets and Metrics. As a second task, we consider conformer generation on GEOM-QM9 (Axelrod & Gomez-Bombarelli, 2022), comprising 133,258 small molecules whose reference ensembles were obtained with CREST (Pracht et al., 2024). We use the splits of Ganea et al. (2021), with 80% of the molecules for training, 10% for validation, and a held-out set of 1,000 molecular graphs for testing. For a molecule with K reference conformers we generate $2K$ samples, capped at 32, and apply the chirality correction of Ganea et al. (2021). Structural accuracy is measured by coverage (COV) and average minimum RMSD (AMR), both reported as recall and precision. Recall quantifies how completely the generated ensemble reproduces the reference conformers and precision how many of the generated conformers are themselves accurate, with a conformer counted as covered when it falls within $\delta = 0.5$ Å. As in the TS setting, these metrics compare geometries and say nothing about where the structures sit on the potential energy surface. We therefore relax all conformers with GFN2-xTB (Bannwarth et al., 2019) and compare Boltzmann-weighted ensemble properties between the generated and reference ensembles, namely the energy E, the minimum energy E_{min}, the dipole moment μ, and the HOMO-LUMO gap $\Delta\epsilon$. Finally, we assess the curvature directly by computing Hessians for up to 32 conformers per molecule and testing whether the generated structures are local minima, that is, whether their spectra are free of negative eigenvalues.

Results. To allow a controlled comparison, we take the publicly available ET-Flow checkpoint and finetune it for 50 further epochs, once with the structural FM loss alone and once with the CurvFM objective. The global structural metrics remain essentially unchanged (Table 3), whereas the energetic metrics improve substantially under residual training (Table 4). For the generated structures, the median deviation in maximum force decreases by 63%, from 15.91 to 5.81 kcal/mol/Å, and the deviation in minimum energy falls from 0.42 to 0.12 kcal/mol. After relaxation with GFN2-xTB, both models converge to nearly identical ensembles, confirming that each recovers the intended

conformational distribution. What separates them is how close the generated structures already are to it. That proximity is a matter of curvature, and Table 14 shows how it changes. The median fraction of generated conformers that are local minima rises to 100%, so a majority of molecules now yield fully valid ensembles, while the mean is marginally lower. Training on curvature therefore places the generated structures nearer to genuine minima. This is reflected in the relaxation itself, where they reach equilibrium in fewer optimization steps, with the median dropping from 5.4 to 4.8 (Table 15).

CurvFM is model-agnostic, since the residual acts on the endpoint estimate rather than on the architecture that produces it. This raises the question of how its effect compares to that of a stronger architecture. We therefore compare ET-Flow with CurvFM against DiTMC (Frank et al., 2025), an independently developed model that reaches better structural accuracy on GEOM-QM9 (Table 16). The two reach comparable physical fidelity, with DiTMC ahead on the ensemble energies and ET-Flow with CurvFM ahead on the maximum force (Table 17). Curvature supervision applied to the weaker model therefore closes the gap in physical fidelity to the stronger architecture, and remains available to it.

Table 3: Molecule conformer generation results on GEOM-QM9 (δ = 0.5Å). Best in bold, second-best underlined.

	Coverage-R (%) ↑		AMR-R (Å) ↓		Coverage-P (%) ↑		AMR-P (Å) ↓	
Model	Mean	Median	Mean	Median	Mean	Median	Mean	Median
ET-Flow	<u>94.95</u>	100.00	<u>0.088</u>	<u>0.037</u>	**91.45**	100.00	**0.120**	<u>0.055</u>
ET-Flow$_{\text{CurvFM}}$	**95.97**	100.00	**0.080**	**0.035**	<u>91.33</u>	100.00	<u>0.121</u>	**0.054**

Table 4: Ensemble property deviations on GEOM-QM9 computed at GFN2-xTB, before and after geometry optimization. Best in bold, second-best underlined.

		$\lvert\Delta\langle E\rangle\rvert$ (kcal/mol) ↓		$\lvert\Delta\langle F_{\max}\rangle\rvert$ (kcal/mol/Å) ↓		$\lvert\Delta\langle\mu\rangle\rvert$ (D) ↓		$\lvert\Delta\langle\Delta\epsilon\rangle\rvert$ (kcal/mol) ↓		$\lvert\Delta\langle E_{min}\rangle\rvert$ (kcal/mol) ↓	
Model	Stage	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
ET-Flow	Generated	<u>0.64</u>	<u>0.49</u>	<u>16.90</u>	<u>15.91</u>	<u>0.16</u>	<u>0.05</u>	<u>1.78</u>	<u>1.18</u>	<u>0.58</u>	<u>0.42</u>
ET-Flow$_{\text{CurvFM}}$	Generated	**0.33**	**0.15**	**6.37**	**5.81**	**0.15**	**0.04**	**1.05**	**0.44**	**0.31**	**0.12**
ET-Flow	Optimized	**0.23**	**0.05**	0.77	0.78	<u>0.15</u>	**0.03**	<u>0.72</u>	<u>0.21</u>	**0.18**	0.02
ET-Flow$_{\text{CurvFM}}$	Optimized	<u>0.27</u>	<u>0.06</u>	0.77	0.78	**0.14**	<u>0.04</u>	**0.65**	**0.20**	<u>0.23</u>	0.02

6 Discussion

Limitations. CurvFM requires a Hessian prediction model for the chemical domain of interest, which currently restricts it to systems covered by the available surrogates. Reference Hessians for training such a model remain expensive to compute, so extending the method to domains with no suitable predictor requires either transferring an existing model out of distribution or generating new reference data. Evaluating the residual adds one surrogate call per training step increasing training time and memory accordingly, although sampling remains unaffected. The residual is also only as informative as the surrogate is smooth. Since both sides are evaluated with the same model, systematic error cancels, but the gradient still relies on the surrogate responding consistently to changes in geometry, which is not guaranteed far from its training distribution. Finally, the gains we observe are consistent but do not place the generated structures at genuine stationary points, so a subsequent relaxation remains necessary before they can be used.

Future work. The Hessian carries more information than the classification of a stationary point. Its spectrum determines a range of observable second-order properties, so the same residual could supervise any of them and enable generation conditioned on such a target, for instance a vibrational spectrum. Another direction concerns the surrogate itself. Since both evaluations of the residual are performed with the same model, its systematic error cancels, and the accuracy the surrogate needs to reach is an open question. Establishing this would clarify whether cheaper or more narrowly trained Hessian models can be used without loss of quality.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, to draft sections of the paper, and to support research execution through code generation for analysis and evaluation scripts. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement the methods presented in this work, support qualitative or thematic data analysis, or interpret results. All AI-assisted work was reviewed by the authors, with AI-assisted text revised as needed and AI-generated code checked for correctness. The experimental design, model development, and interpretation of results were conducted by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All components needed to reproduce our results are described in the paper. The residual and the full training objective are defined in Sec. 4, with a single training step summarized in Algorithm 1. The datasets, splits, and evaluation metrics are specified in Sec. 5, and the metric definitions are given in full in Appendix A.2. Appendix A reports the training setup for both tasks, including the checkpoints we start from, the number of epochs, learning rates, batch sizes, hardware, and the procedure used to select the residual weight and the time-weighting exponent. Additionally, it describes how the reference Hessians for Transition1x-2p3p4p and Transition1x-TMC were computed and how the Hessian prediction model was finetuned on them, together with its accuracy on each dataset. Upon acceptance, the code for training and evaluation will be released on GitHub, and the generated structures, reference Hessians, and model checkpoints will be made available on Zenodo.

REFERENCES

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific data*, 9(1):185, 2022.

Giacomo Baldan, Qiang Liu, Alberto Guardone, and Nils Thuerey. Physics vs distributions: Pareto optimal flow matching with physics constraints. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ajit Banerjee, Noah Adams, Jack Simons, and Ron Shepard. Search for stationary points on surfaces. *The Journal of Physical Chemistry*, 89(1):52–57, 1985.

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.

Jan-Hendrik Bastek, WaiChing Sun, and Dennis Kochmann. Physics-informed diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alex Berlaga, Michael S Jones, and Andrew L Ferguson. Flowback-adjoint: Physics-aware and energy-guided conditional flow-matching for all-atom protein backmapping. *arXiv preprint arXiv:2508.03619*, 2025.

Achinthya Krishna Bheemaguli, Penghao Xiao, and Gopalakrishnan Sai Gautam. Evaluation of foundational machine learned interatomic potentials for migration barrier predictions. *Digital Discovery*, 5(4):1809–1819, 2026.

Andreas Burger, Luca Thiede, Nikolaj Rønne, Varinia Bernales, Nandita Vijaykumar, Tejs Vegge, Arghya Bhowmik, and Alan Aspuru-Guzik. Shoot from the hip: Hessian interatomic potentials without derivatives. *arXiv preprint arXiv:2509.21624*, 2025.

Zhonglin Cao, Mario Geiger, Allan Dos Santos Costa, Danny Reidenbach, Karsten Kreis, Tomas Geffner, Franco Pellegrini, Guoqing Zhou, and Emine Kucukbenli. Efficient molecular conformer generation with SO(3)-averaged flow matching and reflow. In *Forty-second International Conference on Machine Learning*, 2025.

Vlad Cărare, Fabian L Thiemann, Joe D Morrow, David J Wales, Edward O Pyzer-Knapp, and Luke Dicks. Global properties of the energy landscape: a testing and training arena for machine learned potentials. *npj Computational Materials*, 12(1):9, 2025.

Taoyong Cui, Yunhong Han, Haojun Jia, Chenru Duan, and Qiyuan Zhao. A large scale molecular hessian database for optimizing reactive machine learning interatomic potentials. *Scientific Data*, 13(1):37, 2025.

Samir Darouich, Jacob W Toney, Weiliang Luo, Johannes Kästner, Mathias Niepert, and Heather J Kulik. Robust generative transition-state models for unseen chemistry. *Nature Computational Science*, pp. 1–11, 2026a.

Samir Darouich, Vinh Tong, Tanja Bien, Johannes Kästner, and Mathias Niepert. Adaptive transition-state refinement with learned equilibrium flows. *Journal of Chemical Information and Modeling*, 66(4):2154–2165, 2026b.

Samir Darouich, Vinh Tong, Lluís Pastor-Pérez, Tanja Bien, Loay Mualem, and Mathias Niepert. Symdrift: One-shot generative modeling under symmetries. *arXiv preprint arXiv:2605.06140*, 2026c.

Jasper De Landsheere, Maximilian P-P Kovar, Konstantin Mark, Leon Ganser, Leonard Galustian, Johannes Karwounopoulos, Charlotte Gerhaher, and Esther Heid. Toward on-the-fly prediction of reaction energetics for high-throughput screening. *Chemical Physics Reviews*, 7(2), 2026.

Amanda L Dewyer, Alonso J Argüelles, and Paul M Zimmerman. Methods for exploring reaction space in molecular systems. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8 (2):e1354, 2018.

Yuanqi Du, Limei Wang, Dieqiao Feng, Guifeng Wang, Shuiwang Ji, Carla P Gomes, Zhi-Ming Ma, et al. A new perspective on building efficient and expressive 3d equivariant graph neural networks. *Advances in neural information processing systems*, 36:66647–66674, 2023.

Chenru Duan, Yuanqi Du, Haojun Jia, and Heather J Kulik. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. *Nature computational science*, 3 (12):1045–1055, 2023.

Chenru Duan, Guan-Horng Liu, Yuanqi Du, Tianrong Chen, Qiyuan Zhao, Haojun Jia, Carla P Gomes, Evangelos A Theodorou, and Heather J Kulik. Optimal transport for generating transition states in chemical reactions. *Nature Machine Intelligence*, 7(4):615–626, 2025.

Thorben Frank, Winfried Ripken, Gregor Lied, Klaus Robert Muller, Oliver T. Unke, and Stefan Chmiela. Sampling 3d molecular conformers with diffusion transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Leonard Galustian, Konstantin Mark, Johannes Karwounopoulos, Maximilian P-P Kovar, and Esther Heid. Goflow: efficient transition state geometry prediction with flow matching and e (3)-equivariant neural networks. *Digital Discovery*, 4(12):3492–3501, 2025.

Octavian-Eugen Ganea, Lagnajit Pattanaik, Connor W. Coley, Regina Barzilay, Klavs Jensen, William Green, and Tommi S. Jaakkola. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Majdi Hassan, Nikhil Shenoy, Jungyoon Lee, Hannes Stärk, Stephan Thaler, and Dominique Beaini. Et-flow: Equivariant flow-matching for molecular conformer generation. *Advances in Neural Information Processing Systems*, 37:128798–128824, 2024.

Graeme Henkelman and Hannes Jónsson. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.*, 111(15):7010–7022, October 1999. ISSN 0021-9606.

Graeme Henkelman, Blas P. Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.*, 113(22):9901–9904, December 2000. ISSN 0021-9606.

Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. Diffusionpde: Generative pde-solving under partial observation. *Advances in Neural Information Processing Systems*, 37: 130291–130323, 2024.

Christian Jacobsen, Yilin Zhuang, and Karthik Duraisamy. Cocogen: Physically consistent and conditioned score-based generative models for forward and inverse problems. *SIAM Journal on Scientific Computing*, 47(2):C399–C425, 2025.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *Advances in neural information processing systems*, 35:24240–24253, 2022.

Hannes Jónsson, Greg Mills, and Karsten W Jacobsen. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and quantum dynamics in condensed phase simulations*, pp. 385–404. World Scientific, 1998.

Seonghwan Kim, Jeheon Woo, and Woo Youn Kim. Diffusion-based generative ai for exploring transition states from 2d molecular graphs. *Nature Communications*, 15(1):341, 2024.

Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a python library for working with atoms. *J. Phys.:Condens. Matter*, 29(27):273002, 2017.

E.G. Lewars. *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*. Chemistry and Materials Science. Springer International Publishing, 2024. ISBN 9783031514432.

Yunyang Li, Lin Huang, Luojia Xia, Wenhe Zhang, and Mark Gerstein. Elign: Equivariant diffusion model alignment from foundational machine learning force fields. *arXiv preprint arXiv:2601.21985*, 2026.

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

Zhiyuan Liu, Yanchen Luo, Han Huang, Enzhi Zhang, Sihang Li, Junfeng Fang, Yaorui Shi, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. NExt-mol: 3d diffusion meets 1d language modeling for 3d molecule generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shitong Luo, Chence Shi, Minkai Xu, and Jian Tang. Predicting molecular conformation via dynamic graph score matching. *Advances in neural information processing systems*, 34:19784–19795, 2021.

Konstantin Mark, Leonard Galustian, Maximilian P-P Kovar, and Esther Heid. Feynman-kac-flow: Inference steering of conditional flow matching to an energy-tilted posterior. *arXiv preprint arXiv:2509.01543*, 2025.

Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):33, 2011.

Philipp Pracht, Stefan Grimme, Christoph Bannwarth, Fabian Bohle, Sebastian Ehlert, Gereon Feldmann, Johannes Gorges, Marcel Müller, Tim Neudecker, Christoph Plett, et al. Crest—a program for the exploration of low-energy molecular chemical space. *The Journal of Chemical Physics*, 160(11), 2024.

Sereina Riniker and Gregory A Landrum. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling*, 55(12): 2562–2574, 2015.

Austin Rodriguez, Justin S Smith, and Jose L Mendoza-Cortes. Does hessian data improve the performance of machine learning potentials? *Journal of Chemical Theory and Computation*, 21 (14):6698, 2025.

Austin Rodriguez, Justin S Smith, Sakib Matin, Nicholas Lubbers, Kipton Barros, and Jose L Mendoza-Cortes. Projected hessian learning: Fast curvature supervision for accurate machine-learning interatomic potentials. *arXiv preprint arXiv:2603.04523*, 2026.

Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x-a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9(1):779, 2022.

Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. In *International conference on machine learning*, pp. 9558–9568. PMLR, 2021.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Utkarsh Utkarsh, Pengfei Cai, Alan Edelman, Rafael Gomez-Bombarelli, and Christopher Vincent Rackauckas. Physics-constrained flow matching: Sampling generative models with hard constraints. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Brook Wander, Muhammed Shuaibi, John R. Kitchin, Zachary W. Ulissi, and C. Lawrence Zitnick. Cattsunami: Accelerating transition state energy calculations with pretrained graph neural networks. *ACS Catal.*, 15(7):5283–5294, 2025.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, 2025.

Guanchen Wu, Eric C-Y Yuan, Kareem Hegazy, Samuel M Blau, and Teresa Head-Gordon. Machine-learned leftmost hessian eigenvectors for robust transition state finding. *Journal of Chemical Theory and Computation*, 2026.

Guikun Xu, Xiaohan Yi, Peilin Zhao, and Yatao Bian. Energy-guided flow matching enables few-step conformer generation and ground-state identification. *arXiv preprint arXiv:2512.22597*, 2025.

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.

Bangchen Yin, Jian Ouyang, Zhen Fan, Kailai Lin, Hanshi Hu, Dingshun Lv, Weiluo Ren, Hai Xiao, Ji Chen, and Changsu Cao. Hessian-informed machine learning interatomic potential towards bridging theory and experiments. *arXiv preprint arXiv:2603.25373*, 2026.

Eric C-Y Yuan, Anup Kumar, Xingyi Guan, Eric D Hermes, Andrew S Rosen, Judit Zádor, Teresa Head-Gordon, and Samuel M Blau. Analytical ab initio hessian from a deep learning potential for transition state optimization. *Nature Communications*, 15(1):8865, 2024.

Yi Zhang, Peng Wang, and Difan Zou. Physics-informed distillation of diffusion models for pde-constrained generation. *arXiv preprint arXiv:2505.22391*, 2025.

Qiyuan Zhao, Yunhong Han, Duo Zhang, Jiaxu Wang, Peichen Zhong, Taoyong Cui, Bangchen Yin, Yirui Cao, Haojun Jia, and Chenru Duan. Harnessing machine learning to enhance transition state search with interatomic potentials and generative models. *Advanced Science*, 12(34):e06240, 2025.

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.

CURVATURE-AWARE FLOW MATCHING FOR MOLECULAR STRUCTURE GENERATION

ADDITIONAL MATERIAL

A Implementation details **15**

A.1 Training details 15

A.2 Evaluation Metrics 15

B Additional experiments **17**

B.1 Cost of the curvature residual 17

B.2 Transition state generation. 17

B.3 Molecular conformer generation. 21

A Implementation details

A.1 Training details

Hessian prediction model. For Transition1x we use the publicly available EquiformerV2 HIP checkpoint (Burger et al., 2025) without modification, since it is trained at the same level of theory and on the same elements. For Transition1x-2p3p4p and Transition1x-TMC we compute reference Hessians for every reactant, TS, and product structure by finite differences of the analytical forces, using ASE (Larsen et al., 2017) together with the Python interface of GFN2-xTB (Bannwarth et al., 2019). The checkpoint is then finetuned on these Hessians for 400 epochs with a batch size of 5 and a cosine learning rate schedule from 5×10^{-4} to 5×10^{-5}. The accuracy of the resulting surrogates is verified on the whole dataset, giving a Hessian MAE of 0.54, 3.51, and 3.25 kcal/mol/Å^2 on Transition1x, Transition1x-2p3p4p, and Transition1x-TMC. For GEOM-QM9 we reuse the Transition1x-2p3p4p model, which shares the GFN2-xTB level of theory, and confirm its accuracy on a subset of 5000 GEOM-QM9 conformers, which include Fluorine, a element not present in the Transition1x or Transition1x-2p3p4p dataset, still leading to a similar Hessian MAE of 3.37 kcal/mol/Å^2.

Transition state generation. We start from the adapted OA-ReactDiff checkpoint of Darouich et al. (2026a), which replaces the one-hot atom-type encoding of the original LEFTNet backbone (Du et al., 2023) with a learnable embedding of the atomic number supplemented by selected electronic properties, so that elements beyond H, C, N, and O are covered. On each dataset, the checkpoint is first finetuned for 200 epochs with the React-OT flow matching loss at a learning rate of 2.5×10^{-4}. From this shared starting point, both variants are trained for a further 200 epochs at the same learning rate, the baseline with the FM loss alone and CurvFM with the residual of Eq. 4. The weight λ and the exponent p are optimized on the validation set using Bayesian optimization. Training uses a batch size of 16 on a single A100 GPU.

Molecular conformer generation. We follow ET-Flow (Hassan et al., 2024) and use a harmonic prior defined by the connectivity of the molecular graph, removing the center of mass from each sampled configuration to enforce translational invariance. Starting from the publicly available GEOM-QM9 checkpoint, both variants are finetuned for a further 50 epochs, the baseline with the FM loss alone and CurvFM with the residual of Eq. 4. The weight λ and the exponent p are again optimized on the validation set using Bayesian optimization. Training uses a fixed learning rate of 2.5×10^{-4} and a batch size of 64 on a single A100 GPU.

A.2 Evaluation Metrics

Transition state generation. Structural agreement between a generated structure $\hat{\mathbf{x}}_1$ and its reference $\mathbf{x}_1$ is quantified after Kabsch alignment by

$$\mathrm{RMSD}(\hat{\mathbf{x}}_1, \mathbf{x}_1) = \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \hat{\mathbf{x}}_{1,i} - \mathbf{x}_{1,i} \right\|^2 \right)^{1/2}, \tag{7}$$

where N is the number of atoms and $\mathbf{x}_{1,i} \in \mathbb{R}^3$ collects the Cartesian coordinates of atom i. Since the RMSD summarizes the deviation of the structure as a whole, we additionally report the distance matrix mean absolute error (DMAE), which is sensitive to local distortions and requires no alignment,

$$\mathrm{DMAE}(\hat{\mathbf{x}}_1, \mathbf{x}_1) = \frac{1}{N(N-1)} \sum_{i \neq j} \left| \hat{d}_{ij} - d_{ij} \right|, \tag{8}$$

with $d_{ij} = \|\mathbf{x}_{1,i} - \mathbf{x}_{1,j}\|$ the interatomic distance between atoms i and j.

Physical fidelity is assessed on the PES. The deviation in barrier height follows from the electronic energies of the two structures,

$$\Delta E = \left| E(\hat{\mathbf{x}}_1) - E(\mathbf{x}_1) \right|, \tag{9}$$

which coincides with the error in the barrier of the corresponding reaction, since both energies are referenced to the same reactant. How close a generated structure lies to a stationary point is measured

by the largest force component acting on it,

$$F_{\max} = \max_{i,j} \left| f_{ij}(\hat{\mathbf{x}}_1) \right|, \tag{10}$$

with f_{ij} the component $j \in \{x, y, z\}$ of the force on atom i. In contrast to the preceding metrics, $F_{\max}$ is evaluated on the generated structure alone and vanishes exactly at a stationary point.

Curvature is evaluated by comparing the Hessian $H(\hat{\mathbf{x}}_1) \in \mathbb{R}^{3N \times 3N}$ of the generated structure against the Hessian $H(\mathbf{x}_1)$ of the reference TS, both computed at the level of theory of the respective dataset. Since the Hessian is equivariant, the generated structure is rotated into the frame of the reference before the comparison,

$$\text{Hessian MAE} = \frac{1}{(3N)^2} \sum_{i,j} \left| H_{ij}(\hat{\mathbf{x}}_1) - H_{ij}(\mathbf{x}_1) \right|. \tag{11}$$

Writing $\lambda_1 \leq \cdots \leq \lambda_{3N}$ for the eigenvalues of $H(\mathbf{x}_1)$ and $\hat{\lambda}_1 \leq \cdots \leq \hat{\lambda}_{3N}$ for those of $H(\hat{\mathbf{x}}_1)$, the corresponding spectral error is

$$\text{Eigenvalue MAE} = \frac{1}{3N} \sum_{i=1}^{3N} \left| \hat{\lambda}_i - \lambda_i \right|, \tag{12}$$

The eigenvalue MAE treats all modes alike, although their physical roles and magnitudes differ considerably. We therefore resolve the spectrum further, working in mass-weighted coordinates and projecting out translations and rotations with the projector built from the respective geometry, which leaves $3N - 6$ vibrational modes ordered by their eigenvalues $\omega_1 \leq \cdots \leq \omega_{3N-6}$. The lowest of these is the reactive mode, and its direction is compared through the absolute cosine similarity of the corresponding eigenvectors,

$$\text{Neg. Mode Overlap} = \left| \hat{\mathbf{v}}_1 \cdot \mathbf{v}_1 \right|, \tag{13}$$

where the absolute value accounts for eigenvectors being defined only up to sign.

The remaining modes are grouped by their reference wavenumber, with those below 500 cm^{-1} classified as soft, corresponding to bends and torsions, and the rest as stiff, corresponding to bond stretches. For each group we report the relative eigenvalue error

$$\text{MRE} = \underset{i \in \mathcal{S}}{\text{median}} \, \frac{\left| \hat{\omega}_i - \omega_i \right|}{\left| \omega_i \right|}, \tag{14}$$

with $\mathcal{S}$ the index set of the respective group. The median is used rather than the mean because soft modes have small $|\omega_i|$ by construction, so a few of them dominate the mean relative error without reflecting a correspondingly worse prediction. Finally, a generated structure is counted as a valid TS if its spectrum contains exactly one negative eigenvalue after projecting out the six translational and rotational modes.

Molecular conformer generation. For each test molecule with L reference conformers, $K = 2L$ conformers are generated and compared against the reference ensemble. A conformer $C \in \mathbb{R}^{3N}$ assigns Cartesian coordinates to every atom of a given molecular graph, and the quality of an ensemble is assessed by how well the two sets cover one another. Following prior work, this is quantified by the coverage (COV) and the average minimum RMSD (AMR), each reported in a recall and a precision variant, with $\{C_l^*\}_{l=1}^{L}$ the reference conformers, $\{C_k\}_{k=1}^{K}$ the generated ones, and δ an RMSD threshold below which two conformers count as matching.

The recall variants measure how completely the reference ensemble is reproduced,

$$\text{COV-R} = \frac{1}{L} \left| \left\{ l \mid \exists k, \ \text{RMSD}(C_l^*, C_k) < \delta \right\} \right|, \tag{15}$$

$$\text{AMR-R} = \frac{1}{L} \sum_{l=1}^{L} \min_{k} \ \text{RMSD}(C_l^*, C_k), \tag{16}$$

so that a high COV-R and a low AMR-R indicate that few reference conformers are missed. The precision variants exchange the roles of the two sets,

$$\text{COV-P} = \frac{1}{K}\Big|\big\{k \mid \exists l,\ \text{RMSD}(C_k, C_l^*) < \delta\big\}\Big|, \tag{17}$$

$$\text{AMR-P} = \frac{1}{K}\sum_{k=1}^{K} \min_{l}\ \text{RMSD}(C_k, C_l^*), \tag{18}$$

and measure instead how many generated conformers correspond to an actual reference conformer. Recall alone can be satisfied by an ensemble that is broader than the reference, and precision alone by one that is narrower, so the two are reported together.

We additionally compute for both, the generated and the reference ensemble of each molecule, the Boltzmann-weighted average of a property A,

$$\langle A\rangle = \sum_k w_k A_k, \qquad w_k = \frac{\exp(-E_k/k_BT)}{\sum_{k'}\exp(-E_{k'}/k_BT)}, \tag{19}$$

with E_k the electronic energy of conformer k and $T = 298.15$ K. Energies, forces, dipole moments μ and HOMO-LUMO gaps $\Delta\epsilon$ are obtained with GFN2-xTB (Bannwarth et al., 2019). The reported metric is the absolute deviation between the two ensembles,

$$\big|\Delta\langle A\rangle\big| = \big|\langle A\rangle_{\text{gen}} - \langle A\rangle_{\text{ref}}\big|, \tag{20}$$

averaged over the test molecules. The minimum energy E_{min} is treated separately, since it characterizes the lowest-lying conformer rather than the ensemble as a whole and is therefore compared directly rather than through a weighted average. Each metric is evaluated at two stages. In the generated stage the properties are computed on the conformers as produced by the model, which reflects how far the ensemble sits from the PES. In the optimized stage they are computed after relaxing every conformer with GFN2-xTB, which removes local distortions and isolates the conformational composition of the ensemble from the accuracy of the individual geometries.

B Additional experiments

B.1 Cost of the curvature residual

Imposing the residual at every training step is only practical because the surrogate is cheap to evaluate. A Hessian obtained by finite differences with GFN2-xTB takes on average 1.2 s per structure on Transition1x-2p3p4p and 4.1 s on Transition1x-TMC, the larger value reflecting the larger systems. HIP returns the same quantity in 51 and 52 ms, and its near-constant cost reflects the different scaling of the two approaches, since a finite-difference Hessian requires a number of force evaluations proportional to $3N$ while the predictor needs a single forward pass. The resulting overhead during training is modest. On Transition1x-2p3p4p, the baseline requires 23 hours on a single A100 GPU against 30 hours for CurvFM, an increase of 30%, with peak memory rising from 32 to 36 GB. Since the residual is confined to training, sampling remains unaffected.

B.2 Transition state generation.

Bond-resolved error analysis. Bonds are perceived with OpenBabel (O'Boyle et al., 2011) on the reference reactant and product structures, and their union defines the bond graph used throughout. Angles are taken as all triples of atoms connected through a common center, torsions as all chains of four connected atoms, and the non-bonded set as every atom pair not connected by a bond. For a set $\mathcal{S}$ of internal coordinates, the error of a generated structure $\hat{\mathbf{x}}_1$ against its reference TS $\mathbf{x}_1$ is

$$\text{MAE}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\big|\hat{q}_s - q_s\big|, \tag{21}$$

where q_s denotes the corresponding distance, angle, or torsion, and torsional differences are wrapped to $[0^\circ, 180^\circ]$. Table 5 reports the mean and median of this quantity over the structures of each test set. On Transition1x-2p3p4p and Transition1x-TMC the median bond error decreases by 17 and 13% and the angle error by 8% in both cases, while torsions and non-bonded distances change by at most 2% and in some cases marginally for the worse.

Table 5: Structure error at the generated TSs resolved by internal coordinate. Mean and median of the per-structure mean absolute error. Best in bold, second-best underlined.

		Bonded (Å) ↓		Angle (°) ↓		Torsion (°) ↓		Non-bonded (Å) ↓	
Dataset	Model	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Transition1x	React-OT	<u>0.024</u>	<u>0.012</u>	<u>2.46</u>	**1.50**	**8.94**	**4.64**	**0.077**	**0.047**
Transition1x	React-OT$_{\text{CurvFM}}$	**0.024**	**0.012**	**2.45**	<u>1.53</u>	<u>8.94</u>	<u>4.65</u>	<u>0.077</u>	<u>0.047</u>
Transition1x-2p3p4p	React-OT	<u>0.029</u>	<u>0.018</u>	<u>2.89</u>	<u>1.99</u>	**8.01**	**5.03**	**0.082**	<u>0.057</u>
Transition1x-2p3p4p	React-OT$_{\text{CurvFM}}$	**0.027**	**0.015**	**2.78**	**1.84**	<u>8.07</u>	<u>5.06</u>	<u>0.082</u>	**0.056**
Transition1x-TMC	React-OT	<u>0.043</u>	<u>0.032</u>	<u>3.98</u>	<u>3.29</u>	<u>10.36</u>	<u>7.79</u>	<u>0.137</u>	<u>0.112</u>
Transition1x-TMC	React-OT$_{\text{CurvFM}}$	**0.039**	**0.028**	**3.80**	**3.04**	**10.26**	**7.70**	**0.136**	**0.111**

Source of the reference Hessian. Table 6 reports the two choices of reference Hessian, both trained on Transition1x-2p3p4p under identical settings and compared against the purely structural baseline. With the surrogate on both sides, the median barrier error and maximum force decrease by 21 and 21%. With a quantum-chemical reference they increase by 66 and 48% relative to the same baseline, so the residual is not merely uninformative in this case but actively harmful. The structural metrics are affected far less, rising by 2 to 5%, which is consistent with the residual acting on the energetic rather than the geometric part of the objective.

Table 6: Structural and energetic metrics for TS generation. Best in bold, second-best underlined.

		RMSD (Å) ↓		D-MAE (Å) ↓		ΔE (kcal/mol) ↓		F_{max} (kcal/mol/Å) ↓	
Dataset	Model	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Transition1x-2p3p4p	React-OT	**0.196**	<u>0.134</u>	<u>0.083</u>	<u>0.060</u>	<u>5.85</u>	<u>2.53</u>	<u>40.96</u>	<u>33.50</u>
Transition1x-2p3p4p	React-OT$_{\text{CurvFM, HIP}}$	<u>0.197</u>	**0.133**	**0.082**	**0.058**	**5.19**	**2.00**	**32.69**	**26.43**
Transition1x-2p3p4p	React-OT$_{\text{CurvFM, QM}}$	0.199	0.137	0.087	0.062	7.30	4.20	56.34	49.70

The behavior in Table 6 follows from how the surrogate error enters the two residuals. Writing $\epsilon(z) = H_{\text{HIP}}(z) - H_{\text{QM}}(z)$ for the surrogate error at a geometry z, the two choices decompose as

$$H_{\text{HIP}}(\hat{x}_1) - H_{\text{HIP}}(x_1) = \underbrace{H_{\text{QM}}(\hat{x}_1) - H_{\text{QM}}(x_1)}_{\text{oracle residual}} + \underbrace{\epsilon(\hat{x}_1) - \epsilon(x_1)}_{\text{contamination}}, \tag{22}$$

$$H_{\text{HIP}}(\hat{x}_1) - H_{\text{QM}}(x_1) = \underbrace{H_{\text{QM}}(\hat{x}_1) - H_{\text{QM}}(x_1)}_{\text{oracle residual}} + \underbrace{\epsilon(\hat{x}_1)}_{\text{contamination}}. \tag{23}$$

The oracle term is what the residual should measure, namely the change in the true Hessian caused by the error in the endpoint estimate. The two choices differ only in what accompanies it. Eq. 23 carries the full surrogate error, which does not vanish as $\hat{x}_1 \to x_1$, whereas Eq. 22 carries only the difference of the surrogate errors at two nearby geometries, which vanishes whenever ϵ varies smoothly in z, however large ϵ itself may be.

To verify this, we used the one-shot predictions $\hat{x}_1$ on reactions from Transition1x-2p3p4p shown in Figure 2 and compute the Hessian using the GFN2-xTB level of theory. Figure 3 shows $\langle|\epsilon(x_1)|\rangle$, $\langle|\epsilon(\hat{x}_1)|\rangle$ and $\langle|\epsilon(\hat{x}_1) - \epsilon(x_1)|\rangle$ against the RMSD between the two geometries. Below 0.01 Å the two individual errors are nearly equal, at 3.16 and 3.18 kcal/mol/Å^2, while their difference is 0.25, which is 8% of the individual error, so 92% of the surrogate error cancels in Eq. 22.

Bias from the surrogate. Since CurvFM is trained against HIP predictions, the evaluation could in principle reward structures on which the surrogate agrees with itself rather than structures that are physically correct. To test this, we repeat the Hessian evaluation of Table 2 using HIP in place of quantum chemistry for both the generated and the reference structures. The results are reported in Table 7. Both error measures improve under CurvFM on all three datasets, and the valid TS rate rises on each, in agreement with the quantum-chemical evaluation. The absolute values differ, most visibly in the valid TS rate on Transition1x-TMC, where HIP classifies a larger fraction of structures as valid than quantum chemistry does, but the ordering of the two models is the same throughout.

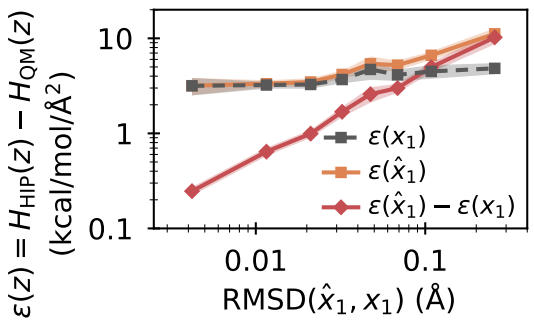

Figure 3: **Surrogate error cancels between the two endpoints.** Mean absolute surrogate error $\epsilon(z) = H_{\mathrm{HIP}}(z) - H_{\mathrm{QM}}(z)$ at the reference TS x_1 and at the endpoint estimate $\hat{x}_1$, together with their difference, binned by the RMSD between the two geometries on Transition1x-2p3p4p. Markers joined by lines are binned means and shaded regions are confidence intervals of the mean.

Table 7: Hessian metrics for TS generation. Hessians of both the generated and the ground-truth structures are computed with the HIP model of the respective dataset. Best in bold, second-best underlined.

		Hessian MAE (kcal/mol/Å^2) ↓		Eigenvalue MAE (kcal/mol/Å^2) ↓		Valid TS (%) ↑
Dataset	Model	Mean	Median	Mean	Median	Mean
Transition1x	React-OT	13.26	**7.78**	25.36	17.61	71.00
Transition1x	React-OT$_{\mathrm{CurvFM}}$	**13.20**	7.93	**24.86**	**16.94**	**74.00**
Transition1x-2p3p4p	React-OT	11.72	8.48	28.12	21.71	61.00
Transition1x-2p3p4p	React-OT$_{\mathrm{CurvFM}}$	**11.04**	**7.49**	**23.51**	**17.08**	**64.00**
Transition1x-TMC	React-OT	11.19	9.35	36.32	29.37	49.00
Transition1x-TMC	React-OT$_{\mathrm{CurvFM}}$	**10.38**	**8.56**	**28.59**	**22.35**	**55.00**

Spectrally resolved Hessian metrics. Table 8 resolves the Hessian error by mode, separating soft from stiff vibrations at a reference wavenumber of 500 cm^{-1} and reporting the overlap of the reactive mode separately. On Transition1x-2p3p4p and Transition1x-TMC both groups improve under CurvFM, the soft modes by 19 and 14% and the stiff modes by 27 and 28% in the median relative error. On Transition1x the improvement is smaller in both groups. The reactive mode changes little on any dataset, with the median overlap rising by less than one point and the mean on Transition1x marginally lower, so the structural objective already reproduces this direction well and the residual acts mainly on the remaining vibrational modes.

Table 8: Hessian metrics for TS generation, computed against ground-truth QM Hessians. Soft modes are assigned by $|\lambda_i| < 500$ cm^{-1}. Best in bold, second-best underlined.

		Soft Eigenvalue MRE (%) ↓		Stiff Eigenvalue MRE (%) ↓		Neg. Mode Overlap (%) ↑	
Dataset	Model	Mean	Median	Mean	Median	Mean	Median
Transition1x	React-OT	**22.57**	11.78	2.53	1.80	**92.35**	99.05
Transition1x	React-OT$_{\mathrm{CurvFM}}$	22.60	**10.44**	**2.45**	**1.68**	92.04	**99.08**
Transition1x-2p3p4p	React-OT	25.04	16.93	3.48	2.61	82.42	95.92
Transition1x-2p3p4p	React-OT$_{\mathrm{CurvFM}}$	**21.89**	**13.79**	**2.68**	**1.90**	**85.29**	**96.57**
Transition1x-TMC	React-OT	22.27	17.30	4.39	3.60	80.91	94.84
Transition1x-TMC	React-OT$_{\mathrm{CurvFM}}$	**19.23**	**14.89**	**3.25**	**2.59**	**82.91**	**95.41**

Downstream behavior. Beyond the static metrics, we measure downstream utility by running a saddle-point optimization from each generated structure at the level of theory of the respective dataset, which is ωB97X/6-31G(d) for Transition1x and GFN2-xTB for Transition1x-2p3p4p and Transition1x-TMC. The difference of two orders of magnitude in the optimization time reported in Table 9 reflects this difference in level of theory rather than a property of the models. The median RMSD after optimization is zero for both variants on all three datasets, so the optimizations recover the reference TS exactly for the majority of reactions. The nonzero means, between 0.08 and 0.27 Å,

come from a minority of cases that converge to a different saddle point, and their share is the same for both variants. The comparison therefore concerns how quickly the intended structure is reached rather than whether it is reached. On Transition1x-2p3p4p the mean number of steps decreases by 18% and the mean optimization time by 23%. On Transition1x the effect is small, with both the mean and median number of steps decreasing slightly. On Transition1x-TMC the median decreases while the mean rises, so a minority of difficult cases requires more iterations under CurvFM. The mean optimization time nevertheless decreases on this dataset as well, which indicates that the individual steps are cheaper, as expected for structures that are closer to the physical geometry and for which the SCF converges more readily.

Table 9: Optimization metrics for TS generation comparing post-optimization TS structures. Best in bold, second-best underlined.

		RMSD (Å) ↓		D-MAE (Å) ↓		Opt. Steps		Opt. Time (s)
Dataset	Model	Mean	Median	Mean	Median	Mean	Median	Mean
Transition1x	React-OT	**0.080**	0.000	0.020	0.000	<u>44.63</u>	<u>39.00</u>	<u>1505.50</u>
Transition1x	React-OT$_{\text{CurvFM}}$	<u>0.090</u>	0.000	0.020	0.000	**44.16**	**37.00**	**1500.70**
Transition1x-2p3p4p	React-OT	0.120	0.000	0.060	0.000	<u>55.70</u>	<u>40.00</u>	<u>8.54</u>
Transition1x-2p3p4p	React-OT$_{\text{CurvFM}}$	0.120	0.000	0.060	0.000	**45.98**	**39.50**	**6.54**
Transition1x-TMC	React-OT	<u>0.270</u>	0.000	<u>0.100</u>	0.000	**116.11**	<u>79.00</u>	<u>20.74</u>
Transition1x-TMC	React-OT$_{\text{CurvFM}}$	**0.260**	0.000	**0.090**	0.000	<u>121.37</u>	**78.00**	**20.45**

Accuracy of the endpoint estimate. Finally, we examine the effect of the endpoint estimate itself. Unrolling two ODE steps rather than one brings $\hat{x}_1$ closer to the true endpoint and thereby narrows Jensen's gap, at the cost of a second model call per training step. Table 10 shows that this yields a further improvement of roughly 11% in both the median barrier error and the median maximum force, reducing them to 1.78 kcal/mol and 23.46 kcal/mol/Å, while the global structural metrics again remain unchanged. The improvement comes at a substantial price, with training time rising by 57% to 47 hours and a corresponding increase in memory. We therefore use a single step throughout, and report this variant as an option where maximal physical fidelity outweighs the cost of training.

Table 10: Structural and energetic metrics for TS generation. Best in bold, second-best underlined.

		RMSD (Å) ↓		D-MAE (Å) ↓		ΔE (kcal/mol) ↓		F_{max} (kcal/mol/Å) ↓	
Dataset	Model	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Transition1x-2p3p4p	React-OT	0.196	<u>0.134</u>	0.083	0.060	5.85	2.53	40.96	33.50
Transition1x-2p3p4p	React-OT$_{\text{CurvFM}}$	0.197	**0.133**	<u>0.082</u>	<u>0.058</u>	<u>5.19</u>	<u>2.00</u>	<u>32.69</u>	<u>26.43</u>
Transition1x-2p3p4p	React-OT$_{\text{CurvFM, nfe=2}}$	0.196	0.136	**0.081**	**0.057**	**4.84**	**1.78**	**28.40**	**23.46**

Formulation of the curvature residual. The residual of Eq. 4 constrains the full Hessian, treating all directions of the spectrum alike. Since the type of stationary point is determined by its lowest modes, a residual restricted to that part of the spectrum could in principle deliver the same information at a fraction of the constraint, which is what the subspace variant of Eq. 5 tests. We set $k = 8$ so that the two lowest vibrational modes are covered, since the six rigid-body modes are retained rather than projected out. Following Burger et al. (2025), we avoid the projection because it introduces numerical instabilities in the backward pass, and the same choice applies to the entrywise residual, which is evaluated on the unprojected matrix throughout. Tables 11 and 12 report the three formulations on Transition1x-2p3p4p, together with the purely structural baseline.

The subspace residual improves on the baseline by 6% in both the median barrier error and the median maximum force, and raises the valid TS rate by 1%. The full residual improves the same quantities by 21% and the valid TS rate by 10% The difference is largest among the stiff modes, where the median relative eigenvalue error falls by 7% under the subspace variant and by 27% under the full one, and it is smallest for the reactive mode, whose overlap changes by less than one point under either formulation.

The combination of both terms behaves almost identically to the full residual. The subspace term therefore adds nothing once the full matrix is constrained, and the small consistent degradation suggests that it competes with the denser signal rather than complementing it.

Table 11: Structural and energetic metrics for TS generation. Best in bold, second-best underlined.

		RMSD (Å) ↓		D-MAE (Å) ↓		ΔE (kcal/mol) ↓		$F_{\max}$ (kcal/mol/Å) ↓	
Dataset	Model	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Transition1x-2p3p4p	React-OT	<u>0.196</u>	0.134	0.083	0.060	5.85	2.53	40.96	33.50
Transition1x-2p3p4p	React-OT$_{\text{CurvFM, full}}$	0.197	0.133	0.082	0.058	**5.19**	**2.00**	<u>32.69</u>	**26.43**
Transition1x-2p3p4p	React-OT$_{\text{CurvFM, subspace}}$	**0.194**	0.133	0.082	0.058	5.62	2.38	38.88	31.64
Transition1x-2p3p4p	React-OT$_{\text{CurvFM, combined}}$	0.199	0.141	0.083	0.060	<u>5.39</u>	<u>2.01</u>	**32.60**	<u>26.89</u>

Table 12: Hessian metrics for TS generation. Hessians of both the generated and the ground-truth structures are computed with quantum chemistry at the level of theory of the respective dataset. Soft modes are assigned by $|\lambda_i| < 500$ cm^{-1}. Best in bold, second-best underlined.

		Soft Eigenvalue MRE (%) ↓		Stiff Eigenvalue MRE (%) ↓		Neg. Mode Overlap (%) ↑		Valid TS (%) ↑
Dataset	Model	Mean	Median	Mean	Median	Mean	Median	Mean
Transition1x-2p3p4p	React-OT	25.04	16.93	3.48	2.61	82.42	95.92	57.00
Transition1x-2p3p4p	React-OT$_{\text{CurvFM, full}}$	**21.89**	**13.79**	**2.68**	**1.90**	**85.29**	**96.57**	**67.00**
Transition1x-2p3p4p	React-OT$_{\text{CurvFM, subspace}}$	24.76	15.49	3.24	2.44	83.42	96.12	58.00
Transition1x-2p3p4p	React-OT$_{\text{CurvFM, combined}}$	<u>23.16</u>	<u>14.12</u>	<u>2.73</u>	<u>1.91</u>	<u>84.26</u>	<u>96.26</u>	<u>66.00</u>

Order of the physical signal. To test whether curvature is required or whether a first-order signal would suffice, we replace the Hessian in Eq. 2 by the forces predicted at the two endpoints. Writing $F_\phi : \mathbb{R}^{3N\times 3} \to \mathbb{R}^{3N\times 3}$ for the forces returned by the same frozen surrogate, the residual becomes

$$\mathcal{R}_{\text{force}}(\hat{x}_1, x_1) = \frac{1}{3N} \sum_{i,j} \left| \left[F_\phi(\hat{x}_1)\right]_{ij} - \left[F_\phi(x_1)\right]_{ij} \right|, \tag{24}$$

so that both sides are again evaluated with the surrogate and its systematic error cancels as before. Since forces and curvature differ in units and magnitude, the weight λ and the exponent p are optimized separately for this variant on the validation set, using the same procedure as for the curvature residual. The results on Transition1x-2p3p4p are reported in Table 13. The force residual does improve on the purely structural baseline, reducing the median barrier error and maximum force by 7% each, but the curvature residual reduces both by 21%. The first-order signal thus recovers about a third of the available gain, and does so even for the maximum force, the quantity it supervises directly.

Table 13: Structural and energetic metrics for TS generation. Best in bold, second-best underlined.

		RMSD (Å) ↓		D-MAE (Å) ↓		ΔE (kcal/mol) ↓		$F_{\max}$ (kcal/mol/Å) ↓	
Dataset	Model	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Transition1x-2p3p4p	React-OT	<u>0.196</u>	0.134	0.083	0.060	5.85	2.53	40.96	33.50
Transition1x-2p3p4p	React-OT$_{\text{CurvFM, force}}$	**0.195**	**0.132**	0.082	0.058	<u>5.62</u>	<u>2.35</u>	<u>37.91</u>	<u>31.07</u>
Transition1x-2p3p4p	React-OT$_{\text{CurvFM, Hessian}}$	0.197	<u>0.133</u>	0.082	0.058	**5.19**	**2.00**	**32.69**	**26.43**

B.3 Molecular conformer generation.

Validity and optimization of generated conformers. Curvature is assessed directly by computing Hessians with GFN2-xTB for up to 32 conformers per molecule and testing whether each is a local minimum, that is, whether its spectrum is free of negative eigenvalues. Table 14 reports the resulting fraction of valid conformers per molecule. The reference ensembles of GEOM-QM9 reach 95.7% on average rather than 100%, although they were generated with the same semi-empirical method, which we attribute to numerical precision and differences in code version between the original CREST run and our evaluation. This value sets the practical ceiling against which the generated ensembles are measured. CurvFM raises the median from 96.5 to 100%, so the majority of molecules yield ensembles in which every conformer is a local minimum, while the mean is marginally lower at 87.6 against 89.0%.

Table 15 reports the number of GFN2-xTB optimization steps required to relax each generated conformer to a local minimum. CurvFM reduces this from 5.4 to 4.8 in the median and from 6.1 to 5.8 in the mean. Since every step corresponds to an energy and force evaluation, the reduction

translates directly into a lower cost for any downstream calculation that begins from the generated ensemble.

Table 14: Validity rate of reference and generated conformers as local Hessian minima on GEOM-QM9. Best in bold, second-best underlined.

	Gen. Valid Equilibrium (%) ↑	
Model	Mean	Median
ET-Flow	**89.0**	<u>96.5</u>
ET-Flow$_{\text{CurvFM}}$	<u>87.6</u>	**100.0**

Table 15: Optimization outcome for generated conformers on GEOM-QM9. Best in bold, second-best underlined.

	Opt. Steps	
Model	Mean	Median
ET-Flow	<u>6.1</u>	<u>5.4</u>
ET-Flow$_{\text{CurvFM}}$	**5.8**	**4.8**

Comparison against a stronger structural baseline. Tables 16 and 17 compare ET-Flow trained with CurvFM against DiTMC (Frank et al., 2025), reproduced from the public checkpoint. DiTMC is the stronger generative model, with a lower median AMR in recall and precision. Despite this, the two are comparable on the ensemble properties, where DiTMC leads by 25% in the median energy deviation and 20% in the minimum energy, and the dipole moment is identical. On the maximum force the ordering reverses, with CurvFM reducing the median deviation by 17% and the mean by 22%. Ensemble properties are averages over the generated conformers and therefore reflect how completely the reference ensemble is recovered, whereas the maximum force is a property of each individual structure. The purpose of this comparison is to place the size of the effect in context rather than to rank the two models, which differ in architecture, training data, and budget. Curvature supervision applied to the weaker generative model is sufficient to reach comparable physical fidelity, and to surpass it where the metric is sensitive to individual structures. CurvFM is architecture-agnostic, as the residual is defined on the endpoint estimate and not on the model that produces it, so it can be combined with DiTMC or any other flow matching model.

Table 16: Molecule Molecular conformer generation. results on GEOM-QM9 (δ = 0.5Å). Best in bold, second-best underlined.

	Coverage-R (%) ↑		AMR-R (Å) ↓		Coverage-P (%) ↑		AMR-P (Å) ↓	
Model	Mean	Median	Mean	Median	Mean	Median	Mean	Median
DiTMC+PE(3)-B$_{\text{reproduced}}$	<u>95.02</u>	100.00	**0.075**	**0.023**	**93.29**	100.00	**0.090**	**0.032**
ET-Flow$_{\text{CurvFM}}$	**95.97**	100.00	<u>0.080</u>	<u>0.035</u>	<u>91.33</u>	100.00	<u>0.121</u>	<u>0.054</u>

Table 17: Ensemble property deviations (mean absolute difference vs. reference) on GEOM-QM9, before geometry optimization. Best in bold, second-best underlined.

		$\lvert\Delta\langle E\rangle\rvert$ (kcal/mol) ↓		$\lvert\Delta\langle F_{\max}\rangle\rvert$ (kcal/mol/Å) ↓		$\lvert\Delta\langle\mu\rangle\rvert$ (D) ↓		$\lvert\Delta\langle\Delta\epsilon\rangle\rvert$ (kcal/mol) ↓		$\lvert\Delta\langle E_{min}\rangle\rvert$ (kcal/mol) ↓	
Model	Stage	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
DiTMC+PE(3)-$B_{\text{reproduced}}$	Generated	**0.27**	**0.12**	<u>8.14</u>	<u>6.96</u>	0.15	0.04	1.05	<u>0.57</u>	**0.25**	**0.10**
ET-Flow$_{\text{CurvFM}}$	Generated	<u>0.33</u>	<u>0.15</u>	**6.37**	**5.81**	0.15	0.04	1.05	**0.44**	<u>0.31</u>	<u>0.12</u>