TANGENTMOL: LOCAL TANGENT CONSISTENCY FOR FEW-STEP 3D MOLECULAR GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion and flow-based models generate 3D molecules by gradually denoising random initial structures over many steps, jointly refining atomic coordinates and chemical structure (atom types, bonds, and charges). Reducing the number of steps lowers sampling cost, but each step must then make a large update, and sample quality degrades when the model's predictions change sharply across noise levels. Consistency training mitigates this for continuous data by encouraging agreement between predictions at nearby noise levels, but it does not naturally extend to chemical structure, which changes through discrete jumps rather than along a continuous path. In this study, we introduce TangentMol, a teacher-free training objective that keeps predictions of both coordinates and chemical structure consistent across nearby noise levels. From the noisy molecule already used in standard training, a single Jacobian–vector product computes how each prediction changes along a shared coordinate–time direction. We use the coordinate tangent to form a regression target and the categorical tangents to form soft probability targets for atom, bond, and charge predictions. These terms are added to the standard denoising loss and leave the network architecture and sampler unchanged. Compared with the same-configuration VEDA-FS baseline, TangentMol reduces Fréchet ChemNet Distance from 4.15 to 0.99 on QM9 (8 steps) and from 28.60 to 13.50 on GEOM-Drugs (16 steps), while raising 4-step QM9 connected validity from 53.3% to 79.3%. At 8 steps on GEOM-Drugs, TangentMol surpasses 16-step VEDA, FlowMol2, and FlowMol3 in connected validity, molecule stability, FCD, and strain energy. In pocket-conditioned generation on CrossDocked, it raises 3D validity from 45.3% to 62.8% at 10 steps over PocketVE.

1 INTRODUCTION

Three-dimensional molecular generation jointly models chemical structure and spatial arrangement, producing candidates for virtual screening, property optimization, and structure-based drug design. High-quality samples must be chemically valid, geometrically plausible, and representative of the data distribution. These criteria are distinct: a valid molecular graph can have a strained conformation, while valid molecules with plausible geometries may still fail to capture the diversity of the data. Large-scale applications such as virtual screening must also meet these requirements at low sampling cost.

Equivariant diffusion and flow-matching models, which generate molecules by iteratively refining random noise over a sequence of sampling steps, have shown strong performance in jointly generating 3D atomic coordinates and discrete molecular structure (Hoogeboom et al., 2022; Irwin et al., 2025; Zhang et al., 2026). However, allocating more sampling steps increases computational cost, making generation quality at small sampling budgets particularly important for high-throughput applications. With only a few sampling steps, each prediction of the clean molecule drives a large update to both continuous coordinates and discrete atom, bond, and charge variables. Errors across these outputs can therefore interact: an incorrect bond prediction can conflict with the predicted geometry, while an otherwise valid graph can acquire an implausible conformation. To make these large updates more reliable, we supervise how predictions respond to a shared change in molecular state and noise level.

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.

In continuous domains, consistency models enable few-step sampling by training predictions at different noise levels along the same sampling trajectory to agree (Song et al., 2023; Lu & Song, 2025), while distillation provides another route (Yin et al., 2024). In molecular generation, teacher-based distillation approaches such as FlashMol (Wei et al., 2026) demonstrate competitive few-step generation, but require a pretrained teacher and an additional multi-stage training pipeline. We instead pursue a single-stage, teacher-free objective for few-step molecular generation.

A central challenge for consistency training is the mismatch in output geometry: coordinates are Euclidean, whereas categorical predictive distributions lie on probability simplices. Moreover, the deterministic probability-flow trajectory used in standard continuous consistency training is not directly available for categorical diffusion, in which atom, bond, and charge types change through random discrete jumps (Sahoo et al., 2025). A local tangent objective offers a way forward: at a single noisy state, it can probe a shared coordinate–time direction with the categorical state held fixed, while supervising coordinate and categorical predictions in their respective output spaces.

To address these challenges, we introduce **TangentMol**, which extends the local tangent view of continuous-time consistency training (Lu & Song, 2025) to the joint coordinate and categorical outputs of 3D molecular generators. From the noisy state already sampled for denoising training, an ordinary forward pass predicts coordinates and categorical logits, while a single Jacobian–vector product (JVP) measures their tangents, that is, how both outputs change along the sampled coordinate–time direction. The coordinate tangent yields a directional regression target and the categorical tangents yield soft probability targets for atom, bond, and charge predictions, matched with KL divergence. Both terms are added to the standard denoising loss.

Our contributions are as follows:

1. **Single-state tangent objective.** TangentMol is a teacher-free consistency objective that supervises coordinate and categorical predictions with one joint JVP at the noisy state already used for denoising, requiring no second noisy state or changes to inference.
2. **Modality-specific targets.** We construct a directional regression target for coordinates and probability targets from corrected logits for categorical predictions, and derive local semi-gradients that show how the two updates differ.
3. **Few-step gains with controlled attribution.** TangentMol improves few-step generation on QM9, GEOM-Drugs, and pocket-conditioned CrossDocked, and a same-configuration baseline separates the effect of the tangent objective from that of the training configuration.

2 Related Work

Mixed continuous–categorical molecular generation. Early diffusion models showed that molecular coordinates and atom identities can be denoised jointly with an equivariant network (Hoogeboom et al., 2022). MiDi added bonds and generated the molecular graph together with its 3D coordinates (Vignac et al., 2023), and GeoLDM moved diffusion into a learned latent space (Xu et al., 2023). EQGAT-diff compared design choices for such joint diffusion models (Le et al., 2024). SemlaFlow introduced an efficient equivariant flow-matching architecture with coordinate and categorical outputs (Irwin et al., 2025), and FlowMol3 used self-conditioning, fake atoms, and train-time geometry distortion to address inference-time distribution drift (Dunn & Koes, 2026). VEDA combined variance-exploding (VE) diffusion over coordinates with masked discrete diffusion over chemical categories (Zhang et al., 2026), and we adopt it as the base model for our unconditional experiments. Pocket-conditioned generators must also keep ligands chemically valid and compatible with the protein pocket (Guan et al., 2023; Peng et al., 2022; Schneuing et al., 2024). TAGMol added target-aware guidance to a mixed equivariant generator (Dorna et al., 2024). EvoEGF-Mol coupled geometry and chemical attributes through exponential-family flows for structure-based drug design (Jin et al., 2026). PocketVE applies VE diffusion to structure-based generation with an emphasis on stability and property guidance (Zhang & Bi, 2026) and serves as the base model for our pocket-conditioned experiments.

Few-step consistency. Consistency Models enforce agreement between predictions at different points of a diffusion trajectory, enabling generation with fewer network evaluations (Song et al., 2023). Continuous-time formulations such as sCM study the local tangent limit using Jacobian–

vector-product (JVP) supervision (Lu & Song, 2025), while MeanFlow learns interval-average velocities through a related total-derivative identity (Geng et al., 2025). Other approaches learn finite jumps explicitly, as in Shortcut Models (Frans et al., 2025), or distill a pretrained diffusion teacher, as in DMD (Yin et al., 2024). These ideas have been applied to molecular scaffold hopping, conformation generation, and context-to-molecule generation (Yoo et al., 2024; Fan et al., 2024; Huang et al., 2026), and FlashMol further adapts DMD to few-step 3D molecular generation (Wei et al., 2026). TangentMol instead uses single-state local tangent supervision without solver rollouts, explicit jump-size conditioning, or a pretrained teacher.

Categorical diffusion and consistency. Discrete diffusion models categorical variables through transitions on finite state spaces, including structured, multinomial, and continuous-time formulations (Austin et al., 2021; Hoogeboom et al., 2021; Campbell et al., 2022; 2024). In molecular generation, such categorical dynamics are typically coupled to a separate coordinate process (Vignac et al., 2023). The Diffusion Duality clarifies that the deterministic probability-flow trajectory used by standard continuous consistency methods is not available in the same form for masked categorical diffusion (Sahoo et al., 2025). Recent discrete consistency methods instead use alternative constructions: Consistent Diffusion Language Models supervise distributions over exact posterior bridges (Amin et al., 2026), while concurrent DBTM learns time-independent discrete transport maps with JVP-based corrections (Tang & Wang, 2026). TangentMol retains a time-conditioned molecular denoiser and uses the same local input direction to construct coordinate regression and categorical probability targets. Our categorical objective also builds on the distributional form of dParallel (Chen et al., 2026).

3 Background and Problem Setup

3.1 Mixed molecular state

We represent a molecule with N atoms as

$$\mathcal{M} = (\boldsymbol{x}, \boldsymbol{z}), \qquad \boldsymbol{z} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}), \qquad \boldsymbol{x} \in \mathbb{R}^{N\times 3}, \tag{1}$$

where $\boldsymbol{x}$ contains atomic coordinates and $\boldsymbol{z}$ collects atom-type, pairwise-bond, and charge categories. For target-aware generation, the model additionally conditions on a fixed protein-pocket representation $\mathcal{P}$, and omitting $\mathcal{P}$ gives unconditional generation.

The molecular generator F_θ is $\mathrm{E}(3)$-equivariant on coordinate outputs and $\mathrm{E}(3)$-invariant on categorical logits, meaning that applying the same rotation or translation to the input molecule (and to the pocket $\mathcal{P}$, when present) transforms the predicted coordinates accordingly and leaves the categorical logits unchanged:

$$(D_\theta^x, Z_\theta^z) = F_\theta(\mathcal{M}_t, t, \mathcal{P}), \tag{2}$$

where, given the noisy molecule $\mathcal{M}_t$ at noise level t (Section 3.2), D_θ^x predicts clean coordinates and Z_θ^z collects the categorical logits, with $p_\theta^z = \mathrm{softmax}(Z_\theta^z)$ applied blockwise. This interface follows the mixed equivariant generators used by SemlaFlow and TAGMol (Irwin et al., 2025; Dorna et al., 2024).

3.2 Continuous and categorical corruption

Coordinates follow the variance-exploding corruption process (Karras et al., 2022)

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + t\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \tag{3}$$

on the zero-center-of-mass subspace (Hoogeboom et al., 2022). VEDA adopts this process for 3D molecular generation (Zhang et al., 2026).

The categorical state $\boldsymbol{z}_t$ is sampled from its time-dependent corruption process, and together with $\boldsymbol{x}_t$ forms the noisy molecular input $\mathcal{M}_t$. For unconditional generation, training samples t from a log-normal distribution and corrupts categorical entries according to a time-dependent mask rate, with the time and mask schedules specified in Section A.2. An ordinary denoising training step forms $\mathcal{M}_t$ from a clean molecule $\mathcal{M}_0$ and evaluates $F_\theta(\mathcal{M}_t, t, \mathcal{P})$ once. The base objective is

$$\mathcal{L}_{\mathrm{base}} = \ell_x(D_\theta^x, \boldsymbol{x}_0) + \ell_z(p_\theta^z, \boldsymbol{z}_0), \tag{4}$$

where ℓ_x is the coordinate regression loss and ℓ_z collects the atom, bond, and charge classification losses. At inference, the unconditional base sampler advances coordinates with an Euler step based on the predicted clean coordinates and samples categorical states from a time-dependent transition kernel (Campbell et al., 2022; 2024) (Section A.2). For target-aware generation, the protein pocket is observed and held fixed while the ligand is corrupted and denoised conditional on it. This setting follows the PocketVE corruption process and sampler (Zhang & Bi, 2026). We denote the sampling budget by the number of model evaluations, NFE, and study generation quality at small NFE. For the unconditional samplers, NFE equals the number of sampling steps.

4 TangentMol: Local Tangent Consistency Across Noise Levels

Ordinary denoising supervises the clean molecular prediction at each sampled noisy state using equation 4. TangentMol augments the base denoising objective with local consistency supervision computed at the same noisy state. We first define the shared log-time tangent and then construct modality-specific coordinate and categorical targets.

4.1 Log-time tangent at a single noisy state

Because VE noise levels span several orders of magnitude, we parameterize local tangent supervision by log noise time $s = \log t$. For the VE training path $\boldsymbol{x}_t = \boldsymbol{x}_0 + t\boldsymbol{\epsilon}$, the sampled coordinate-path direction is $\mathbf{u} := \mathrm{d}\boldsymbol{x}_t/\mathrm{d}t = \boldsymbol{\epsilon}$. The chain rule gives the joint coordinate–time direction

$$\left(\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}s}, \frac{\mathrm{d}t}{\mathrm{d}s}\right) = (t\mathbf{u}, t). \tag{5}$$

An ordinary forward pass produces the predictions optimized by the loss. At the same noisy state, one joint JVP of F_θ along $(t\mathbf{u}, t)$ produces reference outputs and head-specific raw tangents:

$$(D^x_{\mathrm{ref}}, Z^z_{\mathrm{ref}}) = F_\theta(\mathcal{M}_t, t, \mathcal{P}), \tag{6}$$

$$(\tau_x, \tau_z) = \mathrm{JVP}_{(\boldsymbol{x}_t, t)}\, F_\theta\left[(t\mathbf{u}, t)\right]. \tag{7}$$

The categorical state $\boldsymbol{z}_t$ and pocket condition $\mathcal{P}$ are held fixed during the JVP. The reference outputs equal the ordinary predictions numerically but are detached, together with the tangents, before constructing the targets, so optimization differentiates only through the ordinary forward pass. All heads share the input direction and JVP computation, and τ_z collects the tangents of the atom, bond, and charge logits.

The raw tangents have a local consistency interpretation. For an output head D, let D_s be its output along the coordinate–time path and $\tau_D(s) = \mathrm{d}D_s/\mathrm{d}s$ its raw log-time tangent. For a small offset $a > 0$ toward lower noise, Taylor expansion gives

$$D_{s-a} = D_s - a\tau_D(s) + O(a^2). \tag{8}$$

This expansion motivates the direction of the target correction. After tangent normalization (Sections 4.2 and 4.3), the targets are directional surrogates rather than exact finite-time predictions. The coordinate and categorical branches construct them in their respective output spaces below. The logit tangent measures the response to the coordinate–time direction with $\boldsymbol{z}_t$ fixed, not a transition of the categorical state.

4.2 Coordinate tangent consistency

Coordinates admit an additive correction directly in their Euclidean output space, so we construct the coordinate target by shifting the clean-coordinate prediction along a stabilized tangent. Let $\mathcal{M}$ denote the valid-atom mask and let $\mathrm{RMS}_{\mathcal{M}}$ be the molecule-wise RMS over valid scalar coordinates. We stabilize the raw coordinate tangent by molecule-wise RMS normalization:

$$\widetilde{\tau}_x = \frac{\tau_x}{\mathrm{RMS}_{\mathcal{M}}(\tau_x) + c_x}, \tag{9}$$

where $c_x > 0$ is a fixed stabilization constant. This preserves the JVP direction while removing the scale of the raw log-time derivative. For the squared coordinate loss, an explicit offset scale is

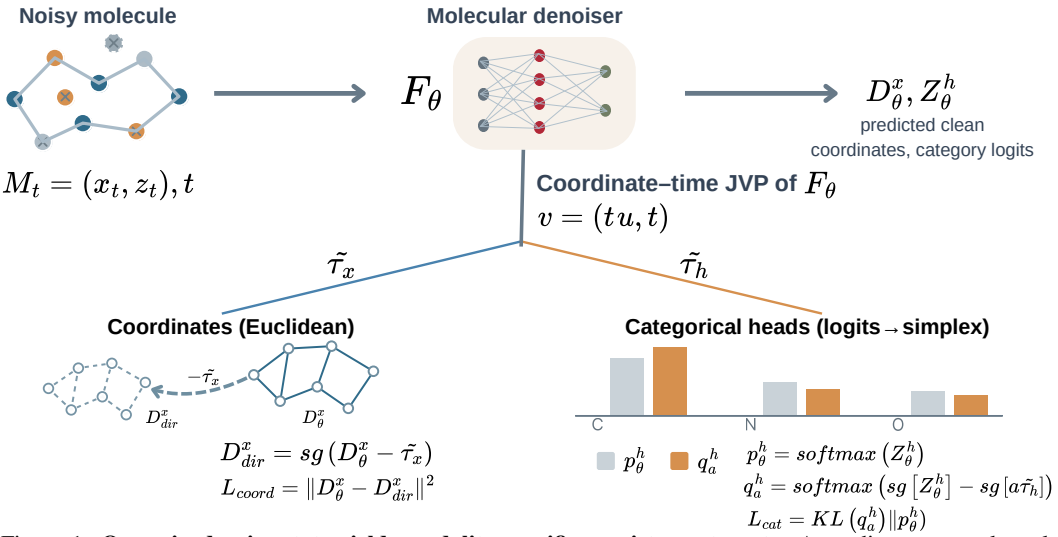


Figure 1: **One mixed noisy state yields modality-specific consistency targets.** An ordinary network evaluation produces coordinate and categorical predictions, while one joint JVP at the same state maps a shared coordinate–time direction to head-specific output tangents. The coordinate tangent defines a local target in coordinate-output space, and each categorical tangent defines a probability target through a logit offset and softmax.

unnecessary because it cancels in the semi-gradient (Section A.5). We therefore use the stabilized tangent directly to define the directional surrogate target

$$D^x_{\text{dir}} = \text{sg}[D^x_{\text{ref}} - \widetilde{\tau}_x]\,, \tag{10}$$

where $\text{sg}[\cdot]$ denotes stop-gradient. The coordinate tangent-consistency loss is

$$\mathcal{L}^x_{\text{tan}} = \|D^x_\theta(\mathcal{M}_t, t, \mathcal{P}) - D^x_{\text{dir}}\|^2_{\mathcal{M}}\,. \tag{11}$$

Its semi-gradient is exactly $2J^{\mathsf{T}}_{\theta,x}\widetilde{\tau}_x$, following the local tangent form of sCM (Lu & Song, 2025).

4.3 Logit tangent consistency for categorical outputs

For categorical predictions, an additive correction in probability space can produce negative entries even if it preserves their sum. We instead correct the logits and apply softmax, ensuring that each target remains a valid probability distribution. Let $\mathcal{H} = \{\text{atom}, \text{bond}, \text{charge}\}$ index the categorical heads. For each head h, let Ω_h denote its valid entries and $K_{h,j}$ the number of classes for entry $j \in \Omega_h$. For each entry, we center the raw logit tangent across classes, $\bar{\tau}_{h,jv} = \tau_{h,jv} - \frac{1}{K_{h,j}}\sum_{w=1}^{K_{h,j}} \tau_{h,jw}$. This removes the component that shifts all logits equally, which leaves the softmax probabilities unchanged but would otherwise distort the subsequent RMS normalization. We then stabilize the centered tangent as

$$\widetilde{\tau}_{h,j} = \frac{\bar{\tau}_{h,j}}{\text{RMS}_v(\bar{\tau}_{h,j}) + c_z}. \tag{12}$$

We apply this normalization separately to the atom, bond, and charge heads, using each head's vocabulary and valid-entry mask.

Unlike the squared coordinate loss, the categorical KL loss yields a semi-gradient that depends on the offset a. We therefore retain $a > 0$ to control the categorical target correction and define

$$q^h_{a,j} = \text{softmax}\big(\text{sg}[Z^h_{\text{ref},j}] - a\,\text{sg}[\widetilde{\tau}_{h,j}]\big)\,. \tag{13}$$

After normalization, a sets the size of this logit offset rather than an exact step in log time. We match the predicted categorical distributions to these targets by aggregating KL losses over valid entries with head-specific weights ρ_h. The factor $1/a$ keeps the semi-gradient of order one as $a \to 0$:

$$\mathcal{L}^z_{\text{tan}} = \sum_{h\in\mathcal{H}} \rho_h \frac{1}{a|\Omega_h|} \sum_{j\in\Omega_h} D_{\text{KL}}\big(q^h_{a,j} \,\big\|\, \text{softmax}(Z^h_{\theta,j})\big)\,. \tag{14}$$

Locally, the two losses yield the following parameter semi-gradients, written per valid entry $j \in \Omega_h$ and with coordinate masking and reductions suppressed:

$$g_x = 2J_{\theta,x}^{\mathsf{T}} \widetilde{\tau}_x, \tag{15}$$

$$g_{h,j} = J_{\theta,h,j}^{\mathsf{T}} C(p_{h,j}) \widetilde{\tau}_{h,j} + O(a), \tag{16}$$

where $p_{h,j} = \mathrm{softmax}(Z_{\theta,j}^h)$, $C(p_{h,j}) = \mathrm{Diag}(p_{h,j}) - p_{h,j}p_{h,j}^{\mathsf{T}}$ is the softmax Fisher matrix, and $J_{\theta,x}$ and $J_{\theta,h,j}$ are the corresponding parameter Jacobians. The coordinate semi-gradient maps the stabilized direction through the parameter Jacobian, whereas the categorical semi-gradient first weights that direction by the local simplex geometry. Section A.5 derives these expressions and explains why we center logit tangents before normalization.

4.4 Anchored TangentMol objective

The core TangentMol objective combines denoising and local consistency:

$$\mathcal{L}_{\mathrm{TangentMol}} = \mathcal{L}_{\mathrm{base}} + \lambda_x \mathcal{L}_{\mathrm{tan}}^x + \lambda_z \mathcal{L}_{\mathrm{tan}}^z, \tag{17}$$

where $\mathcal{L}_{\mathrm{base}}$ anchors the predictions to clean molecular data, preventing trivial locally constant solutions that can satisfy the consistency terms alone. Because the JVP-derived targets are detached, optimization requires no higher-order parameter derivatives. For unconditional models, we additionally add the certainty regularizer of dParallel (Chen et al., 2026) (Section A.7). The joint JVP is used only during training, and inference retains the base model and sampler.

5 Experiments

5.1 Experimental setup

Datasets and backbones. We evaluate unconditional generation on QM9 (Ramakrishnan et al., 2014) and GEOM-Drugs (Axelrod & Gómez-Bombarelli, 2022), and pocket-conditioned generation on CrossDocked2020 (Francoeur et al., 2020). For unconditional generation, TangentMol follows VEDA with the SemlaFlow backbone and VE/annealed denoising parameterization (Irwin et al., 2025; Zhang et al., 2026), and the pocket-conditioned model follows PocketVE with the TAGMol backbone (Dorna et al., 2024; Zhang & Bi, 2026). For unconditional generation, we also construct VEDA-FS, a few-step training configuration of VEDA that uses Muon (Jordan et al., 2024), removes coordinate preconditioning and self-conditioning, and serves as the same-configuration reference for TangentMol (Section 5.4). The additional JVP increases training time by $1.71\times$ on QM9 and $1.43\times$ on GEOM-Drugs on a single A100 (19.8 vs. 11.6 h and 103.5 vs. 72.2 h). Full training settings are given in Section A.1.

Sampling protocol. On QM9, we evaluate 2, 3, 4, 8, and 16 sampling steps with the generalized-arcsin scheduler (Zhang & Bi, 2026) and temperature 0.9. On GEOM-Drugs, we evaluate 4, 8, 16, 30, 50, and 100 steps. At each sampling budget, we generate 10,000 molecules on QM9 and 1,000 on GEOM-Drugs; baseline models use their default sampling settings.

Metrics. We report connected validity, molecule stability, FCD, energy, and strain in the main text, and the appendix additionally reports validity and optimized RMSD. Fréchet ChemNet Distance (FCD; lower is better) measures distributional fidelity to the test-set molecules (Preuer et al., 2018). Validity and stability assess chemical structure, while energy, strain (Halgren, 1996), and optimized RMSD characterize conformational quality. Target-aware evaluation additionally measures pocket compatibility and predicted binding affinity.

5.2 Main low-NFE results

Figure 2 traces quality across sampling budgets, and Table 1 reports matched-budget operating points. Full results are given in Sections A.8 and A.12. On QM9, TangentMol attains lower FCD than VEDA at every shared budget. At 8 steps, it improves all four metrics in Table 1 over every baseline evaluated with our pipeline, and with only 4 steps it still exceeds 8-step VEDA on all four. On GEOM-Drugs at 16 steps, TangentMol improves all four metrics over VEDA, FlowMol2, and

Table 1: Matched-budget operating points on QM9 (8 NFE) and GEOM-Drugs (16 NFE). TangentMol ($\frac{1}{2}$ NFE) uses 4 steps on QM9 and 8 steps on GEOM-Drugs. Bold marks the best matched-budget result among methods evaluated with our pipeline. QM9 TangentMol entries are means over three training runs; standard deviations and all other budgets are given in Tables A3 and A4. A dash denotes an unavailable result.

	QM9, 8 NFE				GEOM-Drugs, 16 NFE			
Method	Conn. valid. ↑	Mol. stab. ↑	FCD ↓	Strain ↓	Conn. valid. ↑	Mol. stab. ↑	FCD ↓	Strain ↓
FlowMol2	72.8	68.3	1.74	44.10	31.3	33.4	29.31	137.14
FlowMol3	–	–	–	–	56.3	55.7	27.17	244.08
VEDA	43.9	36.9	2.98	235.42	31.7	26.5	36.60	265.20
VEDA-FS	92.7	91.6	4.15	19.61	**85.6**	87.2	28.60	39.10
TangentMol	**93.3**	**94.1**	**0.99**	**17.67**	85.4	**92.1**	**13.50**	**32.90**
TangentMol ($\frac{1}{2}$ NFE)	79.3	78.1	1.41	52.86	64.4	67.3	21.70	84.10
Reference, 100 NFE								
SemlaFlow	99.5	99.6	1.50	14.70	92.0	98.1	6.06	65.06
VEDA	99.9	99.9	1.60	8.58	96.8	97.8	28.40	12.60

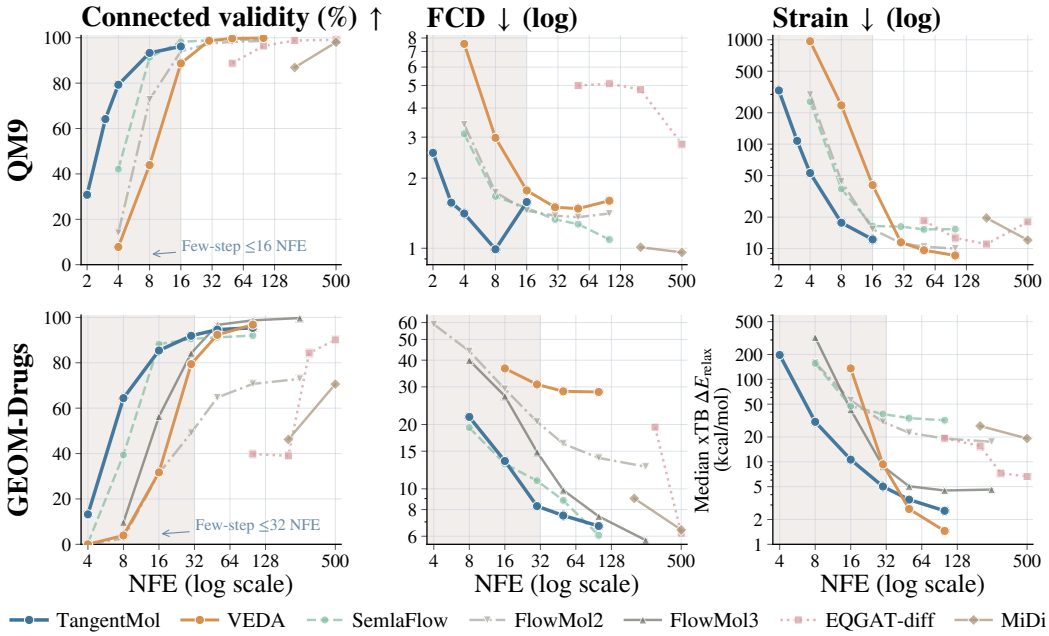


Figure 2: **Quality–efficiency curves across datasets.** The top row shows QM9 and the bottom row shows GEOM-Drugs. Both rows plot connected validity and FCD against NFE, and the right panels show QM9 strain and GEOM-Drugs median GFN2-xTB ΔE_{relax}, respectively. The xTB median is computed for molecules that completed relaxation, and points with fewer than 100 completed relaxations are omitted from that panel.

FlowMol3, and its 8-step results already exceed these three 16-step baselines. The separate GFN2-xTB evaluation (Nikitin et al., 2025) likewise finds a lower median relaxation energy gap at 16 steps (10.63 versus 135.72 for VEDA; Figure 2 and Table A10). Against VEDA-FS, TangentMol mainly improves distributional fidelity, lowering FCD from 4.15 to 0.99 on QM9 and from 28.60 to 13.50 on GEOM-Drugs. On GEOM-Drugs, VEDA-FS already accounts for most of the connected-validity gain over VEDA, and TangentMol further raises molecule stability from 87.2% to 92.1%. Because FlashMol infers bonds after sampling and reports a different validity definition, we list its published results separately in Section A.8.

5.3 Target-aware generalization

We further evaluate TangentMol on pocket-conditioned generation using CrossDocked, with PocketVE as the base target-aware generator (Zhang & Bi, 2026). Both models use the same generalized-arcsin scheduler, and each PocketVE and TangentMol setting generates 100 ligands for each of 100

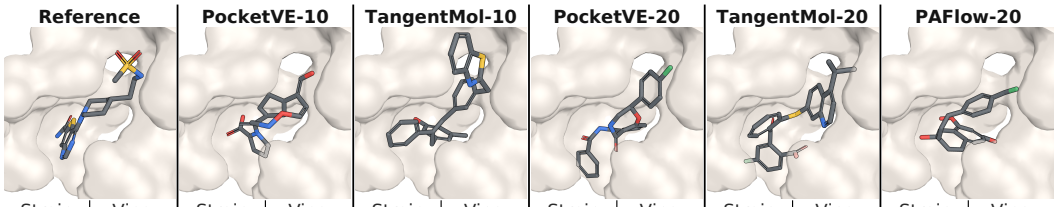


Figure 3: **CrossDocked case study** (PDB 5D7N, chain D). The panels show the reference ligand and examples generated by PocketVE and TangentMol at 10 and 20 steps and by PAFlow at 20 steps in the same pocket. Strain is the GenBench3D MMFF94s strain; Vina is the AutoDock Vina docking score (Vina Dock). Lower values are better for both.

pockets (10,000 requested ligands). Table 2 reports the classifier-free-guidance setting ($\mathrm{CFG} = 5$), together with TargetDiff (Guan et al., 2023), IPDiff (Huang et al., 2024), and PAFlow (Zhou et al., 2025) baselines. The corresponding unguided results are given in Section A.11. At 10 steps, TangentMol raises GenBench3D (Baillif et al., 2024) Validity3D from 45.27% to 62.75% and the clash-free rate from 81.17% to 90.90%, while reducing median strain energy from 1,201.5 to 289.4. These structural gains persist at 20 and 30 steps. It also brings generated ligands closer to the reference distribution: at every matched budget it lowers the local, all-pair, and C–C distance JSDs and the atom-type JS divergence relative to PocketVE, and at 10 steps it is below IPDiff and PAFlow on all four (Table A9). The shared backbone and one network call per step give TangentMol and PocketVE the same theoretical per-step sampling time. Sampling 10,000 ligands with TangentMol takes 27.74 s at 10 steps, versus 1,382.5 s for 1000-step TargetDiff (Zhang & Bi, 2026) and 86.8 s and 203.7 s for 20- and 50-step PAFlow, respectively (Section A.11). Figure 3 shows an illustrative pocket, where TangentMol ligands have lower strain than PocketVE ligands at both 10 and 20 steps and than the 20-step PAFlow ligand.

Table 2: Pocket-conditioned generation on CrossDocked with $\mathrm{CFG} = 5$. Validity3D and clash-free rate are GenBench3D metrics reported as percentages. Strain, AutoDock Vina (Trott & Olson, 2010; Eberhardt et al., 2021) score, and Vina min are medians (lower is better). Bold marks the better value within each matched-step PocketVE/TangentMol pair.

Method	Steps	Validity3D ↑	Clash-free ↑	Strain ↓	QED ↑	SA ↑	Vina score ↓	Vina min ↓
TargetDiff	1000	78.4	86.41	306.0	0.48	0.58	-6.30	-6.83
IPDiff	1000	59.6	91.27	1,248.7	0.53	0.59	-7.01	-7.48
PAFlow	50	45.4	96.38	1,994.0	0.49	0.57	-8.89	-8.97
PAFlow	20	29.3	96.70	3,885.1	0.47	0.57	-8.47	-8.35
PocketVE	10	45.3	81.17	1,201.5	0.52	0.61	-6.29	-6.72
TangentMol	10	**62.8**	**90.90**	**289.4**	**0.58**	**0.66**	**-6.49**	**-6.78**
PocketVE	20	61.8	90.30	388.5	0.59	0.66	**-7.29**	**-7.67**
TangentMol	20	**71.2**	**92.64**	**156.4**	**0.61**	**0.69**	-7.15	-7.48
PocketVE	30	70.0	91.54	234.0	0.60	0.68	**-7.50**	**-7.84**
TangentMol	30	**73.4**	**93.07**	**127.0**	**0.61**	**0.71**	-7.29	-7.57

5.4 What makes TangentMol work?

On both QM9 and GEOM-Drugs, incorrect atom-type and bond predictions coincide with larger geometric errors at intermediate noise levels, supporting joint local supervision of coordinate and categorical predictions (Figure A1). Definitions and uncertainty estimates are provided in Section A.6.

Configuration and objective effects. Table 3 distinguishes the effect of VEDA-FS's training configuration from that of the tangent objectives. On QM9, the configuration changes raise 4-step connected validity from 2.9% to 53.3% and reduce strain from 5540.85 to 128.30. Muon (Jordan et al., 2024) mainly reduces energy and strain, while removing coordinate preconditioning and self-conditioning gives the larger validity gains (Table A5). With VEDA-FS fixed, the coordinate tangent gives the largest FCD reduction on both datasets: from 6.54 to 1.38 on QM9 and from 39.40

Table 3: Training-configuration and objective ablation on QM9 (4 NFE) and GEOM-Drugs (8 NFE). VEDA-FS is VEDA trained with Muon and without coordinate preconditioning or self-conditioning. All rows from VEDA-FS onward share this configuration and sampler within each dataset. TangentMol is VEDA-FS + both tangent terms + certainty regularization. The VEDA rows use the ablation sampler and are not directly comparable to Table 1. Intermediate configuration steps, energy, and larger budgets are given in Section A.9. QM9 TangentMol entries are means over three training runs and match Table 1. A dash marks a variant that was not trained or an FCD that could not be computed.

	QM9, 4 NFE				GEOM-Drugs, 8 NFE			
Configuration / objective	Conn. valid. ↑	Mol. stab. ↑	FCD ↓	Strain ↓	Conn. valid. ↑	Mol. stab. ↑	FCD ↓	Strain ↓
VEDA	2.9	1.4	10.52	5540.85	22.6	17.2	–	480.90
VEDA-FS	53.3	49.1	6.54	128.30	46.5	45.3	39.40	113.60
VEDA-FS + certainty	66.5	62.0	5.96	109.68	–	–	–	–
VEDA-FS + coordinate tangent	74.8	73.6	**1.38**	56.25	47.6	47.2	26.50	123.10
VEDA-FS + logit tangent	63.8	59.4	4.36	115.91	–	–	–	–
VEDA-FS + both tangent terms	78.3	77.9	**1.38**	53.80	52.8	54.3	25.80	100.40
TangentMol	**79.3**	**78.1**	1.41	**52.86**	**64.4**	**67.3**	**21.70**	**84.10**

to 26.50 on GEOM-Drugs. On QM9, it also raises connected validity from 53.3% to 74.8%. Adding the logit tangent raises connected validity to 78.3% and molecule stability from 73.6% to 77.9%. On GEOM-Drugs, the full model further lowers FCD to 21.70 and raises connected validity from 46.5% to 64.4%. At 30 or more steps on GEOM-Drugs, VEDA-FS attains slightly higher validity and lower energy and strain, while TangentMol retains lower FCD (6.70 versus 16.00 at 100 steps; Table A6).

Certainty regularization adds little on top of both tangent terms at 4 steps. At 100 steps on QM9, it raises molecule stability from 94.1% to 98.2% and lowers strain from 16.38 to 9.61; on GEOM-Drugs, it raises molecule stability from 93.1% to 98.0% (Tables A5 and A6). Extra training time alone does not explain the gains: continuing VEDA-FS to 520 epochs, beyond TangentMol's $1.71\times$ training-time cost, raises 4-step connected validity only to 57.2% and lowers FCD only to 6.08. Without tangent normalization, 4-step FCD rises from 1.41 to 2.93, and rescaling to match the average target displacement still gives 2.90 (Section A.10).

6 Limitations and Future Work

TangentMol relies on local network responses, which may be less informative when predictions change sharply across noise levels. The categorical tangent holds the corrupted categorical input fixed and therefore does not define a discrete corruption trajectory. RMS normalization also removes the raw JVP scale, so the virtual targets are directional surrogates rather than calibrated finite-time Taylor predictions. The joint JVP increases training cost, and the stabilization constants and categorical target scale may require retuning across backbones, datasets, or noise schedules.

Our experiments use VE-diffusion denoisers with a log-noise path and clean-coordinate prediction throughout. Extending TangentMol to flow-matching backbones such as FlowMol3 (Dunn & Koes, 2026) requires adapting the shared JVP direction and output targets to their interpolation and categorical transitions. Transfer to protein-structure models also remains open.

7 Conclusion

We introduced TangentMol, a single-state consistency objective that uses one joint log-time JVP to supervise coordinate and categorical predictions of 3D molecular denoisers. On GEOM-Drugs, TangentMol lowers 16-step FCD from 28.60 to 13.50 relative to VEDA-FS, which shares its training configuration and sampler, and 8-step TangentMol surpasses 16-step VEDA, FlowMol2, and FlowMol3 on connected validity, molecule stability, FCD, and strain. The few-step gains also extend to QM9, where 4-step TangentMol exceeds 8-step VEDA on all four metrics, and to pocket-conditioned CrossDocked, where 10-step Validity3D rises from 45.3% to 62.8% over PocketVE. Together, these results show that local tangent supervision improves few-step molecular generation without changing the inference architecture or sampler.

AI USE STATEMENT

We used generative AI tools primarily for manuscript drafting and editing, literature search and summarization, and occasional discussion of research ideas and experimental presentation. Generative AI was also used on a limited basis to assist with code review and debugging. Generative AI was also used to assist in deriving and checking the mathematical derivations in the appendix, including the tangent-offset and categorical semi-gradient analyses (Section A.5). The authors verified every step of these derivations. The core methodology, implementation decisions, experimental configurations, and interpretation of the reported results were developed and determined by the authors. AI-assisted code suggestions were reviewed before use, and all reported numerical results were obtained from experiments conducted by the authors. References and technical claims suggested through generative AI were checked against the original sources. We did not use generative AI to generate training or evaluation datasets, fabricate or modify experimental measurements, or autonomously determine the conclusions of this work. We have reviewed all AI-assisted content and take responsibility for the final manuscript and results.

REPRODUCIBILITY STATEMENT

We provide methodological and experimental details for reproducing TangentMol throughout the paper and appendix. Section 4 specifies the objective and same-state JVP target construction. Appendices A.1–A.4 report dataset-specific training and sampling settings, masking and loss reductions, and the training algorithm, while Appendix A.5 gives the tangent derivations. Section 5.1 reports the evaluation protocols and training-time overhead, and Appendices A.8 and A.11 provide the complete numerical operating points. An anonymized code repository is available at `https://anonymous.4open.science/r/TangentMol-5CA3/`, with training, sampling, and evaluation code and configuration scripts for the unconditional QM9 and GEOM-Drugs experiments.

REFERENCES

Hasan Amin, Yuan Gao, Yaser Souri, Subhojit Som, Ming Yin, Rajiv Khanna, and Xia Song. Consistent diffusion language models. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=64HmKFCaq6`.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, April 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01288-4. URL `https://doi.org/10.1038/s41597-022-01288-4`.

Benoit Baillif, Jason Cole, Patrick McCabe, and Andreas Bender. Benchmarking structure-based three-dimensional molecular generative models using genbench3d: ligand conformation quality matters. *arXiv preprint arXiv:2407.04424*, 2024.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=kQwSbv0BR4`.

Zigeng Chen, Gongfan Fang, Xinyin Ma, Ruonan Yu, and Xinchao Wang. dparallel: Learnable parallel decoding for dLLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=hVOcstAURb`.

Vineeth Dorna, D. Subhalingam, Keshav Kolluru, Shreshth Tuli, Mrityunjay Singh, Saurabh Singal, N. M. Anoop Krishnan, and Sayan Ranu. Tagmol: Target-aware gradient-guided molecule generation. *arXiv preprint arXiv:2406.01650*, 2024.

Ian Dunn and David R Koes. Flowmol3: flow matching for 3d de novo small-molecule generation. *Digital Discovery*, 5(5):2052–2066, 2026.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: new docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891, 2021.

Zhiguang Fan, Yuedong Yang, Mingyuan Xu, and Hongming Chen. Ec-conf: A ultra-fast diffusion model for molecular conformation generation with equivariant consistency. *Journal of Cheminformatics*, 16(1):107, 2024.

Paul G Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B Iovanisci, Ian Snyder, and David R Koes. 3d convolutional neural networks and a crossdocked dataset for structure-based drug design. *Journal of chemical information and modeling*, 60(9):4200, 2020.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OlzB6LnXcS.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uWj4s7rMnR.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kJqXEPXMsE0.

Thomas A Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5-6):490–519, 1996.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Yuhao Huang, Shih-Hsin Wang, Andrea L. Bertozzi, and Bao Wang. RMFlow: Refined mean flow by a noise-injection step for multimodal generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=p072J56yo4.

Zhilin Huang, Ling Yang, Xiangxin Zhou, Zhilong Zhang, Wentao Zhang, Xiawu Zheng, Jie Chen, Yu Wang, Bin Cui, and Wenming Yang. Protein-ligand interaction prior for binding-aware 3d molecule diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Semlaflow – efficient 3d molecular generation with latent attention and equivariant flow matching. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=bee2G6pEh0.

Yaowei Jin, Junjie Wang, Cheng Cao, Penglei Wang, Duo An, and Qian Shi. EvoEGF-mol: Evolving exponential geodesic flow for structure-based drug design. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=bOlF0eLPXa.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tuan Le, Julian Cremer, Frank Noe, Djork-Arné Clevert, and Kristof T Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kzGuiRXZrQ.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LyJi5ugyJx.

Filipp Nikitin, Ian Dunn, David Ryan Koes, and Olexandr Isayev. Geom-drugs revisited: toward more chemically accurate benchmarks for 3d molecule generation. *Digital Discovery*, 4(11): 3282–3291, 2025.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International conference on machine learning*, pp. 17644–17655. PMLR, 2022.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. *Proceedings of machine learning research*, 267: 52584, 2025.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Sophia Tang and Shiyi Wang. Discrete beckmann transport models for one-step language modeling and reasoning. *arXiv preprint arXiv:2609.15903*, 2026.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 560–576. Springer, 2023.

Xinyuan Wei, Zian Li, Shaoheng Yan, Cai Zhou, and Muhan Zhang. Flashmol: High-quality molecule generation in as few as four steps. *arXiv preprint arXiv:2605.07020*, 2026.

Minkai Xu, Alexander S Powers, Ron O Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024.

Kiwoong Yoo, Owen Oertell, Junhyun Lee, Sanghoon Lee, and Jaewoo Kang. Turbohopp: accelerated molecule scaffold hopping with consistency models. *Advances in neural information processing systems*, 37:41157–41185, 2024.

Peining Zhang and Jinbo Bi. Pocketve: Stable and property-guided structure-based drug design with variance-exploding diffusion. *arXiv preprint arXiv:2609.08101*, 2026.

Peining Zhang, Jinbo Bi, and Minghu Song. Veda: Generation of 3d molecules via variance-exploding diffusion with annealing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28346–28354, 2026.

Jingyuan Zhou, Hao Qian, Shikui Tu, and Lei Xu. Prior-guided flow matching for target-aware molecule design with learnable atom number. *arXiv preprint arXiv:2509.01486*, 2025.

A Appendix

A.1 Training hyperparameters

Table A1 summarizes the principal training and TangentMol hyperparameters for the unconditional and pocket-conditioned experiments. Parameters not listed in the table follow the corresponding SemlaFlow- or PocketVE-compatible implementation. The generalized-arcsin parameters listed below describe the TangentMol/PocketVE sampling configurations. We parameterize the categorical mask rate as $m(t) = [(\log t - \log t_{\min})/(\log t_{\max} - \log t_{\min})]^{\alpha}$ for $t \in [t_{\min}, t_{\max}]$.

Table A1: Training hyperparameters for the QM9, GEOM-Drugs, and pocket-conditioned CrossDocked experiments.

Hyperparameter	QM9	GEOM-Drugs	CrossDocked
Training budget	300 epochs	200 epochs	400,000 steps
Optimizer	Muon	Muon	Muon
Learning rate	2×10^{-3}	2×10^{-3}	10^{-3}
Optimizer warm-up steps	2,000	10,000	none
Batch cost	4,096	4,096	-
Batch size	-	-	4
Bond-loss weight	0.5	1.0	0
Mask-rate exponent α	0.5	1.0	1.0
Coordinate consistency weight λ_x	4.0	1.0	0.03
Categorical consistency weight λ_z	1.0	1.0	1.0
Head weights ρ_h (atom, bond, charge)	(0.2,0.5,1.0)	(0.2,1.0,1.0)	(1,0,-)
Time parameterization	log	log	log
Categorical offset a	0.1	0.1	0.1
Coordinate normalization constant c_x	0.1	0.1	0.1
Categorical normalization constant c_z	0.1	0.1	0.1
Tangent warm-up steps (K_x, K_z)	(2,000, 2,000)	(10,000, 10,000)	(10,000, 10,000)
Certainty weight	1.0	0.3	0
Generalized-arcsin coefficient β	2.85	2.85	2.2
Generalized-arcsin exponent p	0.563	0.563	0.57

A.2 Base corruption and sampling process

For the unconditional QM9 and GEOM-Drugs models, training draws the coordinate noise time from $\log t \sim \mathcal{N}(-1.26, 1.41^2)$ and forms $\boldsymbol{x}_t$ using equation 3. At this time, each atom or bond category is retained from the clean molecule or replaced by a prior-sampled category according to the mask rate $m(t)$ defined above; $m(t)$ is clipped to its endpoint values outside $[t_{\min}, t_{\max}]$. The pocket-conditioned experiment retains the PocketVE corruption process (Zhang & Bi, 2026).

For TangentMol's unconditional inference with K sampling steps, the generalized-arcsin scheduler places the noise times on a descending grid; VEDA baseline results use its native scheduler except in the controlled ablation. With $w_i = 1 - i/K$, its grid is

$$u_i = (1-\beta)w_i + \beta\frac{2}{\pi}\arcsin(w_i^p), \qquad t_i = \exp(\log t_{\min} + u_i(\log t_{\max} - \log t_{\min}))\,, \tag{18}$$

where β and p are given in Table A1. For a reverse interval $\Delta_i = t_i - t_{i+1}$ without scheduled noise injection, the coordinate Euler predictor is

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \Delta_i \frac{D_\theta^x(\mathcal{M}_i, t_i, \mathcal{P}) - \boldsymbol{x}_i}{t_i}. \tag{19}$$

When the base sampler injects noise before a step, the same predictor is evaluated at the resulting state and noise time.

The categorical update is stochastic rather than an Euler step on logits. For each current atom or bond category k, the base sampler forms an Euler transition row from a time-dependent jump-rate matrix Q_{θ,t_i}, $\Pi_i(k,\cdot) \simeq \delta_k + \Delta_i Q_{\theta,t_i}(k,\cdot)$, and samples the next category from that row. The rates use the predicted clean-category probabilities, $m(t_i)$, and its derivative. If an Euler interval assigns more than unit mass off the diagonal, the implementation rescales that row before assigning the remaining mass to the current category. These coordinate and categorical updates use the same model evaluation and are unchanged by TangentMol.

A.3 NOTATION AND MASKING CONVENTIONS

For a batch of padded molecules, $r_i = 1$ denotes a real atom and $r_i = 0$ denotes padding. Coordinate losses are averaged over valid atoms and coordinate dimensions. Atom and charge categorical losses are averaged over valid atoms. Bond losses are averaged over valid unordered atom pairs, or over the corresponding symmetric entries when the implementation stores a full adjacency tensor. In every case, padding entries are excluded from both the denoising and TangentMol consistency losses.

The JVP returns each head's output at the evaluation point (its reference prediction) and its raw directional derivative. We use detached copies of these reference quantities to form the consistency targets; the ordinary forward predictions remain attached for the base and consistency losses. EMA parameters, when maintained for sampling, are not used to construct the TangentMol consistency target.

A.4 TANGENTMOL TRAINING ALGORITHM

Algorithm 1 One TangentMol training update

Require: Clean molecule $\mathcal{M}_0$, optional pocket $\mathcal{P}$, model F_θ, optimizer, training step k

1: Sample time t and construct $\mathcal{M}_t$ using the dataset-specific training corruption; let $\mathbf{u} = \mathrm{d}\boldsymbol{x}_t/\mathrm{d}t$ (for VE corruption, $\boldsymbol{x}_t = \boldsymbol{x}_0 + t\boldsymbol{\epsilon}$ and $\mathbf{u} = \boldsymbol{\epsilon}$).
2: Evaluate the ordinary prediction $F_\theta(\mathcal{M}_t, t, \mathcal{P})$ and compute $\mathcal{L}_{\text{base}}$.
3: At the same state, compute the log-time joint JVP along $(t\mathbf{u}, t)$ with respect to $(\boldsymbol{x}_t, t)$, holding $\boldsymbol{z}_t$ and $\mathcal{P}$ fixed, to obtain reference outputs and raw tangents as in equation 7.
4: Detach the JVP reference outputs and tangents; normalize each tangent and apply its warm-up at step k.
5: Form the coordinate and categorical targets in equation 10 and equation 13, then evaluate the masked consistency losses using the ordinary predictions.
6: Combine the losses using equation 44 and equation 43, then update θ.

A.5 OFFSET AND LOG-TIME INTERPRETATION OF TANGENTMOL

This section distinguishes the raw Taylor motivation from the stabilized direction and virtual targets used by the implementation.

Raw tangent and stabilized surrogate. Let $s = \log t$ denote the scalar coordinate in which consistency is enforced. Define

$$D_s := D(\boldsymbol{x}_{t(s)}, t(s)), \qquad \tau_s := \frac{\mathrm{d}D_s}{\mathrm{d}s}, \tag{20}$$

where τ_s is the raw log-time tangent. The first-order expansion for a small backward log-time displacement $a > 0$ is

$$D_{s-a} = D_s - a\tau_s + O(a^2). \tag{21}$$

The actual training target includes both normalization and warm-up. For the coordinate branch, we write these operations explicitly as

$$\widehat{\tau}_s = \frac{\tau_s}{\mathrm{RMS}_{\mathcal{M}}(\tau_s) + c_x}, \qquad \widetilde{\tau}_s = \gamma(k)\widehat{\tau}_s, \qquad \gamma(k) = \min\left(1, \frac{k+1}{K_{\text{warm}}}\right). \tag{22}$$

Here k is the training-step index, and $\widetilde{\tau}_s$ includes the warm-up factor omitted from the main-text normalization formulas. The categorical branch uses the same linear schedule with its own warm-up length, as specified in Table A1. Consider the offset-parameterized coordinate target

$$D_{\text{virt}}(a) \;=\; \mathrm{sg}[D_{\text{ref}} - a\widetilde{\tau}_s]\,, \qquad D_{\text{ref}} = D(\boldsymbol{x}_t, t), \tag{23}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. Unless the normalization factor is exactly one, $D_{\text{virt}}(a)$ is not the strict Taylor approximation to D_{s-a} and does not have the same $O(a^2)$ error guarantee. It is a virtual target guided by a stabilized local direction derived from the raw JVP. The corresponding coordinate residual is

$$r_a = D_\theta(\boldsymbol{x}_t, t) - D_{\text{ref}} + a\widetilde{\tau}_s, \qquad \mathcal{L}_{\text{cont}}(a) = \frac{1}{a}\,\|r_a\|^2_{\mathcal{M}}, \tag{24}$$

where $\|\cdot\|^2_{\mathcal{M}}$ denotes the masked mean used for valid molecular coordinates. The factor $1/a$ preserves the local tangent gradient as the offset changes.

The stop-gradient construction has a semi-gradient identity. Ignoring the masking notation and writing $J_\theta D_\theta$ for the parameter Jacobian, the attached output D_θ and detached reference output D_{ref} have identical values at the same state, so $r_a = a\widetilde{\tau}_s$. Consequently,

$$\nabla_\theta \mathcal{L}_{\mathrm{cont}}(a)|_{D_\theta = D_{\mathrm{ref}}} = 2(J_\theta D)^{\mathsf{T}} \widetilde{\tau}_s, \tag{25}$$

up to the chosen reduction. Because the reference target and tangent are detached, this is a semi-gradient rather than a full gradient penalty on the tangent norm. The offset-free coordinate objective used in the main text,

$$\|D_\theta - \mathrm{sg}[D_{\mathrm{ref}} - \widetilde{\tau}_s]\|^2_{\mathcal{M}}\,, \tag{26}$$

has the same semi-gradient exactly. Thus, a is redundant for the coordinate branch when the attached prediction and detached reference are evaluated at the same state, and the main formulation removes it.

Frozen-discrete logit tangent. For the categorical output collection, the categorical input $\boldsymbol{z}_t$ is fixed during the JVP. The path is therefore

$$G_z(s; \boldsymbol{z}_t) = Z^z_\theta(\boldsymbol{x}_{t(s)}, \boldsymbol{z}_t, t(s), \mathcal{P}), \qquad \tau_z = \frac{\mathrm{d}G_z(s; \boldsymbol{z}_t)}{\mathrm{d}s}. \tag{27}$$

This tangent describes how the categorical logits respond to the shared coordinate–time direction while the corrupted categorical context remains unchanged; it is not a derivative through a discrete jump trajectory. For each categorical head $h \in \mathcal{H}$ and each valid entry $j \in \Omega_h$, the implementation preserves the simplex by applying the virtual displacement in logit space and then normalizing. In particular,

$$\begin{aligned} q^h_{a,j} &= \mathrm{softmax}\big(\mathrm{sg}[Z^h_{\mathrm{ref},j}] - a\,\mathrm{sg}[\widetilde{\tau}_{h,j}]\big)\,, \qquad j \in \Omega_h, \\ \mathcal{L}_{\mathrm{disc}}(a) &= \sum_{h\in\mathcal{H}} \rho_h \frac{1}{a|\Omega_h|} \sum_{j\in\Omega_h} \mathrm{KL}\big(q^h_{a,j} \,\big\|\, \mathrm{softmax}(Z^h_{\theta,j})\big)\,. \end{aligned} \tag{28}$$

We apply this construction blockwise to atom, bond, and charge logits, with valid-atom or valid-pair masks and categorical head weights ρ_h. The tangent is centered over the class dimension before RMS normalization, which removes the class-constant logit direction that has no effect on the softmax. The reference logits and tangent are detached, whereas Z^h_θ remains attached to the computation graph. With the detached target distribution, the per-entry semi-gradient is

$$\frac{\partial \mathcal{L}_{\mathrm{disc},h,j}}{\partial Z^h_{\theta,j}} = \frac{p^h_{\theta,j} - q^h_{a,j}}{a}. \tag{29}$$

Local categorical geometry and tangent centering. Consider one valid entry with K classes, and write z for its current logits, $p = \mathrm{softmax}(z)$, and v for its detached stabilized tangent. For $q_a = \mathrm{softmax}(z - av)$, the target uses detached copies of both z and v. At fixed z and v, smoothness of softmax gives

$$q_a = p - aC(p)v + O(a^2), \qquad C(p) = \mathrm{Diag}(p) - pp^{\mathsf{T}}. \tag{30}$$

The categorical score for class y is $\nabla_z \log p_y = e_y - p$, so its second moment is

$$\mathbb{E}_{y\sim p}[(e_y - p)(e_y - p)^{\mathsf{T}}] = C(p). \tag{31}$$

This identifies $C(p)$ as the Fisher information matrix in logit coordinates. If $A(z) = \log \sum_k \exp(z_k)$, then $\nabla A(z) = p$ and $\nabla^2 A(z) = C(p)$. Consequently,

$$\mathrm{KL}(q_a \| p) = -aq_a^{\mathsf{T}} v - A(z - av) + A(z) = \frac{a^2}{2} v^{\mathsf{T}} C(p) v + O(a^3). \tag{32}$$

This expansion describes the local value of the divergence. During training, the target is detached, giving the logit semi-gradient

$$\nabla_z \left[\frac{1}{a}\,\mathrm{KL}(\mathrm{sg}[q_a] \| \mathrm{softmax}(z))\right] = \frac{p - q_a}{a} = C(p)v + O(a). \tag{33}$$

Thus, differentiating the quadratic expression in equation 32 as a function of the shared logits would not reproduce the training update: that differentiation would also account for the target's dependence on the logits. The update uses the Fisher matrix as a local weighting of the tangent, without applying the inverse Fisher matrix required for a natural-gradient step.

Centering removes a redundant logit direction before its scale enters the normalization. Let $\mathbf{1} \in \mathbb{R}^K$ and $P = I - \mathbf{1}\mathbf{1}^\mathsf{T}/K$. Softmax shift invariance implies

$$C(p)\mathbf{1} = 0, \qquad \mathrm{softmax}(z - a\tau) = \mathrm{softmax}(z - aP\tau). \tag{34}$$

Hence, centering alone leaves the unnormalized probability target unchanged. However, writing $\tau = P\tau + \bar{\tau}\mathbf{1}$ gives

$$\mathrm{RMS}(\tau)^2 = \mathrm{RMS}(P\tau)^2 + \bar{\tau}^2, \qquad \bar{\tau} = \frac{1}{K}\mathbf{1}^\mathsf{T}\tau. \tag{35}$$

Normalizing the raw tangent would therefore let the class-constant component change the magnitude of the probability correction, even though that component has no effect on softmax. Using $P\tau/(\mathrm{RMS}(P\tau) + c_z)$ makes the normalized direction invariant to adding any class-constant component to τ. This justifies the order of centering and normalization in equation 12.

JVP evaluation. We compute one Jacobian-vector product of the network at the same $(\boldsymbol{x}_t, t)$ state. Its reference output supplies D_{ref} (and the categorical reference logits), while its tangent is taken with respect to $s = \log t$. Since $\boldsymbol{x}_t = \boldsymbol{x}_0 + t\mathbf{u}$ for the continuous path (with $\mathbf{u} = \boldsymbol{\epsilon}$ under VE noise),

$$\tau_s = J_x D\,\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}s} + \partial_t D\,\frac{\mathrm{d}t}{\mathrm{d}s} = t\left(J_x D\,\mathbf{u} + \partial_t D\right), \qquad \left(\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}s}, \frac{\mathrm{d}t}{\mathrm{d}s}\right) = (t\mathbf{u}, t). \tag{36}$$

The primal and tangent are detached before constructing the training losses. Consequently, this target construction requires no second noisy state, finite solver step, or solver rollout.

Holding the sampled categorical state $\boldsymbol{z}_t$ fixed, define the conditional mean interpolant direction as

$$\mathbf{v}_t(\boldsymbol{x}; \boldsymbol{z}_t) := \mathbb{E}[\mathbf{u} \mid \boldsymbol{x}_t = \boldsymbol{x}, \boldsymbol{z}_t]\,. \tag{37}$$

Because the raw directional derivative is linear in $\mathbf{u}$, the corresponding conditional mean tangent is

$$\mathbb{E}[\tau_s \mid \boldsymbol{x}_t = \boldsymbol{x}, \boldsymbol{z}_t] = t\left(J_x D_\theta(\boldsymbol{x}, \boldsymbol{z}_t, t)\,\mathbf{v}_t(\boldsymbol{x}; \boldsymbol{z}_t) + \partial_t D_\theta(\boldsymbol{x}, \boldsymbol{z}_t, t)\right). \tag{38}$$

The sampled path direction is therefore distinct from a trajectory obtained by integrating the current model: the raw single-sample JVP is a stochastic local direction whose conditional mean follows the interpolant direction associated with the fixed categorical state.

Under this log-time parameterization, the virtual time associated with a backward log-time offset is

$$\tau_s = \frac{\mathrm{d}D}{\mathrm{d}\log t} = t\frac{\mathrm{d}D}{\mathrm{d}t}, \qquad t_{\mathrm{virt}} = \exp(s - a) = t\exp(-a). \tag{39}$$

Thus, before tangent normalization, a measures a relative noise-scale change: for example, $a = 0.1$ corresponds to the illustrative virtual scale factor $\exp(-0.1) \approx 0.905$, independently of the absolute value of t.

Categorical offset. Both consistency branches use log-time parameterization, but only the categorical branch retains a finite offset magnitude. The network is not conditioned on a, so a defines the categorical training target without changing the inference interface $D_\theta(\boldsymbol{x}, t)$.

A.6 Geometric error conditioned on categorical correctness

We examine the association between categorical prediction errors and geometric reconstruction errors at individual noisy states. We evaluate the QM9 and GEOM-Drugs VEDA models on 10,000 distinct validation molecules each, using non-EMA weights, with self-conditioning disabled and each model's categorical noise schedule. We evaluate 14 noise levels, $t \in \{0.03, 0.04, 0.055, 0.075, 0.1, 0.14, 0.2, 0.3, 0.45, 0.65, 1, 1.5, 2.2, 3\}$. The random seed is reset at each noise level to reuse the sampled priors across evaluations.

For atom-type and charge predictions, the geometric error of atom i is $e_i = \|\widehat{\boldsymbol{x}}_i - \boldsymbol{x}_{0,i}\|_2$, measured in Å. For a bond prediction on pair (i, j), it is the absolute distance error

$$e_{ij} = \left| \|\widehat{\boldsymbol{x}}_i - \widehat{\boldsymbol{x}}_j\|_2 - \|\boldsymbol{x}_{0,i} - \boldsymbol{x}_{0,j}\|_2 \right| . \tag{40}$$

Predicted categorical classes are obtained by argmax and compared with the clean labels. The true-bond analysis includes only unique pairs $i < j$ with a nonzero clean bond class, excluding nonbonded pairs. For each categorical head h, with predicted class $\widehat{y}$ and clean class y_0, we compute

$$\Delta_h(t) = \mathbb{E}[e \mid \widehat{y} \neq y_0, t] - \mathbb{E}[e \mid \widehat{y} = y_0, t]. \tag{41}$$

The empirical means pool valid entries within each correctness group. We estimate percentile 95% confidence intervals using 1,000 paired molecule-level bootstrap replicates, resampling whole molecules and recomputing both group means with the same resampled molecules. This preserves dependence among atoms and bonds within each molecule.

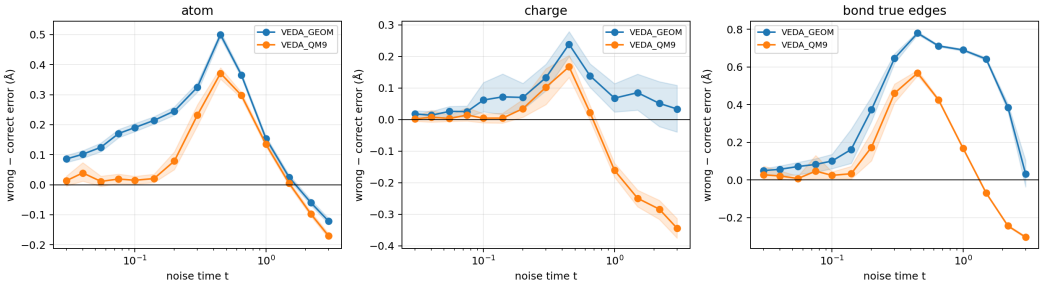


Figure A1: Geometric error conditioned on categorical correctness for VEDA on QM9 and GEOM-Drugs. The panels show the wrong-minus-correct error gap for atom types, charges, and true bonds over the full noise grid; shaded bands denote paired molecule-bootstrap 95% confidence intervals. Positive values indicate larger geometric errors among incorrectly classified entries. Each curve is a within-model diagnostic; absolute gap magnitudes do not rank model quality across the two datasets.

Table A2: Geometric error gaps conditioned on categorical correctness on QM9. Positive Δ indicates greater geometric error for incorrectly predicted categories. All values and molecule-bootstrap 95% confidence intervals are in Å.

	Atom type		True bond	
t	Δ	95% CI	Δ	95% CI
0.20	0.079	[0.051, 0.110]	0.173	[0.102, 0.253]
0.30	0.232	[0.196, 0.267]	0.461	[0.410, 0.510]
0.45	0.371	[0.352, 0.390]	0.568	[0.554, 0.583]
0.65	0.298	[0.288, 0.307]	0.425	[0.419, 0.432]
1.00	0.135	[0.127, 0.143]	0.168	[0.164, 0.173]

Both datasets show a positive intermediate-noise association between atom-type errors and coordinate errors (Figure A1). At $t = 0.45$, the atom gap is 0.371 Å (95% CI $[0.352, 0.390]$) on QM9 and 0.500 Å ($[0.490, 0.510]$) on GEOM-Drugs. The corresponding true-bond gaps are 0.568 Å ($[0.554, 0.583]$) and 0.780 Å ($[0.772, 0.788]$). Table A2 gives additional QM9 values.

The atom gap approaches zero near $t = 1.5$ and becomes negative by $t = 2.2$ in both datasets. The true-bond association persists to higher noise levels on GEOM-Drugs: at $t = 1.5$, its gap remains 0.642 Å ($[0.635, 0.650]$), whereas the QM9 gap is -0.070 Å. The GEOM-Drugs true-bond gap approaches zero only around $t = 3$. Charge also differs between datasets: the QM9 gap changes sign above approximately $t = 0.65$, while the GEOM-Drugs estimate remains weakly positive through most of the grid, with intervals including zero at $t = 2.2$ and $t = 3$. Rare incorrect predictions at low noise levels produce wider intervals, particularly on QM9.

These within-model diagnostics support joint supervision of geometry and categorical predictions, with the associated noise range depending on the modality and dataset. They measure association rather than a causal effect of categorical errors or tangent consistency.

A.7 CERTAINTY REGULARIZATION

The certainty regularizer used in our experiments follows the correctness–certainty separation in dParallel (Chen et al., 2026). For a valid categorical entry $j \in \Omega_h$ of head $h \in \mathcal{H}$, let $y_{h,j}$ be the clean class and $p^h_j = \mathrm{softmax}(Z^h_{\theta,j})$ be the ordinary forward-pass prediction. We define a correctness gate and predictive entropy as

$$g_{h,j} = \mathbf{1}\left[\arg\max_v p^h_{jv} = y_{h,j}\right], \qquad H_{h,j} = -\sum_v p^h_{jv} \log p^h_{jv}. \tag{42}$$

The auxiliary loss is

$$\mathcal{L}_{\mathrm{cert}} = \sum_{h\in\mathcal{H}} \mathrm{Reduce}_{j\in\Omega_h}\left(\mathrm{sg}[g_{h,j}]H_{h,j}\right), \tag{43}$$

where the configured valid-entry reduction is applied separately to the atom, bond, and charge heads before aggregation. Padding entries are excluded, and the native softmax temperature is used. This term encourages confidence only for correct predictions. The training loss used in the reported experiments is

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{TangentMol}} + \lambda_{\mathrm{cert}}\mathcal{L}_{\mathrm{cert}}, \tag{44}$$

with the dataset-specific weight given in Table A1. It uses clean molecular labels and same-state logits and does not participate in the JVP target construction.

A.8 QM9 AND GEOM-DRUGS NUMERICAL RESULTS

Tables A3 and A4 report the full numerical results summarized by the quality–efficiency curves in Figure 2. Rows marked † are literature-reported operating points from FlashMol (Wei et al., 2026), EDM (Hoogeboom et al., 2022), and GeoLDM (Xu et al., 2023). For these external baselines, we only populate NFE, validity, and molecule stability when reported; a dash denotes an unavailable metric. Their original evaluation protocols are retained and are not recomputed with the evaluator used for the other rows.

Comparison with FlashMol. We do not place FlashMol (Wei et al., 2026) in the matched-budget comparison of Table 1 because its reported numbers come from a different model family, training pipeline, and evaluation protocol. We checked each point below against the FlashMol paper and its released code.[1]

Model and training pipeline. FlashMol distills a pretrained GeoLDM teacher (Xu et al., 2023), a latent-space variant of EDM (Hoogeboom et al., 2022) trained with variance-preserving ϵ-prediction and a polynomial noise schedule (`polynomial_2`). The student is initialized from the teacher and trained jointly with a fake-score network and a discriminator, using distribution-matching, adversarial, and Jensen–Shannon losses. Its step counts therefore describe a distilled sampler that depends on a separately pretrained multi-step model. TangentMol instead trains a single network from scratch in one stage and samples with the unchanged base sampler.

Bond-implicit generation. Like EDM and GeoLDM, FlashMol generates atom types and coordinates only. Bonds are not predicted but inferred after sampling from pairwise distances with a lookup table (`get_bond_order` in `qm9/bond_analyze.py`). Its molecule stability therefore checks whether interatomic distances fall into the lookup windows that yield correct valences. TangentMol and the other rows evaluated with our pipeline predict bond types and formal charges directly, and their stability and validity are computed on the predicted molecular graph. The two stability numbers measure different quantities and are not directly comparable.

Validity definition. The released FlashMol evaluation code, inherited unchanged from EDM and GeoLDM, defines validity more permissively than our connected validity. It converts each molecule to SMILES after keeping only its largest fragment (`qm9/rdkit_functions.py`), so disconnected molecules can still count as valid. On GEOM-Drugs, it also restricts every inferred bond to a single bond before RDKit sanitization (`limit_bonds_to_one=True`), and RDKit fills any remaining valence with implicit hydrogens. Under this definition, a molecule fails only if some atom exceeds

[1] `https://github.com/MuLabPKU/FlashMol`

its maximum valence. This is consistent with the reported GEOM-Drugs validity of FlashMol staying above 99.6% at 4, 5, and 8 steps, while molecule stability is not reported for that dataset. The EDM and GeoLDM rows marked † follow the same protocol.

Reported metrics. FlashMol reports atom stability, molecule stability, validity, and valid-and-unique rate, but not connected validity, FCD, energy, or strain. The metrics on which we base our main claims are therefore unavailable for it. FlashMol itself leaves bond-explicit generators such as SemlaFlow out of its comparison for the same reason, arguing that directly generated bonds make stability and validity easy to saturate. We therefore list its reported values only as literature operating points and do not compare them with the values computed by our evaluator. A controlled comparison would require a bond-explicit FlashMol student distilled from the same base model and evaluated with our pipeline, which we leave to future work.

A.9 Full ablation results

Tables A5 and A6 extend the ablation in Table 3. Table A5 lists the intermediate QM9 configuration steps from VEDA to VEDA-FS together with energy at 4, 8, 16, and 100 steps. On QM9, the tangent terms give the largest few-step gains, and certainty regularization mainly acts at larger budgets: without it, molecule stability of VEDA-FS + both tangent terms falls from 95.5% at 16 steps to 94.1% at 100 steps, whereas the full model reaches 98.2%. Table A6 reports the GEOM-Drugs objective ablation at all sampling budgets; all its rows share the VEDA-FS training configuration, sampler, and evaluator used for Table A4. On GEOM-Drugs, the coordinate tangent gives the largest FCD reduction at every budget where FCD is reported. From 16 steps on, certainty regularization slightly raises FCD (12.20 to 13.50 at 16 steps) but raises connected validity from 76.8% to 85.4% and molecule stability from 84.5% to 92.1%. At 30 steps and above, VEDA-FS has higher validity and lower energy and strain than TangentMol, while TangentMol retains lower FCD (6.70 versus 16.00 at 100 steps). The tangent terms and certainty regularization improve validity and stability most clearly at 4 and 8 steps. From 30 steps on, VEDA-FS attains slightly higher validity and lower energy and strain.

A.10 Compute-matched VEDA-FS and tangent normalization

Table A7 reports two additional QM9 controls, evaluated with the sampler and evaluator of Table A5. The compute-matched VEDA-FS continues the 300-epoch VEDA-FS run to 520 epochs with the same constant learning rate, which is $1.73\times$ the original training epochs and exceeds the $1.71\times$ training-time overhead of TangentMol (Section 5.1). The additional training changes VEDA-FS only slightly: at 4 steps, connected validity rises from 53.3% to 57.2% and FCD falls from 6.54 to 6.08, compared with 79.3% and 1.41 for TangentMol. The few-step gains of TangentMol therefore do not come from its additional training compute.

The two no-normalization variants remove the RMS normalization of the coordinate and categorical tangents (Equations (9) and (12)) and keep every other TangentMol setting. Without normalization, the raw tangent scale also changes the average target displacement, making it smaller for coordinates and larger for categorical logits. The rescaled variant corrects this by rescaling the raw coordinate tangent and the categorical offset a so that the average target displacement of each branch matches the default setting, calibrated on the tangent RMS of a trained TangentMol model. Both variants improve over VEDA-FS at 4 steps but fall short of TangentMol: connected validity is 73.5% and 69.9%, FCD is 2.93 and 2.90, and strain is 79.01 and 99.46, compared with 79.3%, 1.41, and 52.86. Because matching the average scale does not recover the default results, the benefit comes from normalizing each molecule's tangent rather than from the overall target magnitude. At 16 steps, both variants reach lower FCD than TangentMol (1.11 and 1.24 versus 1.57) but lower connected validity (94.7% and 92.7% versus 96.1%).

Table A3: QM9 generation results. The SemlaFlow and VEDA blocks provide reference operating points, followed by the TangentMol results. The MiDi, FlowMol2, EQGAT-diff, and SemlaFlow rows are computed with the same evaluator and metric definitions as the TangentMol rows; a dash denotes an unavailable metric. TangentMol entries report means ± standard deviations across three independent training runs with different random seeds; standard deviations for percentage-valued metrics are in percentage points. Connected validity requires a valid molecular graph to form a single connected component. Optimized RMSD is computed on valid molecules, so its evaluated subset varies with connected validity across sampling budgets. Rows marked † are literature-reported and use the EDM validity definition (distance-inferred bonds, largest fragment only), so their Valid column is not comparable to the other rows.

Method	NFE	Connected validity ↑	Valid ↑	Mol. stability ↑	FCD ↓	Opt. RMSD ↓	Energy ↓	Strain ↓
FlashMol†	8	–	96.97	94.87	–	–	–	–
FlashMol†	5	–	92.67	87.05	–	–	–	–
FlashMol†	4	–	90.70	83.89	–	–	–	–
EDM†	1000	–	91.9	82.0	–	–	–	–
GeoLDM†	1000	–	93.8	89.4	–	–	–	–
MiDi	200	86.9	88.8	82.2	1.01	0.331	41.63	19.64
MiDi	500	98.0	98.0	97.6	0.96	0.309	34.38	12.04
FlowMol2	4	14.0	29.2	7.8	3.40	0.367	312.63	297.64
FlowMol2	8	72.8	78.9	68.3	1.74	0.413	67.54	44.10
FlowMol2	16	94.0	95.0	90.2	1.45	0.324	41.35	15.30
FlowMol2	30	97.4	97.7	92.9	1.38	0.279	38.12	11.25
FlowMol2	50	98.2	98.4	92.8	1.36	0.268	36.91	10.41
FlowMol2	100	98.6	98.7	93.3	1.41	0.262	36.62	10.00
EQGAT-diff	50	88.8	90.1	85.6	5.00	0.222	56.75	18.47
EQGAT-diff	100	96.4	96.7	95.4	5.09	0.197	51.18	12.62
EQGAT-diff	200	98.8	98.8	98.5	4.80	0.185	47.59	11.04
EQGAT-diff	500	99.1	99.1	98.9	2.80	0.345	39.56	18.00
SemlaFlow	100	99.5	99.5	99.6	1.50	0.231	39.38	14.70
SemlaFlow	50	99.7	99.7	99.7	1.64	0.224	39.47	14.99
SemlaFlow	30	98.6	98.6	99.4	1.80	0.220	41.95	15.23
SemlaFlow	20	98.5	98.5	99.3	1.83	0.225	44.55	16.15
SemlaFlow	10	96.8	97.1	96.8	2.02	0.249	48.63	23.50
VEDA	100	99.9	99.9	99.9	1.60	0.209	28.44	8.58
VEDA	50	99.7	99.7	99.8	1.48	0.227	31.08	9.63
VEDA	30	98.7	98.7	98.8	1.50	0.253	35.08	11.46
VEDA	16	88.7	90.5	87.2	1.77	0.348	65.89	40.51
VEDA	8	43.9	56.8	36.9	2.98	0.444	261.38	235.42
VEDA	4	7.8	29.8	4.6	7.55	0.260	987.58	969.70
TangentMol	16	96.17 ±0.25	96.27 ±0.25	96.57 ±0.15	1.58 ±0.17	0.3660 ±0.0101	26.58 ±0.79	12.22 ±0.26
TangentMol	8	93.33 ±0.31	93.73 ±0.31	94.13 ±0.32	0.99 ±0.02	0.3463 ±0.0045	38.09 ±0.65	17.67 ±0.09
TangentMol	4	79.30 ±1.25	82.40 ±1.39	78.13 ±1.32	1.41 ±0.05	0.3787 ±0.0081	78.11 ±0.71	52.86 ±0.83
TangentMol	3	64.17 ±0.90	70.97 ±0.86	61.40 ±0.78	1.57 ±0.05	0.4147 ±0.0091	131.43 ±5.40	107.60 ±5.53
TangentMol	2	30.80 ±1.08	46.40 ±0.20	26.43 ±0.97	2.57 ±0.22	0.4210 ±0.0090	347.56 ±13.48	327.14 ±14.06

Table A4: GEOM-Drugs results. The MiDi results follow the Semla-style comparison in the VEDA paper, and all other rows without † are reported using the same evaluator and metric implementation. Rows marked † are literature-reported and use the EDM validity definition, which keeps only the largest fragment and treats every distance-inferred bond as a single bond, so their Valid column is not comparable to the other rows. Lower is better for optimized RMSD, energy, strain, and FCD; percentages are higher-is-better. Optimized RMSD is computed on valid molecules, so its evaluated subset varies with connected validity across sampling budgets. A dash denotes a metric unavailable for that method.

Method	NFE	Connected validity ↑	Valid ↑	Mol. stability ↑	FCD ↓	Opt. RMSD ↓	Energy ↓	Strain ↓
FlashMol†	8	–	99.86	–	–	–	–	–
FlashMol†	5	–	99.69	–	–	–	–	–
FlashMol†	4	–	99.72	–	–	–	–	–
EDM†	1000	–	92.6	–	–	–	–	–
GeoLDM†	1000	–	99.3	–	–	–	–	–
MiDi	200	46.2	54.5	25.4	9.02	1.123	100.48	65.91
MiDi	500	70.6	77.9	40.5	6.43	1.217	90.91	46.64
EQGAT-diff	50	32.1	35.3	30.5	–	0.679	153.56	95.07
EQGAT-diff	100	39.8	42.1	39.7	–	0.704	132.13	64.76
EQGAT-diff	200	39.0	40.4	40.2	–	0.460	120.04	48.55
EQGAT-diff	300	84.3	86.1	85.4	19.44	0.822	90.04	32.09
EQGAT-diff	500	90.1	93.5	93.0	6.20	1.096	68.51	29.35
FlowMol2	4	0.0	0.5	0.0	58.87	0.000	–	–
FlowMol2	8	2.6	12.2	3.5	44.09	0.699	270.13	290.00
FlowMol2	16	31.3	54.2	33.4	29.31	1.197	150.97	137.14
FlowMol2	30	49.1	67.4	44.9	20.61	1.217	95.34	78.38
FlowMol2	50	64.7	81.8	55.0	16.26	1.230	86.61	60.42
FlowMol2	100	70.7	86.8	58.9	13.95	1.230	80.60	48.42
FlowMol2	250	72.9	89.6	58.8	12.70	1.230	74.76	42.56
FlowMol3	8	9.7	24.2	9.3	39.91	0.617	778.85	742.34
FlowMol3	16	56.3	65.5	55.7	27.17	0.935	280.03	244.08
FlowMol3	30	84.2	89.9	88.1	14.87	0.922	95.28	71.39
FlowMol3	50	96.7	99.0	98.8	9.86	0.847	57.03	34.37
FlowMol3	100	98.7	99.9	99.9	7.43	0.804	42.00	19.71
FlowMol3	250	99.7	99.9	99.8	5.75	0.824	39.49	16.29
SemlaFlow	100	92.0	95.0	98.1	6.06	0.844	104.07	65.06
VEDA	100	96.8	98.5	97.8	28.40	0.550	41.70	12.60
VEDA	50	92.3	96.8	95.2	28.60	0.650	52.90	17.30
VEDA	30	79.4	87.2	79.5	30.80	0.740	88.50	48.50
VEDA	16	31.7	51.6	26.5	36.60	0.740	312.90	265.20
VEDA	8	3.9	16.0	2.5	–	0.400	729.40	687.90
VEDA	4	0.0	27.5	0.1	–	0.000	–	–
TangentMol	100	95.4	96.5	98.0	6.70	0.720	52.20	15.00
TangentMol	50	94.6	96.2	98.2	7.50	0.820	55.90	17.30
TangentMol	30	91.9	94.3	97.4	8.30	0.950	58.90	21.90
TangentMol	16	85.4	89.8	92.1	13.50	1.000	70.10	32.90
TangentMol	8	64.4	71.5	67.3	21.70	1.070	125.50	84.10
TangentMol	4	13.2	20.6	11.8	–	0.650	445.40	391.40

Table A5: Full training-configuration and objective ablation on QM9 at 4, 8, 16, and 100 NFE. The + Muon, + remove coordinate preconditioning, and VEDA-FS blocks cumulatively introduce Muon, remove coordinate preconditioning, and remove self-conditioning; all rows from VEDA-FS onward share the resulting configuration, and all rows share one sampler and evaluator. TangentMol entries are means over three training runs. The VEDA rows use different sampling settings from Figure 2 and Table A3 and are not directly comparable to them.

Configuration / objective	NFE	Connected validity ↑	Mol. stability ↑	FCD ↓	Energy ↓	Strain ↓
VEDA	4	2.9	1.4	10.52	5550.00	5540.85
VEDA	8	68.7	61.7	5.22	174.20	136.45
VEDA	16	96.3	95.9	3.17	51.76	15.89
VEDA	100	99.6	99.8	1.43	35.35	9.62
+ Muon	4	3.2	2.8	17.97	2604.10	2594.19
+ Muon	8	74.2	71.1	5.16	260.18	222.98
+ Muon	16	97.9	97.4	3.35	50.68	15.74
+ Muon	100	99.7	99.8	1.82	35.12	9.81
+ remove coordinate preconditioning	4	34.6	29.9	7.11	282.92	248.67
+ remove coordinate preconditioning	8	88.0	86.2	4.89	63.42	26.57
+ remove coordinate preconditioning	16	98.2	98.0	3.34	45.70	11.30
+ remove coordinate preconditioning	100	99.8	99.8	1.79	33.86	9.02
+ remove self-conditioning (VEDA-FS)	4	53.3	49.1	6.54	165.03	128.30
+ remove self-conditioning (VEDA-FS)	8	92.7	91.6	4.15	55.69	19.61
+ remove self-conditioning (VEDA-FS)	16	98.9	98.7	2.91	44.35	11.09
+ remove self-conditioning (VEDA-FS)	100	99.7	99.7	1.60	33.90	9.21
VEDA-FS + certainty	4	66.5	62.0	5.96	147.32	109.68
VEDA-FS + certainty	8	94.4	93.4	4.08	52.65	18.09
VEDA-FS + certainty	16	99.2	99.1	2.99	43.27	11.10
VEDA-FS + certainty	100	99.8	99.8	1.95	32.52	9.10
VEDA-FS + coordinate tangent	4	74.8	73.6	1.38	80.11	56.25
VEDA-FS + coordinate tangent	8	92.0	93.1	1.07	39.23	19.74
VEDA-FS + coordinate tangent	16	95.6	96.5	1.51	27.54	13.45
VEDA-FS + coordinate tangent	100	98.8	98.6	1.89	24.49	10.02
VEDA-FS + logit tangent	4	63.8	59.4	4.36	151.58	115.91
VEDA-FS + logit tangent	8	94.4	93.7	3.90	55.33	18.67
VEDA-FS + logit tangent	16	99.0	99.1	2.85	44.12	11.13
VEDA-FS + logit tangent	100	99.6	99.8	1.62	35.09	9.27
VEDA-FS + both tangent terms	4	78.3	77.9	1.38	78.69	53.80
VEDA-FS + both tangent terms	8	92.8	94.2	1.00	38.82	18.78
VEDA-FS + both tangent terms	16	95.7	95.5	1.49	30.73	15.23
VEDA-FS + both tangent terms	100	97.2	94.1	2.00	32.06	16.38
TangentMol	4	79.3	78.1	1.41	78.11	52.86
TangentMol	8	93.3	94.1	0.99	38.09	17.67
TangentMol	16	96.1	96.5	1.57	26.58	12.22
TangentMol	100	98.8	98.2	2.30	23.78	9.61

Table A6: Full GEOM-Drugs objective ablation. VEDA-FS + $\boldsymbol{x}$ adds the coordinate tangent, VEDA-FS + $\boldsymbol{x}, \boldsymbol{z}$ adds both tangent terms, and TangentMol further adds certainty regularization. Lower is better for optimized RMSD, energy, strain, and FCD; percentages are higher-is-better. FCD is not reported at 4 steps. Uniqueness rounds to 100% and novelty is at least 99.1% in every setting.

Objective	NFE	Connected validity ↑	Valid ↑	Mol. stability ↑	FCD ↓	Opt. RMSD ↓	Energy ↓	Strain ↓
VEDA-FS	4	2.9	7.9	2.2	–	0.410	481.70	447.50
VEDA-FS	8	46.5	48.7	45.3	39.40	0.750	180.40	113.60
VEDA-FS	16	85.6	88.8	87.2	28.60	0.840	89.70	39.10
VEDA-FS	30	92.4	96.5	97.4	20.30	0.880	57.90	20.30
VEDA-FS	50	95.9	98.4	99.1	17.10	0.830	51.80	17.20
VEDA-FS	100	97.1	99.4	99.7	16.00	0.720	36.60	13.40
VEDA-FS + $\boldsymbol{x}$	4	7.9	14.3	4.5	–	0.690	552.20	509.10
VEDA-FS + $\boldsymbol{x}$	8	47.6	56.6	47.2	26.50	1.020	158.00	123.10
VEDA-FS + $\boldsymbol{x}$	16	81.5	84.7	86.0	14.50	0.990	68.10	33.70
VEDA-FS + $\boldsymbol{x}$	30	90.2	92.6	95.9	9.50	0.900	58.20	23.10
VEDA-FS + $\boldsymbol{x}$	50	92.0	92.8	97.2	8.00	0.780	50.60	16.50
VEDA-FS + $\boldsymbol{x}$	100	95.3	96.6	98.0	8.00	0.680	45.10	14.50
VEDA-FS + $\boldsymbol{x}, \boldsymbol{z}$	4	10.4	20.0	8.0	–	0.600	460.20	408.10
VEDA-FS + $\boldsymbol{x}, \boldsymbol{z}$	8	52.8	57.8	54.3	25.80	1.030	139.40	100.40
VEDA-FS + $\boldsymbol{x}, \boldsymbol{z}$	16	76.8	79.8	84.5	12.20	1.020	71.30	35.10
VEDA-FS + $\boldsymbol{x}, \boldsymbol{z}$	30	81.9	83.7	90.8	7.30	0.930	59.50	22.70
VEDA-FS + $\boldsymbol{x}, \boldsymbol{z}$	50	83.4	84.7	92.8	6.80	0.820	50.90	18.00
VEDA-FS + $\boldsymbol{x}, \boldsymbol{z}$	100	88.6	89.0	93.1	6.40	0.720	52.20	15.90
TangentMol	4	13.2	20.6	11.8	–	0.650	445.40	391.40
TangentMol	8	64.4	71.5	67.3	21.70	1.070	125.50	84.10
TangentMol	16	85.4	89.8	92.1	13.50	1.000	70.10	32.90
TangentMol	30	91.9	94.3	97.4	8.30	0.950	58.90	21.90
TangentMol	50	94.6	96.2	98.2	7.50	0.820	55.90	17.30
TangentMol	100	95.4	96.5	98.0	6.70	0.720	52.20	15.00

Table A7: Compute-matched VEDA-FS and tangent-normalization ablation on QM9 at 4, 8, 16, and 100 NFE. VEDA-FS (520 epochs) continues the 300-epoch VEDA-FS run. The no-normalization rows remove the RMS normalization of both tangents; the rescaled row additionally matches the average target displacement of each branch to the default setting. All rows share the sampler and evaluator of Table A5; TangentMol entries are means over three training runs, and the other rows are single training runs.

Configuration / objective	NFE	Connected validity ↑	Mol. stability ↑	FCD ↓	Energy ↓	Strain ↓
VEDA-FS (300 epochs)	4	53.3	49.1	6.54	165.03	128.30
VEDA-FS (300 epochs)	8	92.7	91.6	4.15	55.69	19.61
VEDA-FS (300 epochs)	16	98.9	98.7	2.91	44.35	11.09
VEDA-FS (300 epochs)	100	99.7	99.7	1.60	33.90	9.21
VEDA-FS (520 epochs)	4	57.2	51.6	6.08	160.71	121.43
VEDA-FS (520 epochs)	8	93.6	92.4	4.47	56.03	18.56
VEDA-FS (520 epochs)	16	99.1	99.0	3.16	44.90	11.04
VEDA-FS (520 epochs)	100	99.7	99.7	1.76	35.38	9.46
TangentMol	4	79.3	78.1	1.41	78.11	52.86
TangentMol	8	93.3	94.1	0.99	38.09	17.67
TangentMol	16	96.1	96.5	1.57	26.58	12.22
TangentMol	100	98.8	98.2	2.30	23.78	9.61
TangentMol, no normalization	4	73.5	69.6	2.93	107.60	79.01
TangentMol, no normalization	8	92.6	93.1	1.75	50.85	23.13
TangentMol, no normalization	16	94.7	96.5	1.11	33.25	14.06
TangentMol, no normalization	100	97.6	97.1	2.38	25.88	10.61
TangentMol, no normalization, rescaled	4	69.9	66.5	2.90	127.80	99.46
TangentMol, no normalization, rescaled	8	90.1	91.0	2.02	57.28	29.23
TangentMol, no normalization, rescaled	16	92.7	95.5	1.24	36.20	16.45
TangentMol, no normalization, rescaled	100	96.9	98.1	2.18	26.01	9.63

A.11 CROSSDOCKED OPERATING POINTS

Table A8 reports the guided and unguided CrossDocked evaluations used in the target-aware comparison. All PocketVE and TangentMol rows use the generalized-arcsin $(2.2, 0.57)$ schedule. The CFG=5 rows at 10, 20, and 30 steps use 100 requested ligands for each of 100 pockets (10,000 per row); every other PocketVE and TangentMol operating point uses 10 ligands for each pocket (1,000 per row). Sampling times in Section 5.3 were measured on one A100 GPU with batch size 10, excluding docking and post-processing. The PAFlow rows use its own sampler with 20 and 50 steps and its own classifier guidance rather than classifier-free guidance, with 100 requested ligands for each of the same 100 pockets. We use GenBench3D (Baillif et al., 2024) for 3D structural quality evaluation; Validity3D and clash-free rate are its reported metrics. Molecule stability is a separate valence-based statistic and is included here for completeness, alongside molecular-quality and binding-affinity metrics. With $\mathrm{CFG} = 5$, TangentMol improves Validity3D, clash-free rate, and strain over PocketVE at 10–30 steps; at 50 and 100 steps it retains a higher clash-free rate and lower strain but has lower Validity3D. Without guidance, TangentMol improves Validity3D, clash-free rate, and strain at 10 steps, but falls below PocketVE on Validity3D and clash-free rate at 20 or more steps; it also has lower QED and less favorable Vina scores at every unguided budget, while achieving higher molecule stability throughout. PAFlow attains high clash-free rates and the most favorable Vina scores, but its Validity3D (29.31% at 20 steps and 45.37% at 50 steps) is below every guided TangentMol setting, and its median strain (3,885.1 and 1,994.0) is higher than that of every PocketVE and TangentMol setting.

Table A8: CrossDocked operating points for PocketVE, TangentMol, and PAFlow. PAFlow uses classifier guidance instead of classifier-free guidance, so its CFG entry is a dash. Molecule stability, Validity3D, and clash-free rate are percentages; strain is the median strain energy. Vina Score and Vina Min are reported as medians. Lower is better for strain and the Vina metrics; higher is better for the remaining metrics.

Method	Steps	CFG	N_{req}	Mol. stable	Validity3D	Clash-free	Strain	QED	SA	Vina score	Vina min
PocketVE	10	0	1,000	8.6	36.66	69.31	1,510.5	0.42	0.58	-4.83	-5.59
TangentMol	10	0	1,000	19.8	48.06	74.34	730.7	0.37	0.61	-4.65	-5.24
PocketVE	10	5	10,000	5.9	45.27	81.17	1,201.5	0.52	0.61	-6.29	-6.72
TangentMol	10	5	10,000	17.9	62.75	90.90	289.4	0.58	0.66	-6.49	-6.78
PocketVE	20	0	1,000	19.5	58.27	84.30	530.9	0.49	0.60	-6.17	-6.54
TangentMol	20	0	1,000	31.5	56.31	81.56	511.7	0.39	0.59	-5.44	-5.90
PocketVE	20	5	10,000	13.9	61.75	90.30	388.5	0.59	0.66	-7.29	-7.67
TangentMol	20	5	10,000	28.0	71.21	92.64	156.4	0.61	0.69	-7.15	-7.48
PAFlow	20	–	10,000	0.6	29.31	96.70	3,885.1	0.46	0.57	-8.47	-8.35
PocketVE	30	0	1,000	24.9	66.83	88.06	348.9	0.48	0.62	-6.50	-6.91
TangentMol	30	0	1,000	40.4	62.40	86.40	338.4	0.38	0.60	-5.70	-5.99
PocketVE	30	5	10,000	20.1	69.98	91.54	234.0	0.60	0.68	-7.50	-7.84
TangentMol	30	5	10,000	31.4	73.38	93.07	127.0	0.61	0.71	-7.29	-7.57
PocketVE	50	0	1,000	33.1	77.03	88.74	214.5	0.50	0.63	-6.46	-6.92
TangentMol	50	0	1,000	45.4	68.04	86.86	257.3	0.39	0.59	-5.65	-6.01
PocketVE	50	5	1,000	26.0	77.17	92.44	156.0	0.62	0.71	-7.55	-7.88
TangentMol	50	5	1,000	32.9	74.43	94.08	111.5	0.62	0.73	-7.49	-7.75
PAFlow	50	–	10,000	1.7	45.37	96.38	1,994.0	0.49	0.57	-8.89	-8.97
PocketVE	100	0	1,000	39.8	79.91	91.01	169.5	0.49	0.64	-6.76	-7.12
TangentMol	100	0	1,000	48.1	77.26	89.20	197.6	0.38	0.59	-5.99	-6.31
PocketVE	100	5	1,000	33.5	81.97	93.89	121.6	0.62	0.72	-7.86	-8.14
TangentMol	100	5	1,000	36.2	77.19	94.41	99.3	0.61	0.74	-7.72	-7.90

Table A9 compares generated and reference ligand distributions for the $\mathrm{CFG} = 5$ operating points in Table 2. At every matched budget, TangentMol lowers all four divergences relative to PocketVE; at 10 steps, Local, JSD-All-12Å, JSD-CC-2Å, and atom-type JS decrease from 0.442, 0.121, 0.414, and 0.101 to 0.277, 0.075, 0.280, and 0.089, respectively. At 10 steps, TangentMol is also below IPDiff and both PAFlow budgets on all four metrics, and at 30 steps its distance JSDs approach those of 1000-step TargetDiff.

Table A9: CrossDocked molecular-distribution divergences. TargetDiff and IPDiff rows are reference values from the PocketVE benchmark (Zhang & Bi, 2026); the PAFlow rows are evaluated with the same protocol on the same 100 pockets. PocketVE and TangentMol rows use CFG $= 5$ and the 10,000-ligand runs of Table 2. Local JSD averages the eight short-range pairwise-distance JSDs over 6-6$|1, 2, 4$, 6-7$|1, 2, 4$, and 6-8$|1, 2$. JSD-All-12Å pools intraligand atom-pair distances up to 12 Å, JSD-CC-2Å measures short-range carbon–carbon distances, and Atom is atom-type frequency JS. Lower is better; bold marks the better value within each matched-step PocketVE/TangentMol pair.

Method	Steps	Local ↓	12Å ↓	CC-2Å ↓	Atom ↓
TargetDiff	1000	0.234	0.056	0.213	0.059
IPDiff	1000	0.391	0.098	0.349	0.194
PAFlow	50	0.452	0.141	0.460	0.227
PAFlow	20	0.518	0.175	0.530	0.206
PocketVE	10	0.442	0.121	0.414	0.101
TangentMol	10	**0.277**	**0.075**	**0.280**	**0.089**
PocketVE	20	0.345	0.080	0.325	0.119
TangentMol	20	**0.248**	**0.061**	**0.241**	**0.097**
PocketVE	30	0.316	0.069	0.299	0.121
TangentMol	30	**0.236**	**0.059**	**0.235**	**0.101**

A.12 GEOM-Drugs Revisited xTB results

Table A10 reports the GEOM-Drugs Revisited xTB evaluation. This evaluator is separate from the MMFF-style evaluation in Table A4 because the two protocols use different energy and RMSD definitions. The column layout follows the prior VEDA xTB table (Zhang et al., 2026): bond length is the mean absolute error in units of 10^{-2} Å, and $\Delta E_{\mathrm{relax}}$ and RMSD compare each generated conformer with its GFN2-xTB relaxation. For methods evaluated on five subsets, the second line of each entry is the standard deviation across subsets. SemlaFlow, FlowMol2, FlowMol3, MiDi, and EQGAT-diff use 5,000 molecules in five subsets of 1,000; FlowMol2 is the CTMC sampler. MiDi uses 200 or 500 discrete sampling updates from its 500-step process. EQGAT-diff uses 100, 200, 300, or 500 sampling updates. TangentMol requests 5,000 molecules, whereas VEDA and VEDA-FS request 1,000 each and are reported as single values without standard deviations. Energy and RMSD are omitted when no molecule completes relaxation; the table also includes operating points omitted from the xTB panel of Figure 2.

Table A10: GEOM-Drugs Revisited xTB results. MS, V&C, and atom stability are higher-is-better; all other columns are lower-is-better. The second line is the standard deviation across five evaluation subsets. VEDA and VEDA-FS use 1,000 molecules and report single values.

Method	NFE	MS ↑	V&C ↑	Atom stability ↑	Bond length ($\times 10^{-2}$) ↓	Bond angles ↓	Torsions ↓	Median ΔE_{relax} ↓	Mean ΔE_{relax} ↓	Median RMSD ↓	Mean RMSD ↓
SemlaFlow	4	.000	.008	.690	30.54	16.07	42.23	923.60	1238.33	1.031	1.005
		±.000	±.005	±.008	±.64	±.24	±.84	±171.05	±558.07	±.282	±.251
SemlaFlow	8	.278	.387	.950	8.71	5.18	19.78	155.65	342.17	.664	.737
		±.043	±.050	±.004	±.47	±.26	±1.11	±20.61	±46.37	±.044	±.057
SemlaFlow	16	.914	.873	.998	3.50	2.37	9.63	47.31	99.03	.479	.570
		±.015	±.022	±.000	±.10	±.09	±.54	±5.11	±15.56	±.053	±.040
SemlaFlow	30	.961	.894	.999	3.18	2.15	7.51	37.98	92.49	.352	.474
		±.012	±.019	±.000	±.06	±.07	±.52	±4.64	±21.45	±.058	±.051
SemlaFlow	50	.970	.910	.999	3.03	2.06	6.51	33.81	98.33	.290	.419
		±.008	±.018	±.000	±.06	±.06	±.38	±4.06	±29.09	±.048	±.041
SemlaFlow	100	.977	.923	.999	2.94	1.97	5.67	31.91	88.94	.246	.376
		±.006	±.014	±.000	±.02	±.04	±.39	±3.10	±22.00	±.043	±.043
FlowMol2	4	.000	.000	.667	34.88	17.90	47.03	89.56	89.56	.210	.210
		±.000	±.001	±.003	±.47	±.16	±.70	±179.11	±179.11	±.419	±.419
FlowMol2	8	.010	.030	.851	15.23	10.30	38.66	159.53	194.05	1.127	1.183
		±.003	±.004	±.001	±.28	±.11	±.27	±24.79	±40.07	±.109	±.115
FlowMol2	16	.347	.304	.964	4.78	4.13	24.89	55.60	74.70	1.035	1.092
		±.019	±.011	±.001	±.10	±.06	±.26	±4.09	±3.54	±.027	±.018
FlowMol2	30	.586	.489	.982	2.89	2.66	19.28	30.50	44.21	.999	1.052
		±.005	±.009	±.001	±.11	±.04	±.31	±.93	±1.83	±.014	±.013
FlowMol2	50	.799	.635	.993	1.85	1.99	16.43	22.49	31.58	.964	1.017
		±.003	±.011	±.000	±.04	±.03	±.29	±.47	±1.06	±.015	±.013
FlowMol2	100	.910	.721	.997	1.46	1.72	15.14	19.14	25.89	.974	1.028
		±.005	±.020	±.000	±.02	±.01	±.12	±.40	±.57	±.013	±.008
FlowMol2	250	.949	.755	.999	1.28	1.57	14.56	17.50	23.15	.952	1.001
		±.005	±.014	±.000	±.01	±.01	±.19	±.31	±.46	±.015	±.012
FlowMol3	4	.000	.001	.565	41.41	20.46	48.50	891.28	891.28	.485	.485
		±.000	±.001	±.005	±.48	±.28	±.84	±1091.80	±1091.80	±.621	±.621
FlowMol3	8	.035	.093	.857	16.45	9.98	31.01	321.74	796.49	.838	.865
		±.004	±.009	±.002	±.22	±.06	±.43	±24.26	±124.65	±.036	±.018
FlowMol3	16	.473	.554	.964	5.09	3.78	17.19	42.78	249.01	.679	.748
		±.008	±.009	±.001	±.10	±.06	±.08	±6.51	±24.74	±.009	±.011
FlowMol3	30	.843	.843	.992	1.98	1.72	9.90	8.83	63.18	.552	.622
		±.007	±.010	±.000	±.04	±.03	±.21	±.29	±9.50	±.019	±.015
FlowMol3	50	.983	.965	.999	1.12	1.00	6.02	5.05	25.72	.389	.489
		±.003	±.005	±.000	±.03	±.02	±.16	±.05	±14.68	±.012	±.011
FlowMol3	100	.998	.990	1.000	.99	.87	4.53	4.49	6.73	.313	.421
		±.002	±.002	±.000	±.01	±.01	±.10	±.08	±.69	±.011	±.009
FlowMol3	250	.999	.993	1.000	1.00	.86	4.02	4.61	7.43	.282	.393
		±.001	±.001	±.000	±.00	±.00	±.06	±.06	±2.67	±.006	±.008
EQGAT-diff	100	.397	.414	.925	8.91	7.32	26.08	19.32	37.77	.518	.598
		±.015	±.012	±.003	±.42	±.14	±.36	±.81	±1.72	±.016	±.014
EQGAT-diff	200	.496	.514	.938	7.65	6.64	24.73	15.55	28.69	.492	.551
		±.015	±.018	±.006	±.81	±.42	±1.42	±.78	±1.18	±.012	±.015
EQGAT-diff	300	.832	.825	.993	1.49	1.82	11.78	7.28	14.40	.426	.510
		±.010	±.008	±.000	±.07	±.05	±.17	±.20	±.39	±.013	±.006
EQGAT-diff	500	.903	.828	.997	.97	1.13	8.73	6.64	11.21	.607	.683
		±.008	±.010	±.000	±.02	±.02	±.13	±.17	±.52	±.014	±.014
MiDi	200	.457	.472	.976	3.48	2.53	14.34	27.18	50.34	.905	.968
		±.102	±.111	±.002	±.23	±.15	±1.89	±11.32	±15.06	±.195	±.190
MiDi	500	.856	.717	.996	1.71	1.61	11.08	19.18	30.27	.836	.911
		±.038	±.086	±.000	±.09	±.11	±1.81	±7.99	±10.10	±.202	±.196
VEDA	4	.000	.000	.420	36.78	19.96	51.79	–	–	–	–
VEDA	8	.087	.247	.805	15.72	11.80	36.28	476.68	584.32	.830	.837
VEDA	16	.329	.661	.949	5.61	4.38	20.01	135.72	224.37	.714	.755
VEDA	30	.835	.912	.993	1.37	1.23	10.00	9.33	24.71	.464	.548
VEDA	50	.964	.964	.999	.75	.65	5.92	2.68	7.10	.250	.378
VEDA	100	.991	.982	1.000	.61	.42	3.15	1.45	2.59	.109	.212
VEDA-FS	4	.071	.424	.861	12.78	9.44	28.98	349.80	534.52	.701	.798
VEDA-FS	8	.586	.937	.981	3.11	3.11	18.25	64.87	89.72	.625	.675
VEDA-FS	16	.923	.971	.998	1.23	1.42	12.44	12.02	20.12	.548	.625
VEDA-FS	30	.979	.947	.999	.78	.87	8.36	4.95	8.21	.452	.544
VEDA-FS	50	.987	.977	1.000	.65	.69	6.49	3.08	4.92	.322	.447
VEDA-FS	100	.996	.987	1.000	.59	.54	4.32	1.87	2.87	.188	.306
TangentMol	4	.247	.678	.944	7.83	6.30	24.42	197.37	323.86	.757	.812
		±.045	±.028	±.006	±.36	±.28	±.91	±41.14	±73.74	±.055	±.043
TangentMol	8	.774	.943	.993	2.21	2.21	15.14	30.55	48.84	.656	.713
		±.039	±.014	±.001	±.12	±.12	±.61	±7.28	±8.22	±.056	±.047
TangentMol	16	.907	.958	.997	1.18	1.24	10.64	10.63	17.18	.665	.715
		±.014	±.008	±.000	±.04	±.04	±.56	±1.67	±1.83	±.044	±.043
TangentMol	30	.966	.976	.999	.83	.84	6.87	5.02	7.96	.470	.565
		±.007	±.006	±.000	±.03	±.04	±.55	±1.06	±1.22	±.067	±.061
TangentMol	50	.979	.985	.999	.79	.72	5.16	3.48	5.64	.308	.436
		±.007	±.003	±.000	±.04	±.04	±.36	±.66	±1.00	±.065	±.053
TangentMol	100	.990	.985	1.000	.75	.63	3.70	2.55	4.09	.180	.302
		±.003	±.004	±.000	±.04	±.04	±.27	±.42	±.74	±.026	±.033