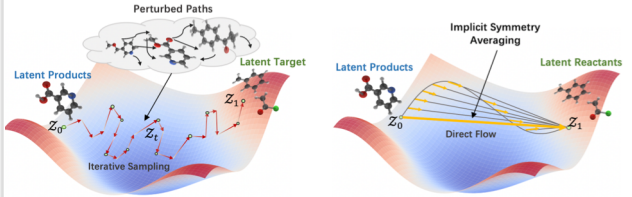
Efficient Retrosynthesis Prediction with Integral Flow Matching and Latent Inversion

Tao Yin [1] · Xiaohong Zhang*[1] · Yinjie Zhu [1] · Jiacheng Zhang [1] · Haotian Zou [1] · Li Huang [1] · JiaJun Cai [1] · Zhibin Zhang [1] · Meng Yan*[1]

Paper

Code

Introduction



Efficiency-Accuracy Bottleneck:

① Recent diffusion and flow-matching models provide accurate, diverse predictions but rely on costly iterative ODE/SDE sampling.

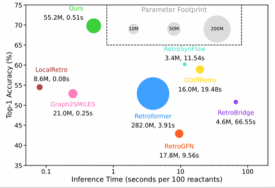
② Notably, simply using fewer steps is unreliable because small latent endpoint errors can cross discrete decoding boundaries and alter molecular graph attributes. This creates a central trade-off between inference speed and prediction quality.

Key Observation:

① Iterative Sampling:
Multiple ODE/SDE steps→ slow inference.

② Naive One-Step Sampling:
Local velocity extrapolation→Unstable decoding.

③ Conclusion:
Efficient retrosynthesis requires finite-interval transport aligned with one-step inference.



Conclusion

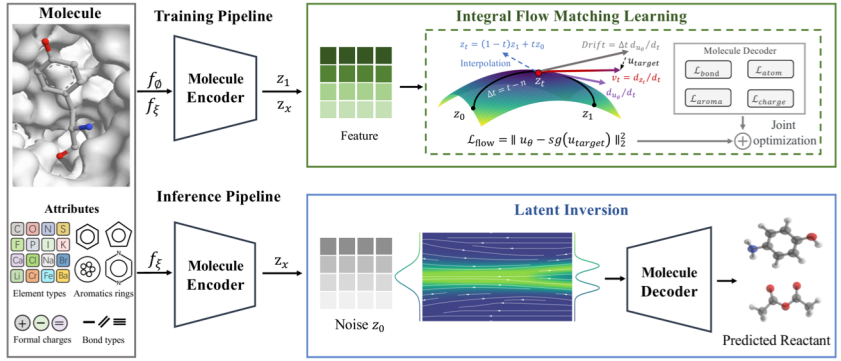
Framework: FlashRetro learns finite-interval transport in a continuous molecular latent space.

Inference: Latent Inversion predicts reactants with one learned transport update and no ODE/SDE solver.

Theory: Endpoint-error and decoder-stability analyses connect transport learning to graph reconstruction.

Results: 69.3% Top-1 accuracy on USPTO-50K, with 22.6×–38.2× speedups over iterative baselines.

Framework & Results



Integral Flow Matching

$z_t = (1-t)z_1 + tz_0, \quad v_t = z_0 - z_1$

Finite-interval transport target

$u_{tgt} = v_t - (t-r)\, du_\theta / dt$

$L_{flow} = E[\, \| u_\theta - sg(u_{tgt}) \|^2_M \,]$

1-NFE Latent Inversion

$\hat{z}_1 = z_0 - u_\theta(z_0, 0, 1 \mid z_x)$

$\|\hat{z}_1 - z_1\|^2 = \|u_\theta - u^*\|^2$

Molecular Reconstruction

$L_{recon} = \Sigma_a \, CE(d_\psi^{(a)}(z_1, z_x), y^{(a)})$

Masked atom, bond, charge and aromaticity losses

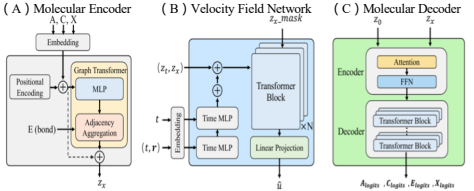


Table 1: Main results on USPTO-50k benchmark. Comparison of Top-k (k = 1, 3, 5) exact match accuracy across different retrosynthesis paradigms. Evaluation covers both unknown and known reaction class settings. Best results within each category are highlighted in **bold**.

	Baselines	Top-k accuracy (unknown class)			Top-k accuracy (known class)		
		k=1	k=3	k=5	k=1	k=3	k=5
Template-Based	MHNreact [Seidl et al., 2022]	51.8	74.6	81.2	-	-	-
	GLN [Dai et al., 2019]	52.5	69.0	75.6	64.2	79.1	85.2
	LocalRetro [Chen and Jung, 2021]	53.4	**77.5**	**85.9**	63.9	**86.8**	**92.4**
	GraphRetro [Somnath et al., 2021]	53.7	68.3	72.2	63.9	81.5	85.2
	RetroCompOser [Yan et al., 2022]	54.5	77.2	83.2	65.9	85.8	89.5
	Dual-TB [Sun et al., 2020]	**55.2**	74.6	80.5	**67.7**	84.8	88.9
Semi-Template	MEGAN [Sacha et al., 2021]	48.1	70.7	78.4	60.7	82.0	**87.5**
	G2Gs [Shi et al., 2020]	48.9	67.6	72.5	61.0	81.3	86.0
	RetroXpert [Yan et al., 2020]	50.4	61.1	62.3	62.1	75.8	78.5
	G^2 Retro [Chen et al., 2023]	51.4	72.1	78.2	59.4	79.4	84.2
	RetroPrime [Wang et al., 2021]	51.4	70.8	74.0	64.8	82.7	85.0
	RetroDiff [Wang et al., 2023]	52.6	71.2	81.0	-	-	-
	GDiffRetro [Sun et al., 2025]	58.9	**79.1**	81.9	**67.6**	**84.1**	86.1
	RetroSynFlow [Yadav et al., 2025]	**60.0**	77.9	**82.7**	-	-	-
Template-Free	Transformer [Vaswani et al., 2017]	43.7	59.7	65.1	-	-	-
	SCROP [Zheng et al., 2019]	43.7	60.0	65.2	59.0	74.8	78.1
	Retroformer [Wan et al., 2022]	52.9	68.2	72.5	64.0	**82.5**	**86.7**
	Graph2SMILES [Tu and Coley, 2022]	52.9	66.5	70.0	-	-	-
	Dual-TF [Sun et al., 2020]	53.6	70.7	74.6	65.7	81.9	84.7
	RetroBridge [Igashov et al., 2024]	50.8	74.1	80.6	-	-	-
	RetroDiffuse [Yadav et al., 2025]	50.2	74.3	**81.2**	-	-	-
	FlashRetro	**69.3**	**78.4**	79.8	**71.8**	80.5	82.1

Analysis

FlashRetro is strongest on localized functional-group transformations, reaching 83.54% Top-1 accuracy on acylation, 85.45% on protections, and 90.48% on functional-group addition.

The decoder exhibits bounded local tolerance. Perturbing the ground-truth reactant latent preserves approximately 69% Top-1 accuracy up to $\sigma = 0.30$. The one-step predicted latent remains stable under small perturbations ($\sigma \leq 0.05$). This confirms that FlashRetro relies on accurate endpoint transport into a locally decodable latent neighborhood.

Table 6: Detailed Top-k performance breakdown by reaction class on the USPTO-50k test set. Reaction fractions are based on the standard dataset distribution.

Reaction Class	Fraction (%)	Top-k Accuracy			
		$k=1$	$k=3$	$k=5$	$k=10$
Heteroatom Alkylation/Arylation	30.3	66.58	82.14	84.79	85.98
Acylation And Related Processes	23.8	83.54	90.97	91.15	91.77
C-C Bond Formation	11.3	41.53	50.12	52.27	54.42
Heterocycle Formation	1.8	42.59	46.30	46.30	46.30
Protections	1.3	85.45	87.27	87.27	87.27
Deprotections	16.5	77.66	80.67	81.74	82.09
Reductions	9.2	77.01	81.16	81.44	82.27
Oxidations	1.6	54.69	64.06	65.62	65.62
Functional Group Interconversion	3.7	66.42	75.37	76.87	79.10
Functional Group Addition	0.5	90.48	95.24	95.24	95.24

Table 9: Robustness when perturbing the ground-truth reactant latent z_1.

Noise Std. (σ)	Top-1	Val@1	Top-3	Val@3	Top-5	Val@5	Top-10	Val@10
0.00	69.30	99.98	78.40	96.82	79.80	90.82	81.10	70.20
0.05	69.48	100.0	78.15	96.61	79.86	90.36	81.18	69.71
0.10	69.44	99.96	78.17	97.19	79.60	91.82	81.00	72.40
0.20	69.56	99.98	78.09	98.88	79.62	96.63	80.98	85.93
0.25	69.28	100.0	78.25	99.70	79.78	98.86	81.14	94.16
0.30	69.02	100.0	77.57	99.95	78.97	99.84	80.16	98.60
0.50	15.12	100.0	17.50	100.0	18.08	100.0	19.08	99.96

Table 10: Robustness when perturbing the predicted latent produced by the 1-step flow.

Noise Std. (σ)	Top-1	Val@1	Top-3	Val@3	Top-5	Val@5	Top-10	Val@10
0.00	69.30	99.98	78.40	96.82	79.80	90.82	81.10	70.20
0.02	69.54	99.96	78.73	97.66	80.06	92.73	81.36	74.46
0.05	69.34	100.0	79.07	99.41	80.52	97.92	82.22	91.21
0.10	53.56	100.0	61.37	100.0	62.90	100.0	64.22	99.40
0.15	3.12	100.0	3.96	100.0	4.23	99.90	4.61	99.95

