ICLR 2026 · GEM WORKSHOP · RIO DE JANEIRO · 27 APRIL 2026

GENERATIVE & EXPERIMENTAL PERSPECTIVES FOR BIOMOLECULAR DESIGN

Synthesis-Constrained Discrete Diffusion for Ionizable Lipid Generation

Rohin Maganti[1,2,†,*], Rahul Maganti[2,†], Mohamad-Gabriel Alameh[2,3] *† Equal contribution · * Correspondence: rohin.maganti@pennmedicine.upenn.edu*

[1]Department of Bioengineering, University of Pennsylvania · [2]Children's Hospital of Philadelphia · [3]Department of Medicine, University of Pennsylvania

ABSTRACT

Ionizable lipids are the critical component of lipid nanoparticles for *in vivo* mRNA delivery, but discovery is bottlenecked by combinatorial library enumeration. Existing ML approaches rank *pre-enumerated* libraries rather than generating novel structures. We introduce **synthesis-constrained discrete diffusion**, the first deep generative model for ionizable lipids that embeds reaction constraints directly into the diffusion process. Three components — **scaffold conditioning**, **region-aware noise**, and **property conditioning** via FiLM with classifier-free guidance — yield 99% chemical validity, 100% scaffold integrity, 62% novelty, and best predicted transfection potency 2× the training mean (*in silico*).

01 BACKGROUND & MOTIVATION

Lipid nanoparticle (LNP) formulations are the leading delivery platform for mRNA therapeutics. The **ionizable lipid** (iL) is the critical component: it complexes mRNA at low pH, remains neutral physiologically, and becomes cationic in endosomes to trigger cargo release.

Despite two decades of medicinal chemistry, iL discovery remains bottlenecked by **combinatorial enumeration**. The Ugi three-component reaction yields tens of thousands of candidates, but current ML approaches (**AGILE**, **LANTERN**, **LiON**) are exclusively *discriminative*: they predict activity for pre-enumerated libraries but cannot propose molecules outside them.

We introduce **synthesis-constrained discrete diffusion**, the first deep generative model for ionizable lipids. By embedding Ugi reaction constraints directly into the diffusion process, we guarantee synthesizability by construction while generating novel structures with high predicted potency.

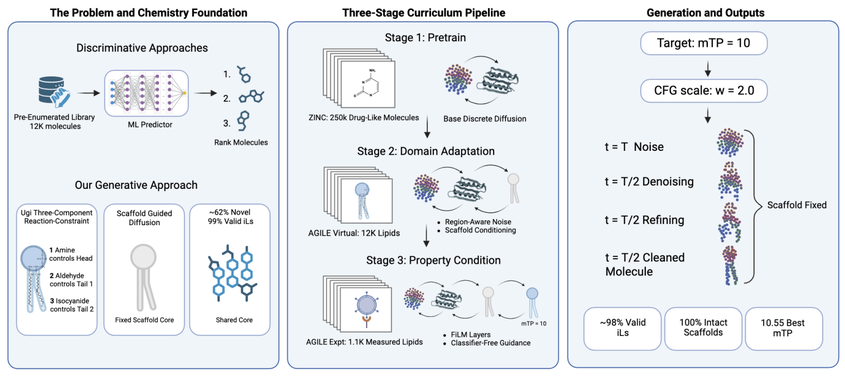

Figure 1. (A) Discriminative approaches rank pre-enumerated libraries; our generative approach discovers novel structures via Ugi scaffold constraints. **(B)** Three-stage curriculum: ZINC250k → AGILE Virtual → AGILE Experimental. **(C)** Generation: 62% novel molecules with intact scaffolds, top candidates achieving 2× training mean mTP.

02 METHOD

2.1 Discrete diffusion on graphs

We operate on discrete state spaces using the D3PM framework [Austin et al., 2021]. The forward process corrupts each node and edge attribute toward a prior marginal m via a categorical transition matrix Q_t:

FORWARD PROCESS

$$q(G_t \mid G_{t-1}) = \prod_{i=1}^{n} \mathrm{Cat}\big(x_t^{(i)};\, x_{t-1}^{(i)} Q_t^{\text{atom}}\big) \prod_{(i,j)\in\mathcal{E}} \mathrm{Cat}\big(e_t^{(ij)};\, e_{t-1}^{(ij)} Q_t^{\text{bond}}\big)$$

The reverse process trains a graph transformer f_θ to predict clean atom and bond types from corrupted inputs. Following DiGress [Vignac et al., 2023], we parameterize f_θ as a direct prediction of G_0 rather than the noise, improving sample quality for molecular graphs.

2.2 Scaffold conditioning

The Ugi reaction produces a conserved bond skeleton — amide bond, ester linkage and α-amino acid core — common to every product. We identify this invariant core C via SMARTS substructure matching (100% match on 1,100 experimental Ugi lipids) and exclude core positions from diffusion:

TRAINING LOSS · NON-CORE POSITIONS ONLY

$$\mathcal{L} = \mathbb{E}_{t,G_0,G_t}\left[\sum_{i\notin C} \ell\big(f_\theta(G_t,t)_i,\, G_0^{(i)}\big)\right]$$

At each reverse step, core bonds are deterministically restored: $E_{t-1}[C] \leftarrow E_C$. Every generated molecule contains an intact Ugi scaffold - synthesizability by construction rather than post-hoc filtering.

2.3 Region-aware noise

Ionizable lipids are chemically heterogeneous: nitrogen-rich **heads**, a fixed Ugi **core**, and carbon-dominated **tails**. A single global prior (78% C, 8% N) cannot capture this. We partition each molecule into three regions with region-specific marginals $m^{(r)}$:

FORWARD KERNEL · REGION-FACTORED

$$Q_t^{(r)} = (1-\beta_t)\, I + \beta_t\, \mathbf{1}\, m^{(r)\top}$$

TABLE 1 · EMPIRICAL ATOM MARGINALS BY REGION

ATOM	HEAD	CORE	TAIL	GLOBAL
C	0.52	0.72	0.94	0.78
N	0.28	0.12	0.003	0.08
O	0.15	0.15	0.05	0.10
other	0.05	0.01	0.007	0.04

2.4 Property conditioning via FiLM

We condition on target mTP y using FiLM [Perez et al., 2018], where scale/shift parameters are injected into each transformer layer. During training, y is replaced with $\varnothing$ with probability 0.15, enabling classifier-free guidance at inference:

CLASSIFIER-FREE GUIDANCE

$$\hat{G}_0 = f_\theta(G_t, t, \varnothing) + w\big[f_\theta(G_t, t, y) - f_\theta(G_t, t, \varnothing)\big]$$

Guidance scale w interpolates between unconditional ($w = 0$) and strong property steering ($w > 1$). We find $w = 2.0$ offers the best validity–potency trade-off.

ALGORITHM 1 · REACTION-CONSTRAINED SAMPLING

Input: f_θ, scaffold bonds E_C, core mask C, regions $r(\cdot)$, target y, guidance w, steps T

Sample G_T from region marginals $m^{(r)}$
Fix core: $E_T[C] \leftarrow E_C$
for $t = T, \ldots, 1$ **do**
  $\hat{G}_0 \leftarrow f_\theta(G_t, t, \varnothing) + w\cdot\big[f_\theta(G_t, t, y) - f_\theta(G_t, t, \varnothing)\big]$
  Sample $G_{t-1} \sim q(G_{t-1} \mid G_t, \hat{G}_0; Q_t^{(r)})$
  Restore core: $E_{t-1}[C] \leftarrow E_C$ *// exact*
return G_0

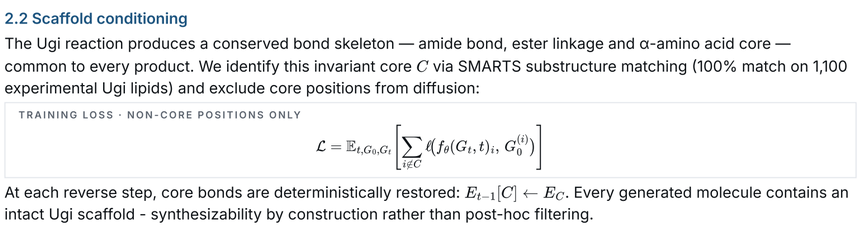

Figure 2. (A) Discrete diffusion on molecular graphs, **(B)** scaffold conditioning freezes Ugi core bonds, **(C)** region-aware noise with distinct head/tail marginals, **(D)** classifier-free guidance for property-conditioned generation.

03 TRAINING CURRICULUM

Only ~1,100 ionizable lipids have experimental transfection data, which is far too few to train from scratch. We use a three-stage curriculum:

STAGE 1 · PRETRAIN	STAGE 2 · ADAPT	STAGE 3 · CONDITION
ZINC250k	**AGILE virtual**	**AGILE experimental**
250 K drug-like	12 K Ugi lipids	1.1 K measured
General chemical grammar: valences, bond statistics, ring closure.	Ugi family & lipid-specific distributions.	mTP-conditioned fine-tuning.
	+ region-aware noise + scaffold conditioning	+ FiLM + classifier-free guidance

Variance decomposition. ANOVA attributes **65%** of mTP variance to *scaffold identity* and **35%** to *within-scaffold chemistry*, motivating property-weighted scaffold selection followed by diffusion over non-core positions.

Architecture. 8-layer graph transformer, hidden dim 256, 7.2 M params; $T = 500$ steps, cosine β-schedule; ~70 GPU-hours on a single A10G.

04 ABLATION STUDY

Removing scaffold conditioning drops scaffold integrity from 100% to 34%. Removing region-aware noise retains scaffold integrity but degrades validity from **99% to 92%**, producing implausible tail motifs:

FAILURE MODES WITHOUT REGION-AWARE NOISE

```
CCCCCCCCC=C=C=C=CCCCCC(=O)OCC…
// cumulene C=C=C=C — chemically unstable
CCCCCCC=C=C=C=C=CCCCCCCCC(=O)OCC…
// extended cumulene — absent from training corpus
```

The two components are complementary: scaffolds ensure reaction validity; region-aware noise ensures chemical plausibility in tail regions where training data is >94% sp3 carbon.

05 RESULTS

We sampled $n = 100$ molecules targeting mTP $= 10$ ($w = 2.0$) and evaluated with a frozen LANTERN ensemble (held-out $R^2 \approx 0.7$). The diffusion model and predictor are **entirely independent**: the model learns molecular distributions during Stages 1–2 with no property labels.

Generation quality

Of 100 sampled molecules, **99% pass RDKit valence checks**, **100% retain the Ugi scaffold**, **91% are unique**, and **62% are novel** (absent from the 13K AGILE library). Among novel candidates, 43% have nearest-neighbor fingerprint similarity below 0.95, confirming genuine structural divergence.

Property optimization

Best predicted mTP is **10.55** — more than **2× the training mean** (4.85). The *median* generated mTP (5.58) exceeds the training median (4.12), indicating a distributional shift toward high-activity regions. Results are robust across 3 seeds × 3 guidance scales (mean validity $93.6 \pm 7.8\%$, mean novelty $66.4 \pm 5.9\%$).

Evaluation integrity

We emphasize **generative–evaluative separation**: the diffusion model never sees mTP labels during Stages 1–2. LANTERN scoring is applied post hoc to a frozen generation, preventing the failure mode where a generator exploits predictor artifacts.

06 CHEMICAL SPACE & CONVERGENT SAR

Despite 62% novelty, the top candidates share **convergent structural features** consistent with learned structure–activity relationships:

- **Dibutylamine heads** — present in 8% of training but enriched to 42% of the top-20 generated candidates
- **C12–C16 saturated tails** with occasional terminal unsaturation — balancing membrane fluidity with endosomal escape
- **Narrow physicochemical window:** TPSA = 71 Å², HBD = 2, cLogP ≈ 12 shared across top 7 candidates

This convergent SAR is non-trivial: the model was never given explicit rules about head-group enrichment or tail-length preference. These patterns emerge from the learned distribution, providing interpretable hypotheses for experimental follow-up.

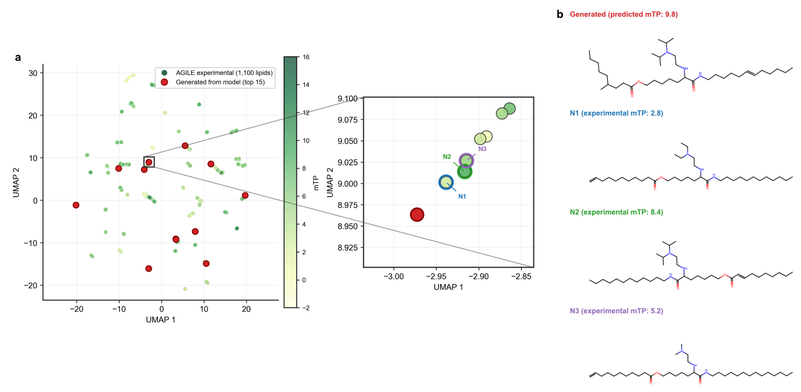

Figure 3. (a) UMAP of generated molecules (red) overlaid on AGILE training data (grey). Generated points cluster in high-density regions but extend into unpopulated chemical space. **(b)** Exemplar: generated candidate (pred. mTP 9.8) vs. its 3 nearest training neighbors (exp. mTP: 2.8, 8.4, 5.2). The 5.6-point mTP range among neighbors with high structural similarity confirms the model is interpolating toward activity maxima, not copying.

07 COMPARISON TO PRIOR WORK

TABLE 2 · CAPABILITY COMPARISON WITH DISCRIMINATIVE BASELINES

CAPABILITY	AGILE	LION	OURS
Predict activity	✓	✓	✓
Generate novel structures	—	—	✓
Guaranteed synthesizability	*library*	*library*	by construction
Property-conditioned generation	—	—	✓
Explore beyond enumeration	—	—	✓

AGILE and LiON rank within a fixed library but cannot escape the enumeration boundary. Our approach is complementary: generate candidates outside the library, then validate with discriminative models or experiment.

08 DISCUSSION & FUTURE DIRECTIONS

Paradigm shift: constrain-then-generate

The dominant approach — generate freely, then filter for synthesizability — discards >90% of candidates. Our scaffold conditioning inverts this: **every generated molecule is synthesizable by construction** because the Ugi bonds are never corrupted.

Generalizability beyond Ugi chemistry

The framework is not Ugi-specific. Any combinatorial reaction with identifiable invariant bonds can be encoded as a scaffold constraint. Immediate extensions include **Michael addition** libraries, **epoxide ring-opening**, and **reductive amination** libraries — all widely used in ionizable lipid discovery with large existing datasets. The scaffold SMARTS pattern is the only reaction-specifc component; all other model components transfer directly.

Toward closed-loop discovery

The natural next step is **active learning**: generate → synthesize → measure → retrain. Multi-property conditioning (pKa, logP, clearance, organ tropism) would enable Pareto-optimal design across the full ADMET landscape.

Limitations & ongoing validation

All reported gains are *in silico* ($R^2 \approx 0.7$); experimental confirmation is required. *In vitro* mTP correlates imperfectly with *in vivo* efficacy; organ tropism, immunogenicity, and formulation parameters remain open. Top candidates are being synthesized for validation at CHOP.

PAPER

REFERENCES

Xu et al., *Nat. Commun.* 2024 — AGILE. Vignac et al., *ICLR* 2023 — DiGress. Mehradfar et al., 2025 — LANTERN. Witten et al., *Nat. Biotech.* 2025 — LiON. Austin et al., *NeurIPS* 2021 — D3PM. Ho & Salimans, 2021 — CFG. Perez et al., *AAAI* 2018 — FiLM.

ACKNOWLEDGMENTS & CONTACT

Children's Hospital of Philadelphia Research Institute · Alameh Lab. Compute via Modal (NVIDIA A10G). Correspondence: rohin.maganti@pennmedicine.upenn.edu · mg.alameh@pennmedicine.upenn.edu