GlycoGen: Glycan Structure Prediction with Crystallizing Flows

Romain Lacombe
Independent Researcher
romain@romainlacombe.com

Abstract

Glycosylation is among the most abundant post-translational modifications in biology, governing cellular signaling and the foundational mechanisms of immunity. Because glycans cannot be sequenced from a genetic template, their structure must be inferred directly from tandem mass spectrometry (MS/MS). Modern instruments acquire 100 to 200 spectra per second, yet the prior open-vocabulary state of the art, GlycoBART, is a 207M autoregressive transformer whose sequential decoding precludes live acquisition and large-scale glycomics. Discrete Flow Matching and masked diffusion remove this sequential bottleneck by generating all token positions in parallel, provided a low-step sampler preserves structural accuracy. We introduce **GlycoGen**, a 50M-parameter Discrete Flow Matching model over fragment spectra, paired with **Crystallizing Flows**, a deterministic sampler that runs out-of-the-box via mixture embeddings with no retraining. Crystallizing Flows integrate per-token distributions on the simplex, snap each position to a one-hot vector once its argmax leaves the mask, and exit early once all positions crystallize. On the 76,774-spectrum CandyCrunch v2 benchmark, GlycoGen reaches **92.67%** top-1 accuracy in 17 passes and **90.19%** in a single pass, surpassing GlycoBART by 11.7 percentage points with a 14× to 61× speedup. Setting a new state of the art in open-vocabulary glycan structure prediction, GlycoGen makes the glycome readable in real time and at scale.

1 Introduction

Nearly all cell-surface and secreted proteins are glycosylated, carrying branched sugar structures that govern cell signaling, self–nonself recognition, and immune regulation. Unlike proteins and nucleic acids, glycans are assembled by competing enzymes without a genetic template; a single monosaccharide composition can form dozens of linkage isomers with identical mass. Tandem mass spectrometry (MS/MS) is the primary experimental probe for resolving these structures. Mapping a fragment-ion spectrum back to its underlying glycan tree is a combinatorial inverse problem, making structure prediction a central computational bottleneck in glycomics—the large-scale, systematic study of glycans (Figure 1).

This bottleneck spans both live instrument control and offline analysis. Modern MS/MS instruments acquire 100 to 200 spectra per second; data-dependent acquisition can steer next-scan selection or collision energy on the fly only when inference finishes within the 5 to 10 millisecond scan interval, as practiced via fast search in proteomics [Schweppe et al., 2020, Furtwängler et al., 2022] and glycoproteomics [Armony et al., 2023, Sutherland et al., 2024]. Even offline, a biomedical characterization campaign generates millions of spectra; slow structure prediction bottlenecks analysis and routinely leaves rare or low-abundance spectra unannotated. Spectral library search is fast only for catalogued structures, while the prior neural state of the art, GlycoBART [Abtheen et al., 2025b], decodes compressed antenna motifs autoregressively one token per forward pass—scaling latency

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

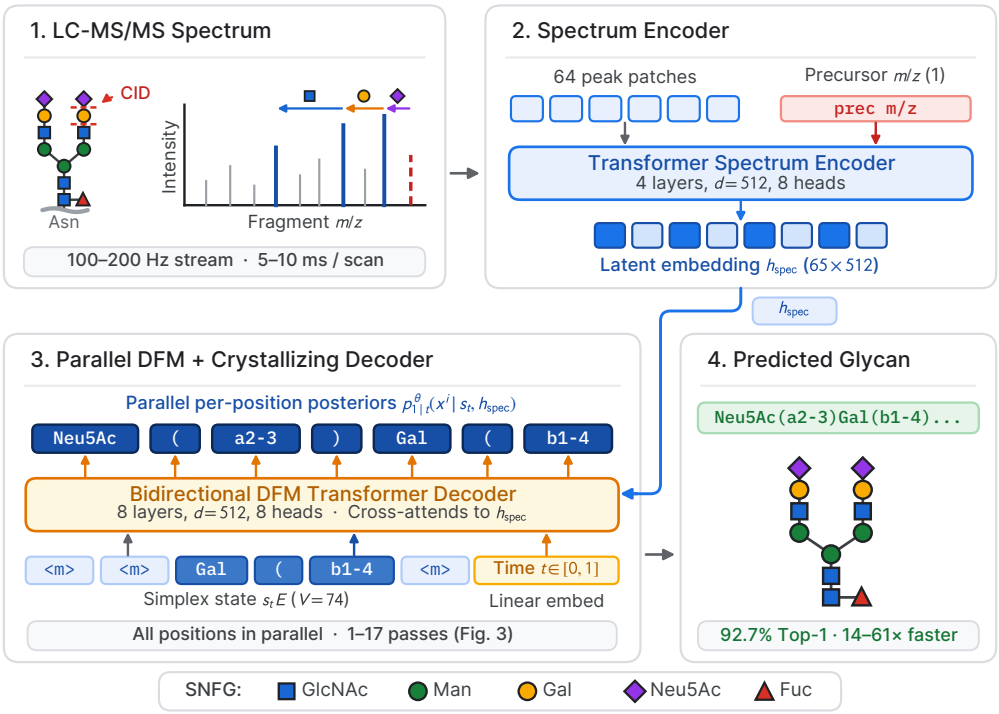


Figure 1: **GlycoGen end-to-end pipeline. (1)** Low-energy CID cleaves glycosidic bonds of branched glycans (SNFG notation [Varki et al., 2015]) into diagnostic monosaccharide losses ($\Delta m/z$) streamed at 100–200 Hz. **(2)** A 4-layer Transformer Spectrum Encoder maps 64 binned intensity/mass-offset patches and one precursor m/z token into $\mathbf{h}_{\text{spec}} \in \mathbb{R}^{65\times512}$. **(3)** An 8-layer bidirectional DFM decoder maps simplex mixture embeddings $\mathbf{s}_t E$ and time $t \in [0, 1]$ to all 74-token atomic IUPAC positions in parallel. **(4)** Crystallizing Flows (Figure 3) reach 92.7% accuracy at a 14×–61× speedup.

linearly with sequence length, losing structural accuracy when reconstructing branched trees, and relying on database snap-back over known glycans. Open-vocabulary glycan MS/MS thus faces a dual bottleneck of sequential latency and lossy structure generation.

To address both challenges, we introduce **GlycoGen**, a 50M-parameter Discrete Flow Matching denoiser [Gat et al., 2024] that predicts open-vocabulary glycan structures directly from full fragment spectra without database lookup. Conditioning on binned peak intensities and sub-bin mass offsets across the entire spectrum, GlycoGen generates atomic IUPAC tokens at all sequence positions simultaneously, decoupling inference latency from glycan length while letting every position attend bidirectionally across sibling branches.

Parallel generation realizes its speed advantage only if a low-step sampler preserves structural accuracy, yet standard Continuous-Time Markov Chain (CTMC) Euler sampling [Gat et al., 2024] draws random categorical commits that degrade sharply at small step budgets. We therefore pair GlycoGen with **Crystallizing Flows** (Figure 3), a parameter-free deterministic sampler that integrates per-position distributions on the probability simplex via mixture embeddings, snaps each position to a one-hot token once its argmax leaves `<mask>`, and exits early once all positions crystallize. **Together, GlycoGen and Crystallizing Flows set a new state of the art in open-vocabulary glycan structure prediction**, surpassing both our matched autoregressive control and GlycoBART in accuracy and speed (Table 1, Figure 4).

Contributions. **(i) GlycoGen: state-of-the-art open-vocabulary glycan structure prediction.** On the 76,774 test spectra of the CandyCrunch v2 benchmark [Urban et al., 2024], single-pass GlycoGen reaches 90.19% top-1 accuracy across three seeds, surpassing greedy GlycoBART (80.92%) by 9.3 percentage points at 61× batched throughput and 21× lower serial latency.

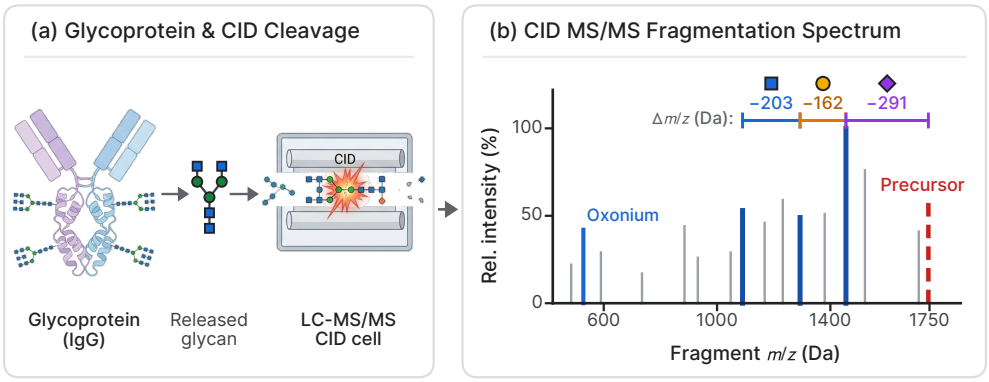


Figure 2: **Biological primer on glycomics LC-MS/MS.** Branched N-glycans released from glycoproteins (e.g., IgG antibodies) enter the collision cell, where low-energy CID cleaves glycosidic bonds into fragment spectra plotting relative intensity against m/z. Sequential mass intervals ($\Delta m/z$) read off monosaccharide losses (-291 Da Neu5Ac, -162 Da Gal, -203 Da GlcNAc) from which GlycoGen reconstructs the glycan.

(ii) Crystallizing Flows: training-free deterministic simplex sampling. This parameter-free sampler reaches **92.67%** top-1 accuracy in 17 effective passes (92.51% in 5 passes), Pareto-dominating standard CTMC Euler at low step counts and transferring without retraining as a drop-in replacement for stochastic sampling on a Masked Diffusion Language Model (MDLM) denoiser [Sahoo et al., 2024].

(iii) Graph-isomorphic generalization and single-spectrum Bayes error ceiling. On test glycans that appear in training only under a different IUPAC branch ordering, GlycoGen reaches 60.6% accuracy (62.1% with MDLM) versus 13.6% for GlycoBART. On seen glycan structures, GlycoGen reaches 93.31% graph accuracy—approaching the Bayes error ceiling of single-spectrum CID, as four in five remaining errors reflect linkage isomers or labile sugar losses that single-stage MS/MS cannot physically distinguish (Section 5).

2 Background and Prior Work

2.1 Glycomics and Tandem Mass Spectrometry

Glycan biology and CID fragmentation. Glycans are central to fundamental functions across biology, from cell–cell recognition and receptor signaling to viral entry, mucosal barrier defense, and the fine-tuning of antibody effector function. Assembled without a genetic template by competing glycosyltransferases and glycosidases, two classes dominate mammalian glycoproteins: asparagine-linked N-glycans built around a conserved chitobiose-trimannosyl core and serine/threonine-linked O-glycans built on mucin cores. In liquid chromatography–tandem mass spectrometry (LC-MS/MS), glycans separated by retention time are isolated at a precursor m/z and fragmented by low-energy collision-induced dissociation (CID). In the nomenclature of Domon and Costello [1988], low-energy CID predominantly cleaves glycosidic bonds (B- and Y-ions) whose $\Delta m/z$ intervals identify lost monosaccharides (Figures 1 and 2), whereas higher-energy cross-ring cleavages (A- and X-ions) needed to separate regioisomers are weak or absent, limiting single-scan isomer resolution.

Glycan representations. Glycan structures are depicted in Symbol Nomenclature for Glycans (SNFG; Varki et al., 2015, Neelamegham et al., 2019) and serialized as condensed IUPAC strings where parentheses specify glycosidic linkages and brackets enclose branches (Figure 1). Because sibling branches lack a unique linear order, distinct strings can encode the same molecular graph (Figure 6a; Appendix D), motivating graph-isomorphism scoring in `glycowork` [Thomès et al., 2021]. Public repositories catalog these structures and spectra across GlyTouCan [Tiemeyer et al., 2017, Fujita et al., 2021], GlyConnect [Alocci et al., 2019], GlycoPOST [Watanabe et al., 2021], and the 500,000-spectrum CandyCrunch corpus [Urban et al., 2024, Abtheen et al., 2025a].

2.2 Prior Work

Discrete Flow Matching and masked diffusion. Continuous diffusion [Ho et al., 2020] and Flow Matching [Lipman et al., 2023] learn velocity fields transporting a simple source to the data distribution. Campbell et al. [2022, 2024] formulated continuous-time categorical flows via CTMCs, and Gat et al. [2024] generalized them into Discrete Flow Matching across arbitrary sources, couplings, and schedules. Related absorbing-state masked diffusion models—D3PM [Austin et al., 2021], SEDD [Lou et al., 2024], MDLM [Sahoo et al., 2024], and MD4 [Shi et al., 2024]—match autoregressive language models at scale [Nie et al., 2025] and outperform them under data scarcity [Prabhudesai et al., 2025]. Unlike simplex-native training relaxations [Dieleman et al., 2022, Davis et al., 2024, Stark et al., 2024], Crystallizing Flows integrate on the simplex purely at inference time atop discretely trained checkpoints (Appendix H).

Inference-time acceleration. Continuous diffusion acceleration uses higher-order ODE solvers [Song et al., 2021, Lu et al., 2022, Karras et al., 2022], distillation [Salimans and Ho, 2022, Song et al., 2023, Luo et al., 2023, Meng et al., 2023, Sauer et al., 2024, Yin et al., 2024, Lacombe and Vaidya, 2024], and rectified flows [Liu et al., 2023]. Discrete unmasking has relied on heuristic confidence schedules such as Mask-Predict [Ghazvininejad et al., 2019] and MaskGIT [Chang et al., 2022] or stochastic CTMC stepping [Sahoo et al., 2024, Gat et al., 2024]; Crystallizing Flows instead pair simplex propagation with a parameter-free argmax commit rule.

Computational mass spectrometry. Classical glycan tools match fragments against rules or curated libraries, including GlycoWorkbench [Ceroni et al., 2008], GlycoDeNovo [Hong et al., 2017], pGlyco3 [Zeng et al., 2021], and StrucGP [Shen et al., 2021]. Neural decoders such as Casanovo [Yilmaz et al., 2022, 2024], InstaNovo [Eloff et al., 2025], and DiffMS [Bohde et al., 2025] predict peptides and small molecules directly from MS/MS spectra. In glycomics, CandyCrunch [Urban et al., 2024] classifies over a fixed 3,500-glycan library, leaving 207M autoregressive GlycoBART [Abtheen et al., 2025b, Lewis et al., 2020] as the sole open-vocabulary neural baseline. Compact IUPAC vocabularies, short branched trees (10–50 tokens), and millisecond scan budgets make glycan MS/MS well suited to parallel flow matching.

3 GlycoGen and Crystallizing Flows

3.1 Mask-Source Discrete Flow Matching

Discrete Flow Matching [Gat et al., 2024] defines a probability path $\{p_t\}_{t\in[0,1]}$ between a tractable source p_0 and the data distribution $p_1 = q$, then trains a denoiser $p^\theta_{1|t}$ whose generating velocity transports samples along that path. We use the mask source $p(x) = \delta_m(x)$ with the linear schedule $\kappa_t = t$, which yields the per-position mixture path

$$p_t(x^i \mid x_0, x_1) = (1-t)\,\delta_m(x^i) + t\,\delta_{x_1^i}(x^i), \tag{1}$$

the masked cross-entropy training loss

$$\mathcal{L}(\theta) \;=\; -\,\mathbb{E}_{t,\,X_1,\,X_t} \sum_{i\in\mathcal{M}_t} \log p^\theta_{1|t}(X_1^i \mid X_t, \mathbf{h}_{\text{spec}}), \quad t\sim\mathcal{U}(0,1), \tag{2}$$

and the closed-form per-position velocity [Gat et al., 2024]

$$u_t^i(x,z) \;=\; \tfrac{1}{1-t}\big[\,p^\theta_{1|t}(x \mid z, \mathbf{h}_{\text{spec}}) - \delta_{z^i}(x)\,\big]. \tag{3}$$

The standard CTMC Euler sampler of Gat et al. [2024], reproduced in Appendix A, integrates this velocity stochastically over a uniform grid of K steps with step size $h = 1/K$. At each step, the state is a discrete token sequence, and each masked position commits to a sampled token with probability $h/(1-t_k)$ and stays frozen thereafter. Because every commit collapses the softmax distribution to a categorical draw and cannot be revised, early errors condition later steps.

Crucially, the training objective in Equation (2) leaves the inference sampler unspecified: any procedure that integrates the velocity field from the all-mask source at $t = 0$ to a clean sequence at $t = 1$ is a valid decoder. This decoupling between training and inference has motivated dedicated research into discrete unmasking schedules and samplers [Ghazvininejad et al., 2019, Chang et al., 2022, Sahoo et al., 2024, Gat et al., 2024]; we build on it directly in Section 3.3.

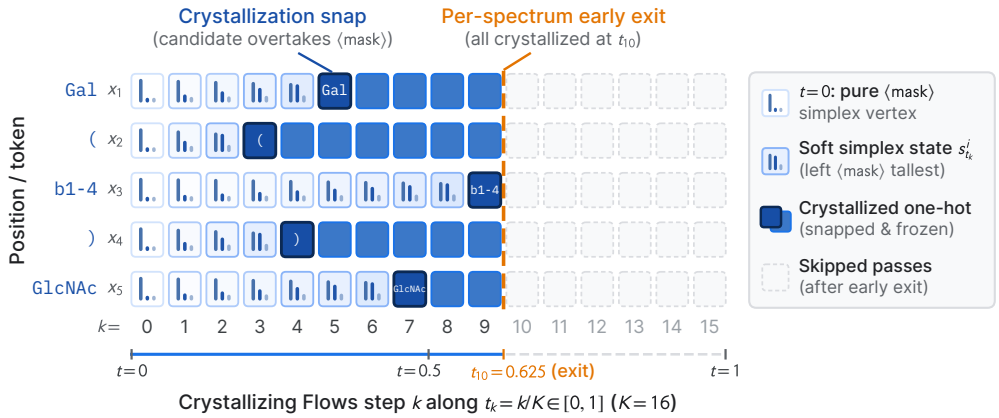


Figure 3: **Crystallizing Flows mechanic.** Per-position state evolution across 16 Crystallizing Flows steps on `Gal(b1-4)GlcNAc`. **Light-blue cells** show soft simplex distributions $\mathbf{s}^i_{t_k} \in \Delta^{V-1}$ whose tallest left bar remains `<mask>` while the candidate bar rises prior to commit. Once the candidate overtakes `<mask>`, the position snaps to a **deep-blue one-hot** token and stays frozen thereafter. Decoding exits early at the **orange dashed line** ($t_{10} = 0.625$) once all positions crystallize.

3.2 GlycoGen Architecture

Two architectural choices separate GlycoGen from prior glycan structure predictors. First, GlycoGen is **open-vocabulary**: rather than classifying over a fixed glycan dictionary [Urban et al., 2024] or predicting compressed antenna motifs whose reconstruction loses branch-attachment accuracy and relies on database snap-back [Abtheen et al., 2025b], GlycoGen decodes directly over a 74-token atomic IUPAC vocabulary (70 monosaccharide, linkage such as `a1-3` or `b1-4`, and branch-bracket `[/]` tokens, plus `<pad>`, `<mask>`, `<bos>`, and `<eos>`). Any chemically valid branched glycan tree in this alphabet can be generated directly. Second, GlycoGen is **non-autoregressive**: whereas causal decoders require $L \approx 33$ sequential forward passes per beam step ($33 \times B$ passes at beam width B), GlycoGen applies full bidirectional self-attention across all $L = 96$ sequence slots simultaneously, refining the entire branched tree in parallel in $K \ll L$ passes (down to $K = 1$).

Each MS/MS spectrum records a variable-length list of fragment-ion $(m/z, \text{intensity})$ peaks produced when the instrument fragments an isolated precursor ion, where low-intensity diagnostic ions often resolve branching or linkage isomers. Rather than truncating to a top-N peak list, we map each spectrum onto a fixed 2,048-bin mass-to-charge grid storing two normalized channels per bin—log-scaled peak intensity and continuous sub-bin mass offset—preserving full spectral coverage without variable-length padding while retaining sub-bin mass precision (Figure 5, Appendix B). The spectrum encoder reshapes these 2,048 dual-channel bins into 64 contiguous patches of 32 bins (64 features per patch), linearly projects each patch to $d = 512$, appends one standardized precursor-m/z token projected to $d = 512$, adds learned positional embeddings, and passes the 65-token sequence through a 4-layer pre-norm GELU Transformer encoder [Vaswani et al., 2017] (8 attention heads, feed-forward width 2,048) to compute $\mathbf{h}_{\text{spec}} \in \mathbb{R}^{65\times512}$ once per spectrum.

The structure decoder embeds the partially masked glycan sequence of length $L = 96$ via token embedding matrix $E \in \mathbb{R}^{74\times512}$ and learned positional embeddings, adds a single linear broadcast embedding of continuous time $t \in [0, 1]$ across all positions (avoiding the adaptive layer-norm overhead of DiT; Peebles and Xie, 2023), cross-attends to $\mathbf{h}_{\text{spec}}$ across an 8-layer bidirectional pre-norm GELU Transformer decoder ($d = 512$, 8 heads, feed-forward width 2,048), and outputs per-position logits over all 74 tokens in parallel.

3.3 Crystallizing Flows Sampler

At inference time, what state we pass from one Euler step to the next determines how much of the denoiser's predictive distribution survives the rollout. Whereas CTMC Euler discards each step's vocabulary distribution in favor of a random categorical draw and commits masked positions by coin flip (Section 3.1), we **propagate each position's full probability distribution on the simplex** and

Algorithm 1 Crystallizing Flows inference.

Require: denoiser $p^{\theta}_{1|t}$, token embedding matrix $E \in \mathbb{R}^{V\times d}$, spectrum embedding $\mathbf{h}_{\text{spec}}$, step budget K

Initialize $\mathbf{s}^i \leftarrow \delta_m$ for non-special $i \in \{1,\dots,L\}$ and one-hot specials on $\mathcal{C} \leftarrow \{\text{special positions}\}$
$t \leftarrow 0$; $\quad h \leftarrow 1/K$
for $k = 0,\dots,K-1$ **do**
  $\mathbf{p} \leftarrow p^{\theta}_{1|t}(\cdot \mid \mathbf{s}E, \mathbf{h}_{\text{spec}})$ ▷ 1 parallel forward pass; $\mathbf{p}^i \in \Delta^{V-1}$
  for each uncrystallized position $i \notin \mathcal{C}$ **do**
    $\mathbf{s}^i \leftarrow \left(1 - \frac{h}{1-t}\right)\mathbf{s}^i + \frac{h}{1-t}\,\mathbf{p}^i$ ▷ simplex Euler step ($i \in \mathcal{C}$ stay frozen)
    if $v^\star \leftarrow \arg\max_v \mathbf{s}^i(v) \neq m$ **then**
      $\mathbf{s}^i \leftarrow \delta_{v^\star}$; $\quad \mathcal{C} \leftarrow \mathcal{C} \cup \{i\}$ ▷ crystallize to one-hot & freeze
    end if
  end for
  if $\mathcal{C} = \{1,\dots,L\}$ **then**
    break ▷ per-spectrum early exit
  end if
  $t \leftarrow t + h$
end for
return $X^i = \arg\max_v\ \mathbf{s}^i(v)$ for $i = 1,\dots,L$

let the model itself decide when to commit. We call this deterministic sampler **Crystallizing Flows** (Figure 3): each undecided position maintains a probability vector over the vocabulary and snaps to a one-hot token the moment its argmax leaves `<mask>`, preserving calibrated uncertainty early in the trajectory while eliminating low-step sampling variance.

Deterministic integration on the simplex. A denoiser trained on discrete one-hot tokens accepts continuous probability vectors at inference time without retraining: multiplying row distributions $\mathbf{s} \in (\Delta^{V-1})^L \subset \mathbb{R}^{L\times V}$ by the token embedding matrix $E \in \mathbb{R}^{V\times d}$ yields mixture embeddings $\mathbf{s}E \in \mathbb{R}^{L\times d}$ that reduce to exact token lookups at every one-hot position. We initialize every non-special position $i \notin \mathcal{C}$ to the one-hot mask distribution $\mathbf{s}^i_{t_0} = \delta_m \in \Delta^{V-1}$ (holding special tokens fixed in $\mathcal{C}$). At step k along $t_k = k/K$, the decoder receives $\mathbf{s}_{t_k}E$, and each uncrystallized position steps forward along the velocity field (Equation (3)):

$$\mathbf{s}^i_{t_{k+1}} \;=\; \left(1 - \tfrac{h}{1-t_k}\right)\mathbf{s}^i_{t_k} \;+\; \tfrac{h}{1-t_k}\,p^{\theta}_{1|t_k}(\cdot \mid \mathbf{s}_{t_k}E, \mathbf{h}_{\text{spec}})^i, \quad i \notin \mathcal{C}. \tag{4}$$

Position i crystallizes as soon as its maximum-probability token leaves the mask, $v^\star = \arg\max_v \mathbf{s}^i_{t_{k+1}}(v) \neq m$: we set $\mathbf{s}^i_{t_{k+1}} \leftarrow \delta_{v^\star}$, add i to the frozen set $\mathcal{C}$, and leave it unchanged on all remaining steps. Subsequent passes therefore attend to crisp one-hot embeddings at crystallized positions and soft mixture embeddings at positions still being refined. Once $\mathcal{C}$ covers all L positions of a spectrum, decoding exits early for that spectrum (Algorithm 1).

Appendix C details why distribution propagation, determinism, clean conditioning at decided positions, and midpoint early exit each contribute to the accuracy gain.

4 Experiments

Dataset, benchmark split, and per-spectrum evaluation. All models train on CandyCrunch v2 [Urban et al., 2024, Abtheen et al., 2025a] (503,446 annotated LC-MS/MS spectra), filtered to the 497,751 rows (98.9%) representable in GlycoBART's antenna notation. Even after CandyCrunch v2 merges consecutive scans within an elution peak, a single LC-MS acquisition records repeated spectra of the same glycan across co-eluting scans, charge states, adduct ions, and separated anomer peaks. Spectrum-level random splitting scatters roughly 85% of these near-duplicate scans into training and 15% into testing, turning structure prediction into chromatographic-peak retrieval and inflating reported accuracy—the regime used in Abtheen et al. [2025b]. We instead construct a leak-free **CandyCrunch v2 benchmark** split by partitioning entire LC-MS files by source filename into 420,977 training and 76,774 test spectra (holding out 10% of training for validation), and retrain GlycoBART from scratch on this split.

At test time, whereas offline sample-level pipelines aggregate 5 to 50 scans per chromatographic peak, we evaluate single-spectrum prediction directly in both batched and serial streaming regimes, predicting one IUPAC structure per spectrum and scoring exact graph isomorphism via `glycowork` [Thomès

Model (operating point)	NFE	ms/spec	spec/s	Accuracy (3 seeds)	Speedup
GlycoGen-DFM + Crystallizing Flows (ours), 50*M*					
K=1 (one-pass)	1	0.046	21,883	$\mathbf{0.9019 \pm 0.0004}$	$\mathbf{61}\times$
K=8 (recommended)	5	0.201	4,975	$\mathbf{0.9251 \pm 0.0006}$	$\mathbf{13.9}\times$
K=32	17	0.675	1,481	$\mathbf{0.9267 \pm 0.0005}$	$\mathbf{4.1}\times$
K=64	34	1.306	766	$\mathbf{0.9269 \pm 0.0006}$	$\mathbf{2.1}\times$
GlycoGen-MDLM + Crystallizing Flows (ours, cross-objective control), 50*M*					
K=1 (one-pass)	1	0.046	21,739	0.8973 ± 0.0017	$61\times$
K=8	5	0.199	5,025	0.9230 ± 0.0012	$14.1\times$
K=32	17	0.676	1,479	0.9247 ± 0.0011	$4.1\times$
GlycoGen-AR (matched-architecture autoregressive control), 50*M*					
beam-1 (top-1)	33	2.21	452	0.9038 ± 0.0005	$1.3\times$
beam-32 (top-1, seed 0)	1,045	72.66	14	0.9039	$0.04\times$
GlycoBART reproduction [Abtheen et al., 2025b], 207*M (per-spectrum)*					
beam-1 (top-1)	23	2.80	357	0.8092 ± 0.0031	$1.0\times$
beam-8 (top-5, seed 0)	180	27.12	37	0.8578	$0.10\times$
beam-32 (top-5, seed 0)	720	91.21	11	0.8635	$0.03\times$

Table 1: **Headline results.** Top-1 per-spectrum structural accuracy on the 76,774 test spectra of the CandyCrunch v2 benchmark (mean $\pm$ SD over three seeds on one H200 GPU in BF16; full sweep in Table 2; high-beam sweeps on seed 0 as top-1 is flat on AR and drops above beam 1 on GlycoBART). Batched speedup compares against greedy GlycoBART (2.80 ms/spec $= 1.0\times$); relative to beam-32 (91.21 ms/spec), 8-step GlycoGen is $454\times$ faster. Unbatched serial streaming latencies appear in Table 7.

et al., 2021] without post-processing (no multi-scan consensus voting, charge-state merging, or reference-database snap-back). In Abtheen et al. [2025b], combining random splitting with multi-scan voting and database snap-back lifts GlycoBART top-1 accuracy to 89.1%, gaining 5–15 percentage points from cross-scan aggregation unavailable live (Appendix E).

Training and reproducibility. We train all headline models on a single 141 GB H200 GPU in BF16 precision. GlycoGen-DFM, GlycoGen-MDLM, and GlycoGen-AR share the exact same optimization recipe: AdamW [Loshchilov and Hutter, 2019] at batch size 1024, base learning rate 3×10^{-4} scaled by $\sqrt{8}$ to 8.49×10^{-4} over 30 epochs with 200-step linear warmup and cosine decay to 3% of peak, weight decay 0.01, and gradient clipping at 1.0. GlycoBART follows Abtheen et al. [2025b] at batch size 160 ($\sqrt{2}$-scaled learning rate; Appendix E). We selected the GlycoGen architecture and schedule via a two-stage **Distillate** automated research exploration [Lacombe, 2026] (single-GPU sweeps followed by 21 runs on A100 and L40S GPUs scaling top candidates to 50M parameters on a CandyCrunch subset); GlycoGen-MDLM and GlycoGen-AR reuse this recipe unchanged.

Models and baselines. We compare four models on identical splits ($K \in \{1, 2, 4, 8, 16, 32, 64\}$ for flow/diffusion; beam widths 1–32 for autoregressive baselines):

- **GlycoGen-DFM + Crystallizing Flows:** our 50M-parameter encoder-decoder DFM model (Section 3.2), trained with uniform continuous time sampling and a linear schedule.
- **GlycoGen-MDLM + Crystallizing Flows:** a matched 50M MDLM denoiser [Sahoo et al., 2024] trained on a 128-step grid with NELBO cross-entropy (Appendix G).
- **GlycoGen-AR:** a matched 46.4M causal autoregressive control trained with teacher-forced cross-entropy at beams 1–32 ($\sim$33 passes/beam; Appendix F).
- **GlycoBART reproduction:** the 207M autoregressive model of Abtheen et al. [2025b], retrained on our CandyCrunch v2 benchmark split. Unlike GlycoGen's atomic IUPAC vocabulary, GlycoBART predicts compressed antenna motifs whose reconstruction back to full IUPAC trees loses 10.3 percentage points of top-1 accuracy from branch-attachment ambiguities (Appendix E) and excludes curly-brace strings (`{...}`, 1.1% of CandyCrunch v2).

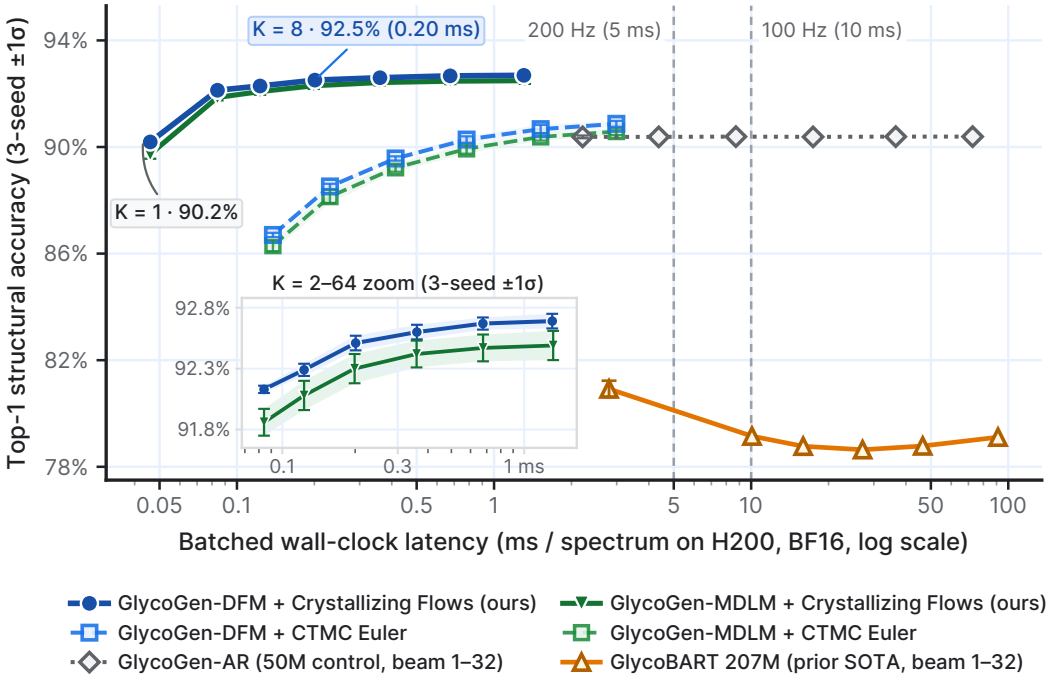


Figure 4: **Crystallizing Flows Pareto-dominate at low step counts on both accuracy and wall-clock cost.** Top-1 per-spectrum structural accuracy on the 76,774 test spectra of the CandyCrunch v2 benchmark (mean $\pm 1\sigma$ over three seeds; inset zooms into K=2–64) versus batched latency on one H200 GPU. DFM with Crystallizing Flows (**deep blue**, ours) and MDLM with Crystallizing Flows (**green**, ours) Pareto-dominate CTMC Euler, 50M GlycoGen-AR, and 207M GlycoBART [Abtheen et al., 2025b]. Dashed lines mark 200 Hz (5 ms) and 100 Hz (10 ms) scan intervals (Table 7).

5 Results

Headline accuracy and ablation of design choices. Across all 76,774 test spectra in the CandyCrunch v2 benchmark, single-pass GlycoGen-DFM reaches **90.19%** top-1 structural accuracy, already surpassing every autoregressive baseline. Refining predictions with Crystallizing Flows lifts accuracy to **92.51%** at 5 effective passes (K=8), **92.60%** at 9 passes (K=16), **92.67%** at 17 passes (K=32), and **92.69%** at 34 passes (K=64) across three seeds (Table 1, Figure 4). Comparing controlled baselines isolates where this 11.7 percentage point improvement over greedy GlycoBART (80.92%) comes from:

1. **Full-spectrum patch encoding and direct IUPAC tokenization (+9.5 pp).** Instead of truncating spectra to top-N discrete peak tokens and predicting lossy antenna strings, encoding the full spectrum as 64 patches of binned intensities and sub-bin mass offsets while generating atomic IUPAC tokens directly lifts GlycoGen-AR from 80.92% to 90.38%, preserving weak diagnostic fragments and eliminating the $\sim$9% antenna loss.
2. **Parallel decoding with flow matching and masked diffusion (+2.1 pp).** Decoding all sequence positions simultaneously rather than one token at a time from left to right lets every position attend bidirectionally to both sides of a branched glycan tree as structure resolves, lifting accuracy from 90.38% on GlycoGen-AR to 92.51% on GlycoGen-DFM at K=8.
3. **Crystallizing Flows (+1.8 to +5.6 pp over Euler).** Propagating full token probability distributions on the simplex instead of drawing random categorical commits adds 5.43 percentage points over standard CTMC Euler at K=2 (5.56 pp on MDLM) and 1.82 percentage points at K=64 (1.91 pp on MDLM).

Inference speed. Decoding all positions in parallel rather than stepping through $\sim$33 autoregressive passes per beam makes GlycoGen one to two orders of magnitude faster than GlycoBART in both batch processing and serial streaming. We benchmarked all models on an identical H200 VM on

Hugging Face Jobs (Tables 1 and 7); as absolute timings vary across GPU generations and serving stacks, these numbers provide one controlled reference point. On this setup, batched GlycoGen runs $61\times$ faster than greedy GlycoBART at $K{=}1$ and $13.9\times$ faster at $K{=}8$ ($454\times$ faster than beam-32), turning multi-day repository runs into minutes. In unbatched serial streaming, end-to-end latency including CPU peak binning drops from 183 ms/spec for greedy GlycoBART to 8.6 ms at $K{=}1$ (116 Hz, matching 100 Hz scan cadence) and 33.8 ms at $K{=}8$.

Comparison with alternative samplers. Crystallizing Flows outperform standard CTMC Euler at every step count from $K{=}2$ to $K{=}64$ (Figure 4). At $K{=}2$, CTMC Euler drops to 86.70%—3.5 percentage points below a single pass—as random categorical commits lock in early errors. Across seven other discrete and simplex samplers on the same checkpoint (temperature-scaled, predictor-corrector, pure soft-propagation, and confidence-ordered unmasking), Crystallizing Flows outperform every alternative at low step budgets, with confidence unmasking matching accuracy only above 33 passes (Appendix H).

Cross-instrument robustness and generalization to unseen IUPAC spellings. Mass spectrometers vary widely in collision energetics and mass resolution, yet single-pass GlycoGen maintains 88.3% to 93.0% top-1 accuracy across all four analyzer families in CandyCrunch v2 (linear ion traps, Orbitraps, Bruker amaZon, and other analyzers; Table 4).

We also tested whether GlycoGen learns chemical graph structure from string supervision or merely memorizes literal training sequences. Because a glycan is a branched tree rather than a linear chain, a single molecule can be written as several distinct valid IUPAC strings depending on which branch is serialized first (Figure 6, Appendix D). Grouping the 76,774 test spectra by chemical graph isomorphism via `glycowork` [Thomès et al., 2021] yields 76,203 spectra matching a training string verbatim, 66 spectra whose true glycan graph appears in training *only* under a different branch ordering, and 505 spectra with completely unseen glycan topologies (Table 3). On those 66 unseen-spelling spectra, GlycoBART drops to 13.6%, failing to recognize known molecules written in a new branch order. Direct IUPAC tokenization lifts GlycoGen-AR to 54.5%, and parallel bidirectional decoding lifts GlycoGen further to **60.6%** (DFM) and **62.1%** (MDLM)—clear evidence that **GlycoGen generalizes across unseen string representations of known chemical graphs**.

Transfer to masked diffusion, early exit, and autoregressive beam saturation. Applied without retraining as a drop-in sampler for our 128-step MDLM denoiser [Sahoo et al., 2024], Crystallizing Flows improve over MDLM's CTMC Euler sampler by 5.56 percentage points at $K{=}2$ and 1.91 percentage points at $K{=}64$, tracking GlycoGen-DFM within 0.21 percentage points from $K{=}4$ upward (Appendix G). Meanwhile, per-spectrum early exit cuts neural forward passes nearly in half (17.4 passes at $K{=}32$ and 33.9 at $K{=}64$). Widening beam search on GlycoGen-AR from beam 1 to beam 32 leaves top-1 accuracy flat at $90.38\% \pm 0.05\%$ despite a $33\times$ latency penalty, showing that wider left-to-right search cannot substitute for iterative parallel refinement.

What explains the remaining errors? On the 76,203 test spectra whose target structure appears in training, GlycoGen reaches 93.31% exact graph accuracy (Table 3), and 99.1% of its 5,627 mismatched predictions at $K{=}32$ are syntactically valid IUPAC strings (over 92% parse into complete chemical graphs in `glycowork`). Roughly four in five valid mismatches reflect physical limits of single-spectrum CID rather than hallucinations (Table 5, Figure 6): 40–42% match monosaccharide composition and differ only in regioisomer linkage (β1-3 vs. β1-4) or linkage wildcard specificity (`a2-?`), which share glycosidic cleavage masses absent cross-ring ions, while $\sim$38% differ by one labile fucose or sialic acid lost in-source before MS/MS, approaching the Bayes error floor of low-energy CID.

6 Limitations and Future Work

6.1 Limitations

Generalization to novel glycan topologies. On the 505 test spectra whose chemical graph never appears in training, all four models score near zero (0.2%–3.6%; Table 3). First, CandyCrunch v2 concentrates roughly 500,000 spectra onto 3,590 unique glycan graphs, limiting structural breadth.

Second, fixed-m/z bin encoding favors whole-spectrum pattern recognition over translation-invariant $\Delta m/z$ reasoning when unfamiliar branches or reducing-end tags shift fragments across bins.

Zero-shot transfer across mass analyzer platforms. While GlycoGen reaches 88.3%–93.0% accuracy across all four analyzer families when present in training, holding out an entire family lowers zero-shot accuracy to 41.4% on linear ion traps and 14.1% on Orbitraps (Table 4); distinct collision energetics mean new platforms (e.g., timsTOF, Orbitrap Astral) require fine-tuning.

Notation scope and crystallization calibration. For direct comparison with GlycoBART, we exclude the 1.1% of CandyCrunch v2 spectra with curly-brace substituents (`{...}`), though atomic IUPAC tokens represent them directly. Under $\kappa_t = t$, positions crystallize at mean probability 0.56 ± 0.06 at K=8 (Appendix D), calibrated to linear interpolation; delayed schedules can use a threshold.

6.2 Future Work

Generalizing to novel glycan structures. Moving from recognition of known topologies to true *de novo* discovery requires building fragmentation physics into both encoder and training data. Replacing fixed-m/z patch embeddings with pairwise peak-difference ($\Delta m/z$) attention makes neutral-loss reasoning invariant to precursor mass shifts and reducing-end tags, while pretraining on simulated fragmentation trees across GlyTouCan [Tiemeyer et al., 2017] can expose the model to tens of thousands of topologies absent from current benchmarks. Pairing compositional pretraining with calibrated simplex probabilities would also let the decoder emit explicit linkage wildcards (`a2-?`) when CID leaves regioisomers underdetermined.

Illuminating the dark glycome across large-scale experimental datasets. Public glycomics and glycoproteomics archives hold tens of millions of raw LC-MS/MS spectra across tissues, species, and disease cohorts that remain unannotated—a "dark glycome" unexplored for lack of a method both fast enough for repository throughput and general enough for open-vocabulary prediction. Sub-millisecond batched decoding with direct IUPAC generation makes exhaustive reanalysis practical on one GPU, turning unannotated repositories into systematic surveys of glycosylation diversity across species and disease.

7 Conclusion

We introduced **GlycoGen** and **Crystallizing Flows** for open-vocabulary glycan structure prediction from tandem mass spectra. Across the CandyCrunch v2 benchmark, GlycoGen reaches 92.67% top-1 accuracy—93.31% on seen structures, approaching the single-spectrum CID Bayes error ceiling—while generalizing to unseen IUPAC branch orderings, transferring without retraining to masked diffusion, and accelerating inference by one to two orders of magnitude. Methodologically, these results show that *de novo* elucidation of branched molecules is poorly served by left-to-right token search: a fragment spectrum probes the entire glycan tree at once, and decoding all positions simultaneously on the probability simplex is both faster and more faithful to the underlying chemistry. Biologically, whereas genomics and proteomics scaled once sequences became routine to read, glycomics has remained bottlenecked by slow, library-bound spectrum interpretation. Making the glycome readable directly from mass spectra in real time and at scale removes a central barrier to mapping how glycosylation governs cellular signaling, immunity, and human health.

Code and data availability. Open-source code, trained three-seed checkpoints, and our CandyCrunch v2 benchmark splits are available at `https://github.com/rlacombe/GlycoGen` and `https://huggingface.co/datasets/rlacombe/candycrunch-v2-benchmark`.

Acknowledgments and Disclosure of Funding

The author thanks Emilia McLaughlin and Marina Lochhead (Kyron Bio) for invaluable discussions on experimental glycobiology workflows that motivated this work, the CandyCrunch and GlycoBART authors for open-sourcing the datasets and pipelines on which it builds, and the maintainers of `glycowork` for the graph-isomorphism evaluation library used throughout. This research was independently self-funded by the author, including LLM token budgets for Distillate auto-research explorations, as well as training and evaluation runs on Hugging Face Jobs GPUs.

References

Ejas Althaf Abtheen, Sriram Neelamegham, and Rudiyanto Gunawan. Transformer-based deep learning for glycan structure inference from tandem mass spectrometry [Dataset]. Zenodo, `https://zenodo.org/records/15741423`, 2025a. DOI: `https://doi.org/10.5281/zenodo.15741423`.

Ejas Althaf Abtheen, Arun Singh, Shyam Sriram, Changyou Chen, Sriram Neelamegham, and Rudiyanto Gunawan. Transformer-based deep learning for glycan structure inference from tandem mass spectrometry. *bioRxiv*, 2025b. doi: 10.1101/2025.07.02.662857.

Davide Alocci, Julien Mariethoz, Alessandra Gastaldello, Elisabeth Gasteiger, Niclas G. Karlsson, Daniel Kolarich, Nicolle H. Packer, and Frédérique Lisacek. GlyConnect: Glycoproteomics goes visual, interactive, and analytical. *Journal of Proteome Research*, 18(2):664–677, 2019. doi: 10.1021/acs.jproteome.8b00766.

Gad Armony, Sven Brehmer, Tharan Srikumar, Lennard Pfennig, Fokje Zijlstra, Dennis Trede, Gary Kruppa, Dirk J. Lefeber, Alain J. van Gool, and Hans J. C. T. Wessels. The GlycoPaSER prototype as a real-time N-glycopeptide identification tool based on the PaSER parallel computing platform. *International Journal of Molecular Sciences*, 24(9):7869, 2023. doi: 10.3390/ijms24097869.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 17981–17993, 2021.

Montgomery Bohde, Mrunali Manjrekar, Runzhong Wang, Shuiwang Ji, and Connor W. Coley. DiffMS: Diffusion generation of molecules conditioned on mass spectra. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 4737–4756, 2025.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 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 *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 5453–5512, 2024.

Alessio Ceroni, Kai Maass, Hildegard Geyer, Rudolf Geyer, Anne Dell, and Stuart M. Haslam. GlycoWorkbench: A tool for the computer-assisted annotation of mass spectra of glycans. *Journal of Proteome Research*, 7(4):1650–1659, 2008. doi: 10.1021/pr7008252.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11315–11325, 2022.

Eldan Cohen and J. Christopher Beck. Empirical analysis of beam search performance degradation in neural sequence models. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 1290–1299, 2019.

Oscar Davis, Samuel Kessler, Mircea Petrache, İsmail İlkan Ceylan, Michael Bronstein, and Avishek Joey Bose. Fisher flow matching for generative modeling over discrete data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 139054–139084, 2024.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Bruno Domon and Catherine E. Costello. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate Journal*, 5(4):397–409, 1988. doi: 10.1007/BF01049915.

Kevin Eloff, Konstantinos Kalogeropoulos, Amandla Mabona, Oliver Morell, Rachel Catzel, Esperanza Rivera-de-Torre, Jakob Berg Jespersen, Wesley Williams, Sam P. B. van Beljouw, Marcin J. Skwark, Andreas Hougaard Laustsen, Stan J. J. Brouns, Anne Ljungars, Erwin M. Schoof, Jeroen Van Goey, Ulrich auf dem Keller, Karim Beguir, Nicolas Lopez Carranza, and Timothy P. Jenkins. InstaNovo enables diffusion-powered de novo peptide sequencing in large-scale proteomics experiments. *Nature Machine Intelligence*, 7(4):565–579, 2025. doi: 10.1038/s42256-025-01019-5.

Akihiro Fujita, Nobuyuki P. Aoki, Daisuke Shinmachi, Masaaki Matsubara, Shinichiro Tsuchiya, Masaaki Shiota, Tamiko Ono, Issaku Yamada, and Kiyoko F. Aoki-Kinoshita. The international glycan repository GlyTouCan version 3.0. *Nucleic Acids Research*, 49(D1):D1529–D1533, 2021. doi: 10.1093/nar/gkaa947.

Benjamin Furtwängler, Nil Üresin, Khatereh Motamedchaboki, Romain Huguet, Daniel Lopez-Ferrer, Vlad Zabrouskov, Bo T. Porse, and Erwin M. Schoof. Real-time search-assisted acquisition on a Tribrid mass spectrometer improves coverage in multiplexed single-cell proteomics. *Molecular & Cellular Proteomics*, 21(4):100219, 2022. doi: 10.1016/j.mcpro.2022.100219.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 133345–133385, 2024.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-Predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121. Association for Computational Linguistics, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

Pengyu Hong, Hui Sun, Long Sha, Yi Pu, Kshitij Khatri, Xiang Yu, Yang Tang, and Cheng Lin. GlycoDeNovo – an efficient algorithm for accurate de novo glycan topology reconstruction from tandem mass spectra. *Journal of the American Society for Mass Spectrometry*, 28(11):2288–2301, 2017. doi: 10.1007/s13361-017-1760-6.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 26565–26577, 2022.

Romain Lacombe. Distillate: A meta-agent harness for autonomous ML research. Working paper, 2026. URL `https://distillate.dev/paper.pdf`.

Romain Lacombe and Neal Vaidya. Accelerating the generation of molecular conformations with progressive distillation of equivariant latent diffusion models. *arXiv preprint arXiv:2404.13491*, 2024. Generative and Experimental Perspectives for Biomolecular Design Workshop, ICLR 2024.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics (ACL), pages 7871–7880. Association for Computational Linguistics, 2020.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 32819–32848, 2024.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 5775–5787, 2022.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14297–14306, 2023.

Sriram Neelamegham, Kiyoko Aoki-Kinoshita, Evan Bolton, Martin Frank, Frédérique Lisacek, Thomas Lütteke, Noel O'Boyle, Nicolle H. Packer, Pamela Stanley, Philip Toukach, Ajit Varki, Robert J. Woods, and The SNFG Discussion Group. Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology*, 29(9):620–624, 2019. doi: 10.1093/glycob/cwz045.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pages 50608–50646, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023.

Mihir Prabhudesai, Mengning Wu, Amir Zadeh, Katerina Fragkiadaki, and Deepak Pathak. Diffusion beats autoregressive in data-constrained settings. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pages 10581–10606, 2025.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 130136–130184, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision (ECCV)*, pages 87–103, 2024.

Devin K. Schweppe, Jimmy K. Eng, Qing Yu, Derek Bailey, Ramin Rad, José Navarrete-Perea, Edward L. Huttlin, Brian K. Erickson, João A. Paulo, and Steven P. Gygi. Full-featured, real-time database searching platform enables fast and accurate multiplexed quantitative proteomics. *Journal of Proteome Research*, 19(5):2026–2034, 2020. doi: 10.1021/acs.jproteome.9b00860.

Jiechen Shen, Li Jia, Liuyi Dang, Yuanjie Su, Jie Zhang, Yintai Xu, Bojing Zhu, Zexuan Chen, Jingyu Wu, Rongxia Lan, Zhifang Hao, Chen Ma, Ting Zhao, Ni Gao, Jieyun Bai, Yuan Zhi, Jun Li, Junying Zhang, and Shisheng Sun. StrucGP: De novo structural sequencing of site-specific N-glycan on glycoproteins using a modularization strategy. *Nature Methods*, 18(8):921–929, 2021. doi: 10.1038/s41592-021-01209-0.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 103131–103167, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252, 2023.

Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3356–3362. Association for Computational Linguistics, 2019.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to DNA sequence design. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 46495–46513, 2024.

Emmajay Sutherland, Tim S. Veth, William D. Barshop, Jacob H. Russell, Kathryn Kothlow, Jesse D. Canterbury, Christopher Mullen, David Bergen, Jingjing Huang, Vlad Zabrouskov, Romain Huguet, Graeme C. McAlister, and Nicholas M. Riley. Autonomous dissociation-type selection for glycoproteomics using a real-time library search. *Journal of Proteome Research*, 23(12):5606–5614, 2024. doi: 10.1021/acs.jproteome.4c00723.

Luc Thomès, Rebekka Burkholz, and Daniel Bojar. Glycowork: A Python package for glycan data science and machine learning. *Glycobiology*, 31(10):1240–1244, 2021. doi: 10.1093/glycob/cwab067.

Michael Tiemeyer, Kazuhiro Aoki, James Paulson, Richard D. Cummings, William S. York, Niclas G. Karlsson, Frédérique Lisacek, Nicolle H. Packer, Matthew P. Campbell, Nobuyuki P. Aoki, Akihiro Fujita, Masaaki Matsubara, Daisuke Shinmachi, Shinichiro Tsuchiya, Issaku Yamada, Michael Pierce, René Ranzinger, Hisashi Narimatsu, and Kiyoko F. Aoki-Kinoshita. GlyTouCan: An accessible glycan structure repository. *Glycobiology*, 27(10):915–919, 2017. doi: 10.1093/glycob/cwx066.

James Urban, Chunsheng Jin, Kristina A. Thomsson, Niclas G. Karlsson, Callum M. Ives, Elisa Fadda, and Daniel Bojar. Predicting glycan structure from tandem mass spectrometry via deep learning. *Nature Methods*, 21(7):1206–1215, 2024. doi: 10.1038/s41592-024-02314-6.

Ajit Varki, Richard D. Cummings, Markus Aebi, Nicole H. Packer, Peter H. Seeberger, Jeffrey D. Esko, Pamela Stanley, Gerald W. Hart, Alan Darvill, Taroh Kinoshita, James J. Prestegard, Ronald L. Schnaar, Hudson H. Freeze, Jamey D. Marth, Carolyn R. Bertozzi, Marilynn E. Etzler, Martin Frank, Johannes F. G. Vliegenthart, Thomas Lütteke, Serge Perez, Evan Bolton, Pauline Rudd, James Paulson, Minoru Kanehisa, Philip Toukach, Kiyoko F. Aoki-Kinoshita, Anne Dell, Hisashi Narimatsu, William York, Naoyuki Taniguchi, and Stuart Kornfeld. Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology*, 25(12):1323–1324, 2015. doi: 10.1093/glycob/cwv091.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017.

Yu Watanabe, Kiyoko F. Aoki-Kinoshita, Yasushi Ishihama, and Shujiro Okuda. GlycoPOST realizes FAIR principles for glycomics mass spectrometry data. *Nucleic Acids Research*, 49(D1): D1523–D1528, 2021. doi: 10.1093/nar/gkaa1012.

Melih Yilmaz, William E. Fondrie, Wout Bittremieux, Sewoong Oh, and William S. Noble. De novo mass spectrometry peptide sequencing with a transformer model. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 25514–25522, 2022.

Melih Yilmaz, William E. Fondrie, Wout Bittremieux, Carlo F. Melendez, Rowan Nelson, Varun Ananth, Sewoong Oh, and William Stafford Noble. Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nature Communications*, 15(1):6427, 2024. doi: 10.1038/s41467-024-49731-x.

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

Wen-Feng Zeng, Wei-Qian Cao, Ming-Qi Liu, Si-Min He, and Peng-Yuan Yang. Precise, fast and comprehensive analysis of intact glycopeptides and modified glycans with pGlyco3. *Nature Methods*, 18(12):1515–1523, 2021. doi: 10.1038/s41592-021-01306-0.

A Mask-source CTMC Euler update

For reference, Algorithm 2 states the standard CTMC Euler update for mask-source DFM (the mask-source specialization of Algorithm 1 of Gat et al., 2024). At the final step $k = K - 1$, $t = 1 - 1/K$ so $h/(1-t) = 1$, committing every remaining masked position within K forward passes.

Algorithm 2 CTMC Euler update for mask-source DFM (Gat et al., 2024, Algorithm 1).

Require: denoiser $p^{\theta}_{1|t}$, spectrum embedding $\mathbf{h}_{\mathrm{spec}}$, step budget K

Initialize $X^i \leftarrow m$ for non-special $i \in \{1, \dots, L\}$ and fixed specials on $\mathcal{C} \leftarrow \{\text{special positions}\}$
$t \leftarrow 0$; $\quad h \leftarrow 1/K$
for $k = 0, 1, \dots, K-1$ **do**
  $\mathbf{p} \leftarrow p^{\theta}_{1|t}(\cdot \mid X, \mathbf{h}_{\mathrm{spec}})$ ▷ 1 parallel forward pass; $\mathbf{p}^i \in \Delta^{V-1}$
  for each masked position $i \notin \mathcal{C}$ with $X^i = m$ **do**
    $r \sim \mathrm{Uniform}(0, 1)$
    if $r < h/(1-t)$ **then**
      $X^i \sim \mathbf{p}^i$ ▷ commit a sampled token and freeze
    end if
  end for
  $t \leftarrow t + h$
end for
return X

B Spectrum dual-channel binning and patch encoding

Raw tandem mass spectra vary widely in peak density, and low-intensity fragment ions frequently carry diagnostic branching or linkage information that top-N peak filtering discards. Following the binned CandyCrunch v2 representation [Urban et al., 2024], we discretize each spectrum across $B_{\mathrm{mz}} = 2{,}048$ uniform bins spanning the glycan fragment mass window alongside a continuous sub-bin mass-offset vector recording the exact intra-bin m/z remainder of peaks falling in each bin. We normalize both vectors $\mathbf{x} \in \mathbb{R}^{2048}_{\geq 0}$ per spectrum via $\tilde{\mathbf{x}} = \log(1+\mathbf{x})/(\max_b \log(1+x_b) + 10^{-6})$, compressing dynamic range while preserving faint peaks and Orbitrap-grade mass offsets.

Rather than feeding 2,048 individual bin tokens into quadratic self-attention, the spectrum encoder groups every $C = 32$ contiguous bins into $N_{\mathrm{patch}} = 64$ non-overlapping patches and concatenates the intensity and sub-bin offset channels within each patch to form a 64×64 patch matrix (Figure 5). A linear layer $\mathbb{R}^{64} \to \mathbb{R}^{512}$ projects the 64 patches into model width $d = 512$, while a separate linear map $\mathbb{R}^1 \to \mathbb{R}^{512}$ projects the scalar precursor m/z (z-score standardized using training-set mean and standard deviation) as token 65. Adding learned positional embeddings in $\mathbb{R}^{65\times 512}$ and applying 4 pre-norm GELU Transformer encoder layers (8 heads, feed-forward width $2{,}048$) computes $\mathbf{h}_{\mathrm{spec}} \in \mathbb{R}^{65\times 512}$ in a single encoder pass cached across all K decoder steps.

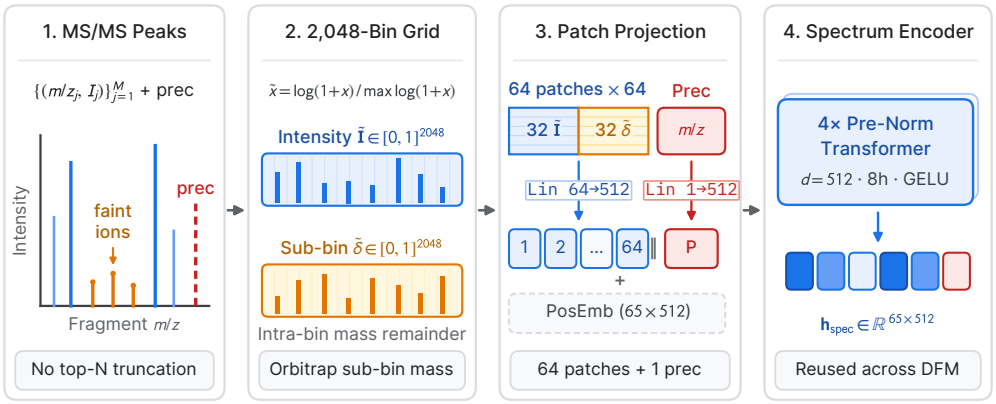


Figure 5: **Dual-channel spectrum binning and patch encoding.** Variable-length fragment peak lists $\{(m/z_j, I_j)\}_{j=1}^{M}$ map without top-N peak truncation onto a 2,048-bin mass-to-charge grid carrying two normalized channels per bin: log-scaled peak intensity $\tilde{\mathbf{I}} \in [0,1]^{2048}$ and continuous sub-bin mass remainder $\tilde{\boldsymbol{\delta}} \in [0,1]^{2048}$. Reshaping into 64 contiguous patches of 32 bins (64 features per patch), projecting to $d = 512$, and appending one standardized precursor-m/z token yields a fixed 65-token sequence encoded once per spectrum into $\mathbf{h}_{\text{spec}} \in \mathbb{R}^{65\times512}$.

C Crystallizing Flows: mechanistic intuitions

Distribution propagation preserves information. The per-position softmax over the 74-token vocabulary (70 atomic IUPAC tokens plus 4 special tokens) carries up to $\log_2 74 \approx 6.2$ bits of information, compared to zero bits once a token commits to a single value. CTMC Euler discards this distribution at every commit step and replaces a calibrated belief with a single Bernoulli draw; once committed, positions remain frozen and cannot recover lost probability mass. Crystallizing Flows instead preserve the full simplex state until the model's own argmax leaves `<mask>`, collapsing only positions on which the denoiser has formed a clear consensus.

Determinism eliminates sampling variance. CTMC Euler injects per-step Bernoulli noise at rate $h/(1-t)$. Sampling variance compounds across positions and steps; at small step budgets K, the trajectory drifts far from the posterior mean and accuracy regresses below a single forward pass, dropping 3.5 percentage points to 86.70% across three seeds at K=2 (86.76% on seed 0; Tables 2 and 10). Crystallizing Flows take the closed-form posterior-mean update at every step without random draws, improving accuracy monotonically with K.

Clean conditioning at decided positions sharpens remaining positions. Crystallized positions feed the decoder discrete one-hot tokens, while uncrystallized positions feed soft mixture embeddings. Trained on discrete one-hot inputs, the decoder treats crystallized positions as observed context and concentrates attention on resolving the remaining uncertain positions, tightening the conditional sub-problem on each commit.

Why early exit averages roughly half the step budget. Under the linear schedule $\kappa_t = t$, a trajectory starting at $\mathbf{s}_0^i = \delta_m$ whose denoiser places zero mass on `<mask>` retains exact mask mass $\mathbf{s}_{t_k}^i(m) = 1 - t_k$ at step $t_k = k/K$. Because $\mathbf{s}_t^i(m) \geq 0.5$ for all $t \leq 0.5$, no position can leave `<mask>` before the midpoint. On the first step past $t = 0.5$ ($t_{k+1} = 0.5 + 1/K$, forward pass $\lfloor K/2 \rfloor + 1$), mask mass drops below 0.5 and every confident position whose leading token exceeds $1 - t_{k+1}$ crystallizes immediately (91.6% of commits at K=8 occur on pass 4 and the remaining 8.4% on pass 5; Section D). Across all 76,774 test spectra, decoding therefore averages roughly $K/2 + 1$ effective passes, delivering measured wall-clock speedups from 1.28× at K=4 to 1.84× at K=64.

Why no unmasking schedule or temperature is needed. MaskGIT-style top-N schedulers require choosing how many positions to commit per step, and temperature samplers require tuning a softmax temperature. Crystallizing Flows require neither: commits trigger directly when a position's simplex argmax overtakes `<mask>`, calibrating unmasking timing automatically from the flow schedule.

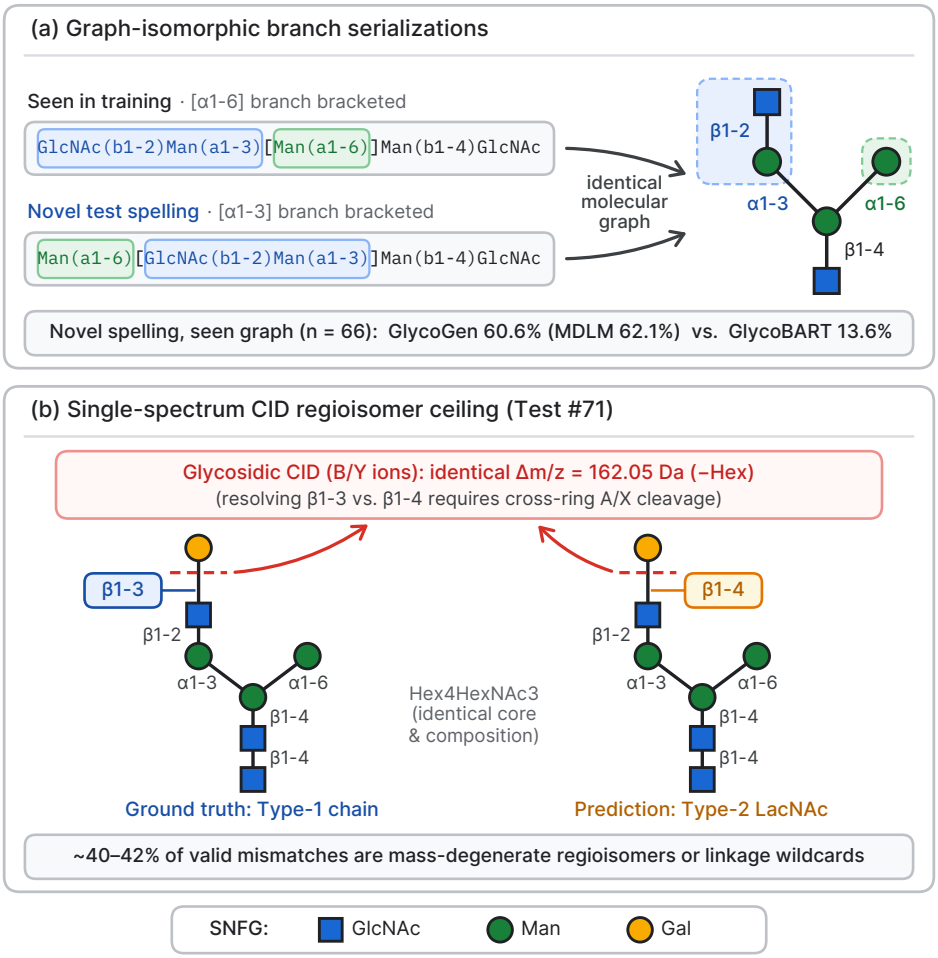


Figure 6: **Graph-isomorphic generalization and the single-spectrum CID regioisomer ceiling.** **(a)** Distinct IUPAC bracket traversal orders serialize the exact same branched SNFG molecular graph; GlycoGen generalizes across unseen branch spellings (60.6% DFM, 62.1% MDLM vs. 13.6% GlycoBART on n=66). **(b)** Terminal β1-3 vs. β1-4 galactose regioisomers (#71) yield identical glycosidic CID neutral losses ($\Delta m/z = 162.05$ Da), illustrating the physical single-spectrum CID ambiguity behind ∼40–42% of valid mismatches.

D Multi-seed replication, generalization, and qualitative error analysis

Full three-seed replication across step budgets. Table 2 reports the three-seed replication (initialization seeds 0, 1, and 2) across step budgets $K \in \{1, 2, 4, 8, 16, 32, 64\}$ for GlycoGen-DFM and MDLM under both standard CTMC Euler and Crystallizing Flows, matching the curves and $\pm 1\sigma$ bands in Figure 4.

Structure-level generalization breakdown. Table 3 splits the 76,774 test spectra into three mutually exclusive regimes by comparing each target against the training set via exact IUPAC string matching and `glycowork` graph isomorphism: 76,203 spectra whose exact IUPAC string appears in training, 66 spectra whose glycan graph appears in training only under a different IUPAC branch ordering, and 505 spectra whose glycan graph is absent from training under any string encoding.

Cross-instrument robustness and held-out analyzer retraining. Table 4 reports per-instrument accuracy on the 76,774 test spectra using the true analyzer metadata counts in X_{test}, alongside leave-one-instrument-out retraining runs in which all linear ion trap spectra (220,762 training rows) or all

K	Eff. NFE (CF)	DFM + CF	DFM + CTMC Euler	MDLM + CF	MDLM + CTMC Euler
1	1.00	0.9019 ± 0.0004	0.9019 ± 0.0004	0.8973 ± 0.0017	0.8973 ± 0.0017
2	2.00	0.9213 ± 0.0003	0.8670 ± 0.0008	0.9186 ± 0.0011	0.8630 ± 0.0021
4	2.99	0.9229 ± 0.0005	0.8852 ± 0.0010	0.9208 ± 0.0012	0.8814 ± 0.0025
8	5.05	0.9251 ± 0.0006	0.8957 ± 0.0008	0.9230 ± 0.0012	0.8922 ± 0.0015
16	9.16	0.9260 ± 0.0006	0.9028 ± 0.0012	0.9242 ± 0.0011	0.8993 ± 0.0018
32	17.40	0.9267 ± 0.0005	0.9067 ± 0.0016	0.9247 ± 0.0011	0.9038 ± 0.0014
64	33.89	0.9269 ± 0.0006	0.9087 ± 0.0011	0.9249 ± 0.0012	0.9058 ± 0.0013

Table 2: Three-seed top-1 structural accuracy (mean $\pm$ standard deviation over initialization seeds 0, 1, and 2) across the 76,774 spectra of the CandyCrunch v2 benchmark test set. Eff. NFE (CF) reports mean forward passes under Crystallizing Flows early exit; CTMC Euler evaluates all K passes.

Model	Seen str. (76,203)	Novel str., seen graph (66)	Novel graph (505)
GlycoGen-DFM (K=1)	0.9081 ± 0.0004	0.6061 [0.485, 0.715]	0.0079 [0.003, 0.020]
GlycoGen-DFM + CF (K=32)	$\mathbf{0.9331 \pm 0.0005}$	0.5758 [0.456, 0.688]	0.0178 [0.009, 0.033]
GlycoGen-MDLM + CF (K=32)	0.9311 ± 0.0011	**0.6212** [0.500, 0.729]	0.0020 [0.000, 0.011]
GlycoGen-AR (beam-1)	0.9102 ± 0.0005	0.5455 [0.426, 0.660]	0.0059 [0.002, 0.017]
GlycoBART (beam-1, 3 seeds)	0.8152 ± 0.0031	0.1364 [0.073, 0.239]	**0.0356** [0.023, 0.056]

Table 3: Top-1 structural accuracy disaggregated by target novelty on the 76,774 spectra of the CandyCrunch v2 benchmark test set. Bracketed intervals on smaller subsets denote 95% Wilson confidence intervals.

Orbitrap spectra (113,200 training rows) are excluded from training and validation and evaluated zero-shot.

Instrument family (`trap`)	Test n	In-dist. (K=1)	In-dist. (K=8 CF)	Held-out 0-shot (K=1)
Linear ion trap (`linear`)	39,446	0.8829	0.9061	0.414 [0.409, 0.419]
Orbitrap (`orbitrap`)	18,055	0.9299	0.9508	0.141 [0.136, 0.146]
Bruker amaZon (`amazon`)	13,046	0.9291	0.9525	—
Other (`other`)	6,227	0.8890	0.9074	—

Table 4: Per-instrument accuracy on X_{test} (n=76,774) and zero-shot leave-one-analyzer-out retraining on linear ion trap and Orbitrap (95% Wilson CIs). Matched MDLM zero-shot retraining scores 0.404 on held-out linear ion trap and 0.157 on held-out Orbitrap.

Pre-commit confidence calibration and qualitative examples. Instrumenting all 7,084,712 token-crystallization events across X_{test} at K=8 confirms how Crystallizing Flows differ from one-shot decoding: whereas single-pass (K=1) max-softmax confidences are bimodal (mean 0.739, median 0.910, with 33.8% above 0.99 and 34.1% below 0.50), multi-step simplex commits fire as soon as consensus emerges off `<mask>` (mean pre-commit simplex probability 0.562 ± 0.060, median 0.617, bounded above at 0.625 by the linear Euler step; 91.6% of commits occur at step 4 and 8.4% at step 5). Figure 6 and Table 5 illustrate representative test-set predictions, graph-isomorphic generalization across branch orderings, and the physical CID regioisomer ceiling.

E GlycoBART reproduction and per-spectrum evaluation

Setup. We reproduce the 207-million-parameter GlycoBART architecture from Abtheen et al. [2025b] (12 encoder layers, 12 decoder layers, hidden width 768, 16 attention heads, and feed-forward dimension 3072) and retrain it from scratch on a run-stratified partition of CandyCrunch [Urban et al., 2024]. The published GlycoBART split is a random spectrum-level partition, which allows spectra from the same LC-MS run and biological sample to appear in both training and test sets. Under such a split, a model can exploit memorized retention times, ionization patterns, and matrix effects. We instead group spectra by source filename before partitioning via

Regime (X_{test} idx)	Ground-truth IUPAC	GlycoGen + CF prediction
Exact branched hit (#15964)	Neu5Ac(a2-?)Gal(b1-4)GlcNAc(b1-2) Man(a1-3)[Man(a1-6)]Man(b1-4) GlcNAc(b1-4)[Fuc(a1-6)]GlcNAc	Exact match (monosialylated, core-fucosylated N-glycan).
Novel spelling, seen graph (#10631)	GlcNAc(b1-2)Man(a1-6)[Man(a1-3)] [GlcNAc(b1-4)]Man(b1-4) GlcNAc(b1-4)GlcNAc	GlcNAc(b1-2)Man(a1-?)[GlcNAc(b1-4)] [Man(a1-?)]Man(b1-4)GlcNAc(b1-4)GlcNAc (isomorphic bisecting graph).
Regioisomer / wildcard (#71)	Gal(b1-3)GlcNAc(b1-2)Man(a1-3) [Man(a1-6)]Man(b1-4) GlcNAc(b1-4)GlcNAc	Gal(b1-4)GlcNAc(b1-2)Man(a1-?) [Man(a1-?)]Man(b1-4)GlcNAc(b1-4)GlcNAc ($\text{Hex}_4\text{HexNAc}_3$ regioisomer).
Unseen graph fallback (#72)	GlcNAc(b1-2)Man(a1-3)[ManOMe(a1-6)] Man(b1-4)GlcNAc(b1-4)GlcNAc	Fuc(a1-2)Gal(b1-4)GlcNAc(b1-6) [Neu5Ac(a2-3)Gal(b1-3)]GalNAc (rare ManOMe $\rightarrow$ common motif fallback).

Table 5: Qualitative examples from X_{test} illustrating exact branched recovery, graph-isomorphic generalization across branch orderings, single-spectrum CID regioisomer ambiguity, and unseen-topology prior fallback.

GroupShuffleSplit(test_size=0.15, random_state=42), ensuring that no biological run contributes spectra to both training and evaluation. The resulting CandyCrunch v2 split contains 420,977 training and 76,774 test spectra over disjoint filename sets. We retain GlycoBART's antenna-form tokenization and train with AdamW following the published recipe: an initial learning rate of 1×10^{-5} scaled by $\sqrt{2}$ for our batch size of 160, no weight decay, gradient clipping at 1.0, and a full 30-epoch budget without early stopping.

Evaluation methodology: per-spectrum, no aggregation. We evaluate all models per-spectrum: each MS/MS spectrum is scored independently against its annotated glycan structure using the same glycowork graph-isomorphism evaluation as CandyCrunch and GlycoBART. We apply no consensus voting across the 5 to 50 scans a single glycan generates within one sample, no snap-back to a curated glycan database, and no aggregation across charge states. This protocol matches real-time, single-spectrum acquisition, where a structure must be inferred from each fragmentation event as it arrives rather than assembled offline across an entire LC-MS run. By contrast, the published 89.1% top-1 and 93.1% top-5 figures in Abtheen et al. [2025b] are evaluated at the sample level after per-sample deduplication, charge-state combination, and snap-back to the curated df_glycan database. Those aggregations raise accuracy by 5 to 15 percentage points when individual scans are noisy. We retain the stricter per-spectrum metric throughout so that our GlycoBART reproduction, GlycoGen, and autoregressive baselines are compared head-to-head on intrinsic per-spectrum capability rather than offline ensemble voting or external database lookup.

Headline results. On the 76,774 spectra of the leak-free CandyCrunch v2 test set, our reproduced GlycoBART reaches $80.92\% \pm 0.31\%$ top-1 structural accuracy under greedy decoding (beam 1 across three seeds) and 86.35% top-5 accuracy at beam 32. At beam 1, antenna-form structural accuracy reaches 91.24%; the roughly 10 percentage point drop to full IUPAC accuracy stems from the lossy antenna-to-IUPAC reconstruction inherent to GlycoBART's tokenization. Because the structural-equivalence ceiling under linkage normalization is 99.65%, any antenna-form decoder under this pipeline faces a practical IUPAC top-1 ceiling near 91%. Batched inference at beam 1 takes 2.80 milliseconds per spectrum on a single H200 GPU in BF16 at batch size 128, corresponding to roughly 23 forward passes per spectrum.

Pareto frontier and the beam-search curse. Sweeping beam widths from 1 to 32 (Table 6) reveals a non-monotonic relationship between beam width and top-1 accuracy. Greedy decoding achieves the highest top-1 accuracy on the primary seed at 81.18%; accuracy drops to 79.15% at beam 2, 78.77% at beam 4, and 78.64% at beam 8 before partially recovering to 78.78% at beam 16 and 79.11% at beam 32. By contrast, top-5 accuracy rises monotonically from 81.18% to 86.35%. This divergence exemplifies the beam-search curse [Stahlberg and Byrne, 2019, Cohen and Beck, 2019]: wider beams find higher-likelihood completions, but for sequences with ambiguous glycan linkage symbols, higher

likelihood does not guarantee greater structural accuracy. Chasing top-5 accuracy also carries a steep latency penalty: beam 32 is 32.6× slower than beam 1 (91.21 versus 2.80 milliseconds per spectrum) for a 0.30 percentage point top-5 gain over beam 16 and a 2.07 percentage point loss in top-1 accuracy relative to greedy decoding. Even against GlycoBART's 5-candidate oracle ceiling (86.35% top-5 at beam 32), 8-step GlycoGen leads by 6.2 percentage points in top-1 accuracy (92.51% at 0.201 ms/spec) at 454-fold lower batched latency.

beam	top-1	top-5	top-1 (mass-rerank)	antenna top-1	antenna top-5	ms/spec	rel. cost
1 (greedy)	**0.8118**	0.8118	0.8118	**0.9124**	0.9124	2.80	1×
2	0.7915	0.8275	0.7933	0.8852	0.9183	10.07	3.6×
4	0.7877	0.8482	0.7909	0.8804	0.9258	15.93	5.7×
8	0.7864	0.8578	0.7908	0.8782	0.9249	27.12	9.7×
16	0.7878	0.8605	0.7926	0.8792	0.9256	46.48	16.6×
32	0.7911	**0.8635**	0.7961	0.8822	**0.9281**	91.21	32.6×

Table 6: GlycoBART beam sweep on the 76,774 spectra of the CandyCrunch v2 benchmark test set (primary seed), evaluated per-spectrum. Top-1 accuracy is non-monotonic in beam width and peaks at greedy decoding, whereas top-5 improves monotonically. Mass reranking checks precursor-mass plausibility across all beam candidates and selects the highest-likelihood mass-consistent prediction, adding at most 0.5 percentage points at any beam width.

Post-processing pipelines. We also tested two post-processing pipelines on the GlycoBART outputs: first, a reranking pass combining precursor-mass and fragment-match scores across all beam candidates; second, the published `filter_glycans_glycobart` pipeline of Abtheen et al. [2025b] applied verbatim. The first did not improve top-1 accuracy because composite heuristic scores ranked correct structures lower than the model's own likelihood. The second is incompatible with per-spectrum evaluation because it deduplicates predictions across scans from the same biological sample, collapsing per-spectrum evaluation into sample-level aggregation. We therefore report raw per-spectrum model accuracy throughout.

Streaming-compatible baseline and unbatched end-to-end latency. The published GlycoBART pipeline is inherently offline: beam-32 decoding takes 91 milliseconds per spectrum even when batched on an H200 GPU, compared with a 5 to 10 millisecond scan interval on modern 100 to 200 Hz instruments, and sample-level voting requires waiting for the full LC-MS run to finish. Table 7 reports unbatched serial end-to-end streaming latency from a raw peak list to a decoded IUPAC string across 1,000 test spectra. CPU peak-binning and normalization add only 0.13 milliseconds per spectrum, allowing single-pass GlycoGen at 8.6 milliseconds per spectrum to keep pace with 100 Hz acquisition while beam-1 GlycoBART requires 183 milliseconds per spectrum.

Model / operating point	CPU preproc. (ms)	GPU forward (ms)	Total E2E (ms)	Max rate
GlycoGen K=1 (one-pass)	0.13 ± 0.03	8.5 ± 0.3	$\mathbf{8.6 \pm 0.3}$	116 Hz
GlycoGen K=8 (CF)	0.13 ± 0.03	33.7 ± 2.5	$\mathbf{33.8 \pm 2.5}$	30 Hz
GlycoGen K=32 (CF)	0.13 ± 0.03	109.4 ± 10.2	109.5 ± 10.2	9.1 Hz
GlycoBART beam-1 (greedy)	0.13 ± 0.03	183 ± 99	183 ± 99	5.5 Hz
GlycoBART beam-8	0.13 ± 0.03	240 ± 110	240 ± 110	4.2 Hz
GlycoBART beam-32	0.13 ± 0.03	256 ± 107	256 ± 107	3.9 Hz

Table 7: Unbatched serial end-to-end streaming latency on a single H200 GPU in BF16, reported as mean and standard deviation over 1,000 randomly sampled test spectra. CPU preprocessing includes peak filtering, intensity normalization, and mass-to-charge histogram binning.

F GlycoGen-AR: matched-architecture autoregressive control

We train a 46.4-million-parameter autoregressive sequence-to-sequence model that matches GlycoGen-DFM in every dimension except the three that define autoregressive modeling: causal

self-attention in the decoder, no time embedding, and teacher-forced next-token cross-entropy. This control isolates the contribution of Discrete Flow Matching from architecture and parameter count.

Training. We use the same CandyCrunch v2 split, optimization recipe, and H200 hardware as GlycoGen-DFM: AdamW at batch size 1024 for 30 epochs, a peak learning rate of 8.49×10^{-4} with cosine decay to 3% of peak after 200 warmup steps, and BF16 mixed precision. Training completes 11,100 steps in 49.5 minutes on a single H200 GPU (4.34 training steps per second excluding validation). The best checkpoint (step 10,500, epoch 28.4) reaches 93.30% top-1 structural accuracy at beam 4 on a fixed 3,000-spectrum validation subset and 92.74% on the full 42,097-spectrum validation set.

Test-set Pareto sweep. Table 8 evaluates the best checkpoint on the full 76,774-spectrum CandyCrunch v2 benchmark test set across beam widths 1, 2, 4, 8, 16, and 32 using the same Hugging Face beam-search implementation as GlycoBART. Without key-value caching, decoding latency scales linearly with beam width. At beam widths 16 and 32, we cap the batch size at 1,024 and 512 and disable the cuDNN scaled-dot-product attention backend to avoid graph execution limits at very large effective attention batches.

beam	Top-1 struct. acc.	NFE/spec	ms/spec	batch
1	0.9035	32.68	2.21	4096
2	0.9038	65.34	4.37	4096
4	0.9039	130.68	8.72	4096
8	0.9038	261.35	17.41	4096
16	0.9039	522.69	36.50	1024
32	0.9039	1045.39	72.66	512

Table 8: GlycoGen-AR Pareto sweep on the 76,774 spectra of the CandyCrunch v2 benchmark test set (primary seed). Top-1 accuracy varies by only 0.04 percentage points across a 32-fold increase in beam width, forward passes, and wall-clock time. The autoregressive Pareto frontier degenerates to a vertical line at roughly 33 forward passes.

Why beam search confers no gain. Token-level accuracy at the converged checkpoint reaches 99.61%, corresponding to a 0.4% per-token error rate. When the model makes an error, it is overconfident: the ground-truth token rarely appears among the top beam candidates at that position, leaving beam search no path to recover it. Moreover, because `glycowork` graph isomorphism treats equivalent IUPAC branch permutations and wildcard linkages as identical, most distinct strings across a beam collapse to the same molecular graph. A 30-spectrum diagnostic confirms that beam width 4 produces 4.0 distinct strings on average yet bounds the oracle top-4 gain at just 3.3 percentage points, consistent with the 0.04 percentage point gain observed across the full test set.

Implications. Wider beams multiply latency and forward passes without improving structural accuracy. Autoregressive decoding pays a floor of roughly 33 sequential forward passes per spectrum that beam search cannot reduce. Improving the speed-accuracy trade-off on this architecture therefore requires replacing sequential decoding altogether: single-pass GlycoGen cuts forward passes 32-fold at comparable accuracy, and Crystallizing Flows lift accuracy by another 2.3 percentage points in 5 to 17 passes (Section 5).

G MDLM cross-objective ablation

We train a matched-architecture 50-million-parameter Masked Diffusion Language Model denoiser [Sahoo et al., 2024] on the same dataset, vocabulary, optimizer, and learning-rate schedule as GlycoGen-DFM. Only the time distribution and loss weighting change: time t is sampled uniformly from the 128-step discrete grid $\{1/128, 2/128, \ldots, 1\}$ instead of continuous $\mathcal{U}(0, 1)$, and the cross-entropy loss uses the NELBO weighting of Sahoo et al. [2024]. On the primary seed, the MDLM checkpoint reaches 92.17% validation and 89.89% test top-1 accuracy in a single pass, trailing the DFM checkpoint by 0.34 percentage points under the same training compute budget.

MDLM + Crystallizing Flows Pareto sweep. Table 9 compares standard CTMC Euler against Crystallizing Flows on the MDLM checkpoint across $K \in \{1, \ldots, 64\}$. Once we account for the 0.34 percentage point single-pass offset between the trained checkpoints, MDLM with Crystallizing Flows tracks DFM with Crystallizing Flows within 0.16 percentage points at every step budget on seed 0, showing that Crystallizing Flows lift any mask-source denoiser at inference time without continuous-time DFM retraining.

K	MDLM CTMC Euler	MDLM + CF	Δ vs CTMC Euler	Eff. NFE (CF)
1	0.8989	0.8989	—	1.00
2	0.8642	0.9184	+5.42 pp	2.00
3	—	0.9170	—	2.02
4	0.8811	0.9205	+3.94 pp	2.99
8	0.8928	0.9228	+3.00 pp	5.05
16	0.9005	0.9243	+2.38 pp	9.16
32	0.9051	0.9248	+1.97 pp	17.41
64	0.9054	0.9250	+1.96 pp	33.90

Table 9: MDLM with Crystallizing Flows on the 76,774 spectra of the CandyCrunch v2 benchmark test set (primary seed). Crystallizing Flows lift the MDLM denoiser by 1.96 to 5.42 percentage points over CTMC Euler across all step budgets from 2 to 64, matching the gain profile observed on Discrete Flow Matching.

Two-step sampling already reaches the low-step shoulder. At 3 nominal steps, both MDLM with Crystallizing Flows (91.70% at 2.02 effective passes) and DFM with Crystallizing Flows (91.90% at 2.02 effective passes) sit slightly below their 2-step counterparts (91.84% and 92.14% at 2.00 effective passes) at the same wall-clock cost. Nearly every spectrum crystallizes within the first two Euler steps at $K = 3$, so the third step adds little refinement. Because 2 steps strictly dominate 3 steps on both objectives, we omit $K = 3$ from the Pareto curves in Figure 4.

Off-grid continuous-time inference. Although the MDLM denoiser is trained on a 128-step discrete time grid, Crystallizing Flows take continuous Euler steps that land off the training grid whenever K does not divide 128. Rounding t to the nearest training-grid point before each forward pass produces the exact same test accuracy to four decimal places as continuous off-grid evaluation across $K \in \{1, 2, 4, 8, 16, 32\}$ and differs by only 0.01 percentage points at $K = 64$. The learned conditional distribution is smooth enough in t that continuous-time sampling on a discrete-time checkpoint incurs no calibration penalty.

H Sampler ablations

We compare Crystallizing Flows against seven alternative inference algorithms on the same trained 50-million-parameter GlycoGen-DFM checkpoint across step budgets from 1 to 64. Table 10 reports the main accuracy comparison across effective function evaluations, and the paragraphs below summarize each alternative.

Velocity-explicit Euler. This variant removes the absorbing-state constraint of CTMC Euler and draws a categorical sample from the full per-position mixture $\big(1 - \frac{h}{1-t}\big)\delta_{z^i} + \frac{h}{1-t}\, p^\theta_{1|t}$ at every position and step. At 4 steps (5 forward passes), it reaches 90.27% accuracy, gaining 1.66 percentage points over standard CTMC Euler but trailing Crystallizing Flows by 2.24 percentage points.

Velocity with final-step argmax. Replacing the final categorical draw at $t \approx 1$ in velocity-explicit Euler with a deterministic argmax removes boundary resampling noise and lifts 4-step accuracy to 91.07%.

All-argmax CTMC. This variant keeps the stochastic Bernoulli commit schedule of CTMC Euler, but commits each chosen position to the argmax of $p^\theta_{1|t}$ rather than a categorical sample. Removing token sampling noise lifts 4-step accuracy to 91.65%, though random commit timing still leaves it 0.86 percentage points behind Crystallizing Flows.

NFE	GlycoGen-AR	CTMC Euler	Soft prop.	Greedy conf.	Crystallizing Flows
1	—	0.9023	0.9023	0.9023	0.9023
3	—	0.8676	0.9214	0.9115	**0.9230**
5	—	0.8861	0.9250	0.9187	**0.9251**
9	—	0.8964	0.9249	0.9226	**0.9260**
17	—	0.9042	0.9252	0.9257	**0.9266**
33	0.9035	0.9080	0.9247	**0.9277**	**0.9268**

Table 10: Top-1 per-spectrum structural accuracy across samplers on the 76,774 spectra of the CandyCrunch v2 benchmark test set (primary seed). Crystallizing Flows are reported at their effective forward-pass count under per-spectrum early exit ($K \in \{1, 4, 8, 16, 32, 64\}$); non-early-exit baselines at $K \in \{2, 4, 8, 16, 32\}$ include a terminal $t{=}1$ argmax cleanup pass ($K{+}1$ NFE). The GlycoGen-AR column shows our matched-architecture autoregressive control, bounded below by the mean sequence length of roughly 33 tokens and flat at 90.39% across beam widths up to 32 (Appendix F). Bold indicates the Pareto-optimal sampler at each budget within the ± 0.18 percentage point 95% Wilson confidence interval at 33 passes.

Soft propagation without crystallization (and temperature sweep). Propagating the simplex distribution via Equation (4) without snapping committed tokens to one-hot vectors reaches 92.50% at 4 steps (5 passes) and saturates there; varying the per-step softmax temperature $\tau \in \{0, 0.5, 1.0\}$ changes accuracy by less than 0.5 percentage points across all step counts. Adding discrete one-hot commits allows Crystallizing Flows to strictly dominate pure soft propagation from $K{=}8$ upward while also exiting early.

Greedy confidence (MaskGIT top-N). Following Chang et al. [2022], this deterministic sampler commits the $N_k = \lceil n_{\text{remaining_masks}} / n_{\text{remaining_steps}} \rceil$ most confident masked positions to their argmax at each step. It is the strongest alternative baseline, matching Crystallizing Flows within 0.09 percentage points at 33 or more forward passes (92.77% at 32 steps and 92.78% at 64 steps, within the test-set noise floor), but trails across the low-step regime up to 17 passes.

Greedy selection with categorical commits. Combining MaskGIT top-N position selection with categorical sampling at the chosen positions compounds sampling noise and drops 4-step accuracy to 89.18%, performing worse than greedy argmax at every step budget.

Heun's second-order predictor-corrector. Averaging velocity predictions across predictor and corrector steps following Karras et al. [2022] plateaus near 91.5% across all step budgets from 2 to 64 while doubling neural function evaluations to $2K$. Averaging two steps smooths the predictive distribution just as the denoiser sharpens around a token commit.

Per-spectrum budget routing. As an orthogonal latency optimization, we can route each spectrum to a step budget K based on its minimum token confidence on the initial pass at $t = 0$. Routing with a confidence threshold of 0.7 averages 2.46 forward passes and reaches 92.61% accuracy, offering a complementary knob that can be combined with Crystallizing Flows.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 claims match the experimental protocol in §4 and empirical results in §5 and Appendices D–H. Headline accuracy and batched throughput in Table 1 and Table 2 are measured on a single H200 GPU across three independent training seeds on the CandyCrunch v2 benchmark test split (n=76,774), complemented by 1,000-spectrum serial streaming latencies (Table 7) and sampler ablations (Table 10); scope boundaries (seen vs. unseen glycan topologies, cross-analyzer transfer, and single-spectrum CID regioisomer limits) are stated explicitly in §1, §2.1, §5, and §6.1.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in dedicated §6.1, covering near-zero generalization to completely novel glycan graph topologies (n=505, Table 3), zero-shot degradation on held-out mass-analyzer families requiring target-instrument fine-tuning (Table 4), single-benchmark evaluation with exclusion of the 1.1% curly-brace (`{...}`) subset for GlycoBART comparability, and crystallization probability calibration under linear schedules (Appendix D), alongside the physical single-spectrum CID isomer ceiling in §5.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical and algorithmic and states no novel theorems requiring formal proofs; the mask-source DFM formulation in §3.1 follows Gat et al. [2024], and §3.3 and Appendix C provide the Crystallizing Flows algorithm and mechanistic analysis.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §3.1–§3.3 and Appendix B specify the DFM objective, 2,048-bin dual-channel patch encoder, 50M Transformer dimensions, 74-token atomic IUPAC tokenizer, and Algorithm 1; §4, §5, and Appendices D–H specify the leak-free run-grouped split (420,977 train with 10% validation / 76,774 test, `random_state=42`), optimizer, learning-rate schedule, batch size, BF16 precision, and step/beam sweep grids. Complete code, split builders, and converged checkpoints are publicly released at `https://github.com/rlacombe/GlycoGen` and `https://huggingface.co/datasets/rlacombe/candycrunch-v2-benchmark`.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Open-source training, evaluation, baseline, and split-verification code is publicly available under the MIT License at `https://github.com/rlacombe/GlycoGen`, alongside the aligned CandyCrunch v2 benchmark splits (`https://huggingface.co/datasets/rlacombe/candycrunch-v2-benchmark`) and converged model checkpoints on Hugging Face. Raw CandyCrunch v2 data is publicly archived at Zenodo record 15741423 [Abtheen et al., 2025a].

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Architecture and tokenization details are specified in §3.2 and Appendix B; the leak-free train/validation/test split, `glycowork` graph-isomorphism metric, AdamW optimization recipe, two-stage Distillate hyperparameter selection, and matched baselines (`GlycoGen-MDLM`, `GlycoGen-AR`, and `GlycoBART`) are specified in §4 and Appendices D–H, with inference step and beam grids in §5.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Headline accuracies are reported as mean $\pm 1\sigma$ standard deviation across three independent random initialization/training seeds (seeds 0, 1, and 2) in Table 1, Figure 4, §5, and Table 2 (Appendix D). Finite-sample 95% Wilson binomial confidence intervals are reported for full-test sampler comparisons (± 0.18 pp at n=76,774, Table 10 in Appendix H) and for novel-topology and held-out-analyzer subsets in Table 3 and Table 4 (Appendix D), while Table 7 (Appendix E) reports serial latency mean $\pm 1\sigma$ standard deviation over $1{,}000$ randomly sampled test spectra.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: §4, §5, and Appendices E–G specify the compute environment (single 141 GB NVIDIA H200 GPU in BF16 on Hugging Face Jobs) and per-run wall-clock times ($\sim$45–50 min per 50M GlycoGen/AR seed and $\sim$9 h per 207M GlycoBART seed). Total reported experiment compute is $\sim$40–45 H200 GPU-hours across all multi-seed, held-out-analyzer, and evaluation runs, plus $\sim$20 GPU-hours on single A100 and L40S GPUs for preliminary Distillate sweeps (§4).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The dataset is publicly licensed mass-spectrometry data from a peer-reviewed source; no human-subjects work was conducted.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive applications in biologics quality control, therapeutic engineering, and repository-scale glycomics are discussed in §1, §6.2, and §7, and physical limits of single-spectrum CID isomer resolution are discussed in §5 and §6.1. We see no plausible negative societal or misuse pathway specific to analytical MS/MS glycan structure prediction.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: The model predicts glycan structures from analytical MS/MS spectra and does not generate text, code, or any content with misuse risk.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: CandyCrunch v2 [Urban et al., 2024, Abtheen et al., 2025a] is hosted on Zenodo (record 15741423, CC BY 4.0); `glycowork` [Thomès et al., 2021] and GlycoBART (`glycotrans`) [Abtheen et al., 2025b] are released under the MIT License; and Hugging Face `transformers` is licensed under Apache-2.0. All creators and licenses are credited and respected.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: As linked in §7, GlycoGen training and evaluation code, baseline reproductions, split-alignment utilities, and converged checkpoints (DFM, MDLM, and AR) are publicly released under the MIT License at `https://github.com/rlacombe/GlycoGen`, and the aligned CandyCrunch v2 benchmark splits are released on Hugging Face (`https://huggingface.co/datasets/rlacombe/candycrunch-v2-benchmark`) under CC BY 4.0 matching upstream CandyCrunch v2, with README documentation and reproduction commands.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human-subjects research was conducted.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human-subjects research was conducted; IRB approval is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: As noted in §4 and the Acknowledgments, we used the Distillate automated-research harness [Lacombe, 2026] for preliminary architecture and learning-rate sweeps, and used LLM coding and research agents to run data processing, model training, hyperparameter and sampler sweeps, evaluation scripts, figure generation, and manuscript editing. All mathematical formulations, algorithm designs, experimental protocols, and scientific claims were directed and verified by the human author.