LIEFLOWER: CONTROLLING TARGET PROTEIN DYNAMICS VIA LIE-GUIDED DISCRETE FLOWS

Manvitha Ponnapati[1], **Tong Chen**[2], **Yinuo Zhang**[3], **Pranam Chatterjee**[2,4,†]

[1]MIT Media Lab
[2]Department of Computer and Information Science, University of Pennsylvania
[3]Duke-NUS Medical School
[4]Department of Bioengineering, University of Pennsylvania

[†]Corresponding author: pranam@seas.upenn.edu

ABSTRACT

Many diseases, including cancer and neurodegeneration, arise from mutation-driven changes in protein dynamics that shift conformational ensembles toward dysfunctional regimes rather than from a single aberrant structure. Existing generative protein design methods largely condition on static structures or sequence-derived features, leaving no mechanism to design sequences that actively modulate non-equilibrium motion. We introduce **LieFlower**, a dynamics-conditioned generative framework that integrates discrete flow matching with Lie group representations of molecular dynamics to control protein behavior at the sequence level. We encode protein trajectories using truncated log-signatures, elements of a free Lie algebra that capture causal ordering and non-equilibrium flux, and we quantify conformational shifts by projecting log-signature differences along a global reference direction defined through a holonomy-aware Baker-Campbell-Hausdorff composition. This construction yields scalar targets that measure directional dynamical change, which we predict directly from sequence using a surrogate model and use to guide discrete generative flows without requiring simulation at inference time. We evaluate LieFlower by generating both targeted mutations and peptide binders that induce directional shifts in protein dynamics, including peptide-mediated modulation of mutant *p53*, demonstrating a general framework for dynamics-aware sequence design beyond static structure-based models.

1 INTRODUCTION

Generative modeling has become a central tool for protein and peptide design, enabling data-driven discovery of binders and modulators across diverse targets (Watson et al., 2023; Pacesa et al., 2025; Stark et al., 2025b; Chen et al., 2025a; Bhat et al., 2025; Hong et al., 2025). Most existing approaches condition on static structural snapshots or learned sequence embeddings, which limits their applicability to systems whose function depends on conformational dynamics rather than a single stable structure. This includes intrinsically disordered proteins (Uversky, 2025), allosteric receptors (Qing et al., 2020), fusion oncoproteins (Vincoff et al., 2025), and aggregation-prone proteins (Watson et al., 2025), where binding can induce substantial, functionally relevant shifts in conformational state that are not captured by snapshot-based models.

Discrete generative models such as discrete diffusion and discrete flow matching provide a complementary framework for sequence-level design by formulating generation as transport over discrete sequence space (Shi et al., 2024; Sahoo et al., 2024; Austin et al., 2021; Gat et al., 2024; Stark et al., 2024; Tang et al., 2025b; Davis et al., 2024). These models have enabled objective-guided generation under constraints such as affinity, specificity, and developability (Tang et al., 2025a; Chen et al., 2025c; Tang et al., 2025c; Chen et al., 2025b). However, existing conditioning signals are typically static or sequence-derived, and do not explicitly encode protein conformational dynamics, limiting their ability to design sequences that induce targeted dynamical state changes.

A dynamics-aware conditioning space should represent the full conformational evolution of a protein rather than isolated intermediates. Rough path theory provides such a representation by encoding

trajectories through their signatures, ordered collections of iterated integrals that are invariant to time reparameterization and capture causal interactions (Lyons, 2014; Lyons et al., 2007). The log-signature lies in a free Lie algebra and yields a finite-dimensional representation that includes antisymmetric non-equilibrium flux terms such as Lévy areas (Diamantakis et al., 2023). Recent work has shown that log-signatures can encode protein dynamics at scale and improve downstream prediction tasks (Qin et al., 2025), and related diffusion models have begun to explore generative modeling in log-signature and Lie group spaces (Barancikova et al., 2025; Bertolini et al., 2025).

In this work, we introduce **LieFlower**, a generative framework that integrates discrete flow matching with Lie algebraic representations of molecular dynamics to design sequence-level interventions that modulate protein conformational behavior. We formulate sequence design as a conditional transport problem guided by a target dynamical regime, enabling both targeted mutations and peptide binders to be generated within a unified framework. To avoid explicit protein-peptide or protein-variant dynamics simulation during generation, we train a surrogate model that predicts a scalar score quantifying the directional alignment between sequence-induced dynamics and a specified target log-signature, computed via a holonomy-aware Baker-Campbell-Hausdorff composition. We use this score to guide discrete generative flows during sampling, biasing generation toward sequences that induce protein dynamics aligned with the desired conformational regime.

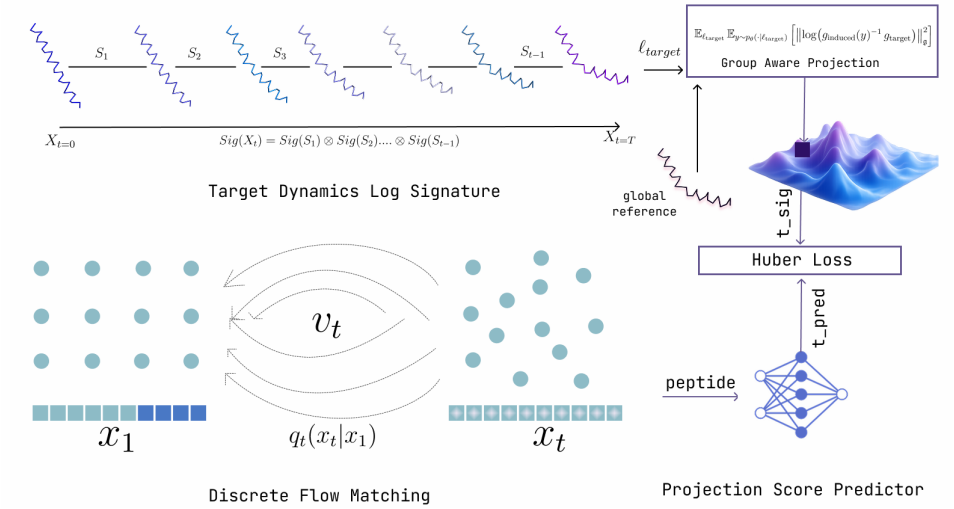

Figure 1: **LieFlower framework for dynamics-conditioned sequence generation.** Molecular target dynamics are encoded as a truncated log-signature $\ell_{\text{target}} = \log S(X) \in \mathfrak{g}_{\leq p}$ in the free Lie algebra, capturing causal ordering and non-equilibrium flux through antisymmetric components such as Lévy areas. Conditioned on this representation, a discrete flow matching model transports a simple sequence prior through reverse-time dynamics to generate peptide sequences whose predicted induced dynamics match the target state. A surrogate dynamics predictor maps generated sequences to induced log-signatures, and a Lie algebraic holonomy loss enforces consistency at the group level by penalizing discrepancies between the induced and target dynamics under the Lie exponential and logarithm maps.

2 Dynamics-Conditioned Sequence Design

We formalize LieFlower as a generative control problem over discrete biological sequences conditioned on continuous protein dynamical states. A molecular dynamics trajectory is represented as a continuous path $X : [0, T] \to \mathbb{R}^{3N}$, where N is the number of atoms, observed at discrete times $X_{\text{discrete}} = (X(t_1), \ldots, X(t_M)) \in (\mathbb{R}^{3N})^M$. We assume X has finite p-variation, which ensures well-defined signature integrals (Hambly & Lyons, 2010).

Path signatures and log-signatures. The signature of a path X over $[0, T]$ is the infinite collection of iterated integrals (Lyons, 2014; Lyons et al., 2007), forming an element of the tensor algebra

$T(\mathbb{R}^d) = \bigoplus_{k=0}^{\infty} (\mathbb{R}^d)^{\otimes k}$:

$$S(X) := \left(1, X^{(1)}, X^{(2)}, \ldots\right), \tag{1}$$

with

$$X^{(k)} = \int_{0<t_1<\cdots<t_k<T} dX_{t_1} \otimes \cdots \otimes dX_{t_k}.$$

The signature is invariant to time reparameterization and satisfies Chen's identity under path concatenation (Chen, 1954), making the space of signatures a Lie group under tensor product.

Taking the logarithm yields the *log-signature*

$$\ell(X) := \log S(X) \in \mathfrak{g}, \tag{2}$$

where $\mathfrak{g} = \mathrm{Lie}(\mathbb{R}^d)$ is the free Lie algebra. In practice, we work with the degree-p truncation

$$\ell_p(X) := \Pi_{\leq p} \ell(X) \in \mathfrak{g}_{\leq p}. \tag{3}$$

Second-order log-signature terms correspond to antisymmetric components such as Lévy areas,

$$\mathrm{Levy}_{i,j}(X) = \frac{1}{2}\left(\int_0^T X_t^i \, dX_t^j - \int_0^T X_t^j \, dX_t^i\right), \tag{4}$$

which quantify persistent cyclic flux and provide a signature of non-equilibrium dynamics.

Sequence-induced dynamics and surrogate modeling. A sequence $y \in \mathcal{A}^L$ (e.g., a mutation or peptide) induces a modified protein trajectory $X^{(y)}$ with log-signature

$$\ell_{\mathrm{induced}}(y) := \ell_p\left(X^{(y)}\right). \tag{5}$$

Direct simulation of $X^{(y)}$ is expensive, so we approximate the unknown dynamics map

$$F : \mathcal{A}^* \to \mathfrak{g}_{\leq p}, \qquad F(y) = \ell_p\left(X^{(y)}\right), \tag{6}$$

using a differentiable surrogate $\Phi_\psi(y) \approx F(y)$ trained on a dataset

$$\mathcal{D}_{\mathrm{dyn}} = \{(y^{(i)}, z^{(i)})\}_{i=1}^{N_{\mathrm{dyn}}}, \qquad z^{(i)} \approx \ell_p(X^{(y^{(i)})}), \tag{7}$$

by minimizing

$$\mathcal{L}_{\mathrm{dyn}}(\psi) = \mathbb{E}_{(y,z)\sim\mathcal{D}_{\mathrm{dyn}}}\left[\|\Phi_\psi(y) - z\|_{\mathfrak{g}}^2\right]. \tag{8}$$

Conditional sequence generation via discrete flow matching. LieFlower models the conditional law $p_\theta(y \mid \ell_{\mathrm{target}})$ using discrete flow matching. Let x_1 denote the clean sequence and x_t a corrupted version drawn from a forward noising process $q_t(x_t \mid x_1)$. A time-dependent vector field

$$v_\theta(x_t, t, \ell_{\mathrm{target}}) = \mathrm{NN}_\theta(x_t, t, \ell_{\mathrm{target}}) \tag{9}$$

is trained to match the target flow $u_t(x_t \mid x_1)$ by minimizing

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,x_1,x_t,\ell_{\mathrm{target}}}\left[\|v_\theta(x_t, t, \ell_{\mathrm{target}}) - u_t(x_t \mid x_1)\|^2\right]. \tag{10}$$

Sampling proceeds by integrating the learned reverse-time flow from a simple prior at $t = 1$ to $t = 0$, yielding a sequence y conditioned on ℓ_{target}.

Lie algebraic holonomy consistency. To evaluate whether a generated sequence produces dynamics consistent with desired target dynamics, we compare the induced and target log-signatures upto a truncated level. However, a direct comparison would require the surrogate model to predict the full truncated log-signature $\Phi_\psi(y) \in \mathfrak{g}_{\leq p}$ vector, which is typically very high-dimensional. Instead, we compare induced and target dynamics using low-dimensional quantities derived from their relative position in the Lie group.

Let

$$g_{\mathrm{induced}}(y) := \exp(\Phi_\psi(y)), \qquad g_{\mathrm{target}} := \exp(\ell_{\mathrm{target}}), \tag{11}$$

and we define the holonomy score as

$$\Delta(y; \ell_{\mathrm{target}}) = \log\left(g_{\mathrm{induced}}(y)^{-1} g_{\mathrm{target}}\right) \in \mathfrak{g}_{\leq p}. \tag{12}$$

Rather than matching Δ exactly during training, we summarise it using two scalar targets: its graded norm $\|\Delta\|_w$ on $\mathfrak{g}_{\leq p}$, which measures the overall magnitude of the dynamical shift, and its signed projection $t^\star$ along a fixed reference direction R.

mdCATH. Each file in mdCATH corresponds to a single protein domain drawn from the mdCATH database. For every domain, 5 replica trajectories are available at five temperatures from 320K to 450K.We extract the Cα backbone path from each replica, compute its depth-p log-signature, and take the Fréchet mean across replicates at each temperature. To define a shared reference frame for measuring temperature-induced conformational change, we compute the per-domain shift between low- and high-temperature log signatures and aggregate these shifts across all domains to obtain a global reference R. Each domain's response is then measured as the shift from this global low-temperature baseline, and the signed projection $t^\star$ of this shift along R serves as the regression target.

p53-peptide binder Each instance corresponds to a candidate peptide bound to a mutant p53 complex. Molecular-dynamics trajectories are first ran for the mutant (e.g. R175H) and wild-type p53 in isolation, and the reference direction R is defined as the BCH difference between their Fréchet mean log-signatures, capturing the dominant mode by which the mutation perturbs p53 conformational dynamics relative to wild type. The holonomy score Δ measures the shift of the bound complex away from the unbound mutant.The signed projection $t^\star$ of Δ along R quantifies the extent to which the peptide restores wild-type-like dynamics, and serves as the regression target.

We train the surrogate models to predict the scalar $t^\star$ directly from amino-acid sequence, bypassing the need to compute full log-signature. Each sequence is embedded using a frozen protein language model, and the resulting representation is passed through a small regression head to produce a scalar prediction $\hat{t}$. Training minimises a Huber loss between $\hat{t}$ and $t^\star$. This architecture is shared across both settings; only the training data and the construction of R differ.

3 Results

3.1 Datasets

To ensure generalizability of our results, we trained surrogate models to predict log-signatures conditioned on protein sequence and temperature using the mdCATH dataset (Mirarchi et al., 2024). mdCATH contains 5,398 diverse protein domains modeled with a state-of-the-art classical force field and simulated in five replicates of 500 ns at temperatures of 320, 348, 379, 413, and 450 K. We use this dataset to evaluate the ability of log-signatures to encode temperature-dependent and time-dependent variations in protein molecular dynamics across diverse structural classes.

To demonstrate the design of peptide binders that induce specific conformational regimes, we generated 10 ns molecular dynamics trajectories of peptides in complex with the p53 R175H mutant protein. This set includes previously studied p53 binders as well as newly generated peptides from BoltzGen (Stark et al., 2025a). Protein-peptide complex structures were predicted using AlphaFold2 (Jumper et al., 2021), relaxed, and subjected to explicit-solvent molecular dynamics simulations. As shown in Figure 3b, these trajectories exhibit measurable geometric variation between wild-type and peptide-bound mutant conformations even at the 10 ns timescale.

3.2 Log-signatures capture protein dynamics across timescales

We randomly selected five protein domains from the mdCATH dataset (Mirarchi et al., 2024) and computed truncated log-signatures up to depth 2, uniformly sampling 50 C_α backbone atoms and encoding their trajectories across simulation frames. For each trajectory, we computed the magnitude of the associated Lévy area terms across different temperatures and temporal windows. As shown in Figure 3a and 3b, the resulting non-zero mean Lévy areas indicate that log-signatures capture persistent, directional components of protein motion that vary systematically with both temperature and timescale. In the p53 system, log-signature distances further distinguish peptide-bound mutant and wild-type trajectories even at the 10 ns simulation timescale, as shown in Figure 3c, with corresponding geometric differences reflected in the radius of gyration (Figure 3d).

3.3 Surrogate Modeling of Temperature-Induced Protein Conformational Shifts in Lie-Algebraic Log-Signature Space

As illustrated in Figure 4a, for each protein in the mdCATH dataset, we quantify shift in conformation induced by temperature change by computing the signed projection $t^\star$ of the dynamical shift between low-temperature (320 K) and high-temperature (450 K) molecular dynamics trajectories. We uniformly subsample 50 Cα atoms along the protein backbone and extract up to 10 ns of aligned frames per replicate across 5 independent simulation replicates at each temperature. Depth-2 log-signatures

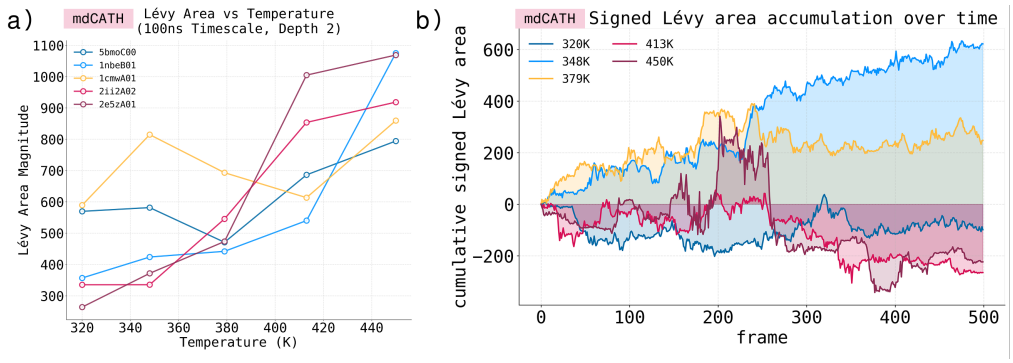

Figure 2: Log-signatures capture protein dynamics across timescales. (a) Lévy area magnitude computed from five randomly selected MD trajectories from the mdCATH database, demonstrating that log-signatures capture directional, non-equilibrium components of motion across diverse protein domains. (b) Cumulative Lévy area magnitude as a function of temperature, showing that log-signatures encode temperature-dependent timescale information.

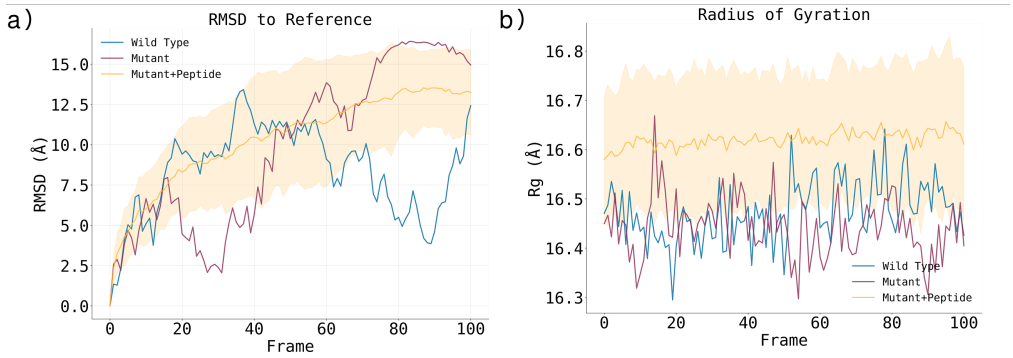

Figure 3: (a) Log-signature RMSD to a reference trajectory computed for peptide-bound mutant and wild-type p53, together with the unbound target, demonstrating that log-signatures distinguish dynamical regimes even on the 10 ns timescale. (b) Radius of gyration (Rg) for a subset of trajectories in p53 dataset, illustrating corresponding geometric differences in conformational compactness across trajectories.

are computed for each replicate trajectory and aggregated via the Fréchet mean under the depth-2 BCH approximation.

To enable sequence-conditioned prediction of $t^\star$ without requiring molecular dynamics simulation, we train a supervised surrogate on the mdCATH dataset using the frozen `esm2_t33_650M_UR50D` encoder (Lin et al., 2023). As shown in Table 1, the mdCATH surrogate achieves a Spearman correlation of 0.605 and Pearson correlation of 0.680 under 5-fold cross-validation, indicating that the model reliably captures conformational shift induced by temperature using just the sequence information.

3.4 SURROGATE MODELING OF PEPTIDE-INDUCED CONFORMATIONAL SHIFTS IN LIE-ALGEBRAIC LOG-SIGNATURE SPACE

For the p53 protein-peptide system, we quantify the conformational shift induced by peptide mutation relative to the wild-type complex by computing the signed projection $t^\star$ of the dynamical shift between each mutant and wild-type peptide-protein trajectory. All Cα atoms of the protein are retained, and up to 10 ns of aligned frames are extracted per replicate across independent simulation replicates for each complex. Depth-2 log-signatures are computed for each replicate trajectory and aggregated via the Fréchet mean under the depth-2 BCH approximation. To enable sequence-conditioned prediction of $t^\star$ without requiring molecular dynamics simulation, we train a supervised surrogate on the p53 peptide dataset using the frozen `esm2_t12_35M_UR50D` encoder (Lin et al., 2023). As shown in

Task	Metric	Result
p53-Peptide Binder	Pearson r	0.230 ± 0.105
	Spearman ρ	0.240 ± 0.102
MD-CATH	Pearson r	0.680 ± 0.042
	Spearman ρ	0.605 ± 0.034

Table 1: Sequence-level correlation between surrogate-predicted and ground-truth dynamical scores. For p53, all Cα atoms were used to compute the log signature. The mdCATH result is reported for the 50 Cα atom subsampling case.

Table 1, the p53 surrogate achieves a modest Spearman correlation of 0.240 and Pearson correlation of 0.230 under 5-fold cross-validation (Table 1), reflecting the limited training set of approximately 400 peptide sequences with only 3 simulation replicates each, constrained by the computational cost of generating 10 ns molecular dynamics trajectories per complex.

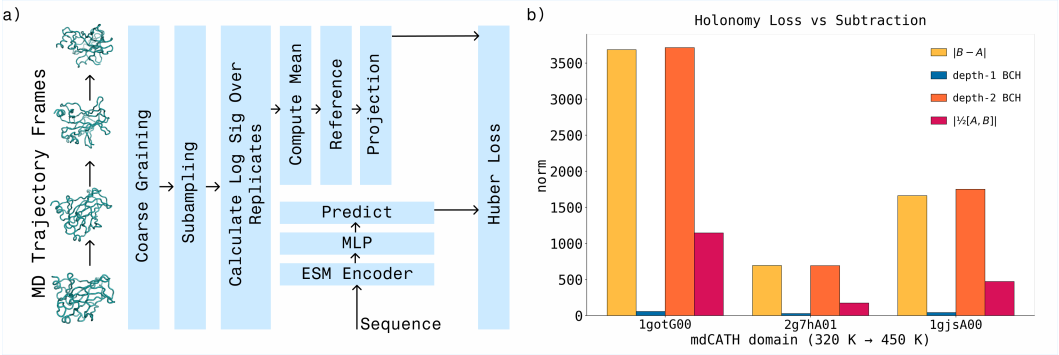

Figure 4: Surrogate models for predicting log-signatures. (a) Schematic of the surrogate architecture. Log-signatures are computed from molecular dynamics trajectories and projected along a global reference direction to obtain a scalar conformational regression target. (b) Comparing naive subtraction of log-signatures against the components of structured holonomy loss.

3.4.1 Dynamics-Conditioned Sequence Design for Enhanced Thermostability via Surrogate-Guided Discrete Flow Matching Generation

To evaluate whether dynamics-conditioned guidance can steer sequence generation toward thermostable variants, we trained a surrogate model on mdCATH simulations to predict log-signatures that capture the shift of conformational dynamics between low (320K) and high (450K) temperatures. The objective was to generate sequences whose dynamical profiles remain conserved across temperatures, reflecting robust folding behavior under thermal perturbation. Using this surrogate to guide discrete flow sampling, we generated a set of 48 50-mer sequences predicted to maintain stable dynamics at elevated temperature. As shown in Figure 5b, LieFlower-guided sampling produces sequences with consistently higher predicted stability scores compared to the unguided baseline, indicating that the surrogate model effectively biases generation. To further assess thermostability beyond surrogate predictions, we evaluated generated sequences using DeepAbst (Jung et al., 2023), a model trained to predict protein melting temperatures. We observe a modest increase in predicted melting temperature for sequences generated under dynamics-conditioned guidance relative to unguided designs, suggesting that conditioning on dynamical signatures at elevated temperature translates to measurable improvements in predicted thermostability. While the observed shift is moderate, these results provide evidence that dynamics guidance, by representing molecular dynamics trajectories as log-signatures, offers a viable route toward desired conformation-directed design.

3.4.2 LieFlower Generates Peptide Binders that Stabilize Mutant p53 Dynamics

p53 is a central tumor suppressor whose R175H mutation disrupts the native folding landscape and shifts the conformational ensemble toward dysfunctional states (Chiang et al., 2021). Prior work has suggested that peptide binders capable of stabilizing wild-type-like conformational dynamics may

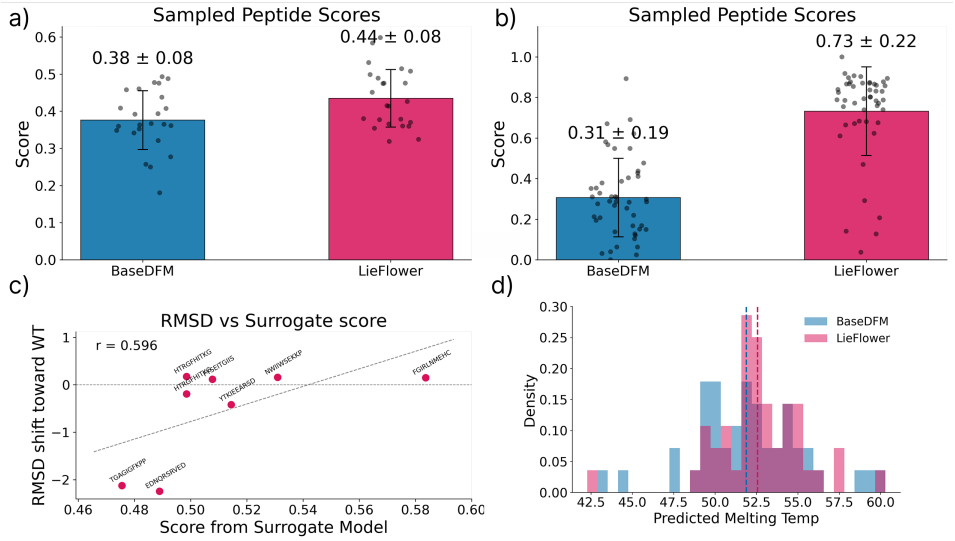

Figure 5: Surrogate-guided generation induces desired dynamics. (a) Surrogate model predictions of the peptide-induced dynamics signature in p53. Samples generated with guidance exhibit dynamics signatures closer to the target, whereas samples from the baseline flow matching model yield poorer predictions. (b) Analogous surrogate model predictions for the thermostability signature for samples of a 50mer protein, showing improved agreement when guidance is applied. (c) Correlation between surrogate-predicted scores and shift toward wild-type stability (Pearson $r = 0.596$), indicating that higher surrogate scores are associated with greater recovery of wild-type-like conformation upon peptide binding to mutant p53. (d) Distribution of predicted melting temperatures for samples generated with and without guidance ($n = 48$ each), demonstrating a modest improvement in predicted thermal stability when the guidance signal is incorporated.

partially restore function by reshaping this ensemble rather than enforcing a single static structure (Lei et al., 2021). Motivated by this view, we trained a surrogate model to predict the log-signature of p53 conformational dynamics induced by peptide binding and used this model to guide peptide generation toward signatures associated with wild-type p53. As shown in Figure 5a, peptides sampled from the dynamics-conditioned discrete flow achieve higher predicted scores than unguided samples, where higher scores correspond to greater similarity to wild-type conformational log signature. This confirms that surrogate-guided sampling can generate peptide sequences predicted to induce wild-type-like dynamics in mutant p53. To validate these predictions beyond the surrogate model, we selected the top 24 generated peptide sequences and folded protein-peptide complexes using AlphaFold (Jumper et al., 2021). Predicted structures were relaxed, and the top 8 complexes were subjected to 10ns molecular dynamics simulations. As shown in Figure 5c, a subset of LieFlower-generated peptides produces conformational signatures that are shifted toward the wild-type reference relative to the mutant baseline. These results indicate that conditioning generation on log-signatures of conformational dynamics provides sequence-level control over the target conformational ensemble, offering a complementary route to modulating challenging targets where static structure alone is insufficient to modulate functional behavior.

4 Discussion

We show that LieFlower guides protein sequence design using signals derived directly from molecular dynamics rather than static structural proxies. We find that log-signatures capture temperature- and timescale-dependent features of conformational motion, and that projecting these dynamics along a holonomy-derived reference direction provides a practical control signal for guiding discrete generative models. At the same time, our framework has limitations: the surrogate compresses complex trajectories, truncated log-signatures may miss higher-order or long-timescale effects, and guidance relies on predicted rather than observed dynamics, which can introduce bias outside the training distribution. Despite these caveats, our results suggest that explicit, sequence-level control

of protein dynamics is achievable within a discrete generative framework, offering a path beyond structure-centric design.

REFERENCES

Osama Abdin, Satra Nim, Han Wen, and Philip M Kim. Pepnn: a deep attention model for the identification of peptide binding sites. *Communications biology*, 5(1):503, 2022.

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

Barbora Barancikova, Zhuoyue Huang, and Cristopher Salvi. Sigdiffusions: Score-based diffusion models for time series via log-signature embeddings, 2025. URL https://arxiv.org/abs/2406.10354.

Marco Bertolini, Tuan Le, and Djork-Arné Clevert. Diffusion generative modeling on lie group representations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Jom8tNYuQI.

Suhaas Bhat, Kalyan Palepu, Lauren Hong, Joey Mao, Tianzheng Ye, Rema Iyer, Lin Zhao, Tianlai Chen, Sophia Vincoff, Rio Watson, et al. De novo design of peptide binders to conformationally diverse targets with contrastive language modeling. *Science Advances*, 11(4):eadr8638, 2025.

Anton Bushuiev, Roman Bushuiev, Petr Kouba, Anatolii Filkin, Marketa Gabrielova, Michal Gabriel, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, and Josef Sivic. Learning to design protein-protein interactions with enhanced generalization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xcMmebCT7s.

Kuo-Tsai Chen. Iterated integrals and exponential homomorphisms. *Proceedings of The London Mathematical Society*, pp. 502–512, 1954. URL https://api.semanticscholar.org/CorpusID:121288919.

Leo Tianlai Chen, Zachary Quinn, Madeleine Dumas, Christina Peng, Lauren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sophia Vincoff, Lin Zhao, et al. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, pp. 1–9, 2025a.

Tong Chen, Yinuo Zhang, and Pranam Chatterjee. Areuredi: Annealed rectified updates for refining discrete flows with multi-objective guidance. *arXiv preprint arXiv:2510.00352*, 2025b.

Tong Chen, Yinuo Zhang, Sophia Tang, and Pranam Chatterjee. Multi-objective-guided discrete flow matching for controllable biological sequence design. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025c. URL https://openreview.net/forum?id=8YIMLoHP9J.

Yen-Ting Chiang, Yi-Chung Chien, Yu-Heng Lin, Hui-Hsuan Wu, Dung-Fang Lee, and Yung-Luen Yu. The function of the mutant p53-r175h in cancer. *Cancers*, 13(16):4088, 2021.

Oscar Davis, Samuel Kessler, Mircea Petrache, İsmail İ Ceylan, Michael Bronstein, and Avishek J Bose. Fisher flow matching for generative modeling over discrete data. *Advances in Neural Information Processing Systems*, 37:139054–139084, 2024.

Theo Diamantakis, Darryl D. Holm, and Grigorios A. Pavliotis. Variational principles on geometric rough paths and the lévy area correction. *SIAM Journal on Applied Dynamical Systems*, 22(2): 1182–1218, June 2023. ISSN 1536-0040. doi: 10.1137/22m1522164. URL http://dx.doi.org/10.1137/22M1522164.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385, 2024.

Ben Hambly and Terry Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, 171(1):109–167, March 2010. ISSN 0003-486X. doi: 10.4007/annals.2010.171.109. URL http://dx.doi.org/10.4007/annals.2010.171.109.

Lauren Hong, Tianzheng Ye, Tian Z Wang, Divya Srijay, Howard Liu, Lin Zhao, Rio Watson, Sophia Vincoff, Tianlai Chen, Kseniia Kholina, et al. Programmable protein stabilization with language model-derived peptide guides. *Nature Communications*, 16(1):3555, 2025.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Felix Jung, Kevin Frey, David Zimmer, and Timo Mühlhaus. Deepstabp: A deep learning approach for the prediction of thermal protein stability. *International Journal of Molecular Sciences*, 24 (8), 2023. ISSN 1422-0067. doi: 10.3390/ijms24087444. URL https://www.mdpi.com/1422-0067/24/8/7444.

Jiangtao Lei, Mengqiang Cai, Yun Shen, Dongdong Lin, and Xiaohua Deng. Molecular dynamics study on the inhibition mechanisms of reacp53 peptide for p53–r175h mutant aggregation. *Physical Chemistry Chemical Physics*, 23(40):23032–23041, 2021.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Terry Lyons. Rough paths, signatures and the modelling of functions on streams, 2014. URL https://arxiv.org/abs/1405.4537.

T.J. Lyons, M. Caruana, and T. Lévy. *Differential Equations Driven by Rough Paths: École D'été de Probabilités de Saint-Flour XXXIV-2004*. Number no. 1908 in Differential Equations Driven by Rough Paths: École D'été de Probabilités de Saint-Flour XXXIV-2004. Springer, 2007. ISBN 9783540712848. URL https://books.google.com/books?id=HWMrAAAAYAAJ.

Antonio Mirarchi, Toni Giorgino, and Gianni De Fabritiis. mdcath: A large-scale md dataset for data-driven computational biophysics. *Scientific Data*, 11(1):1299, 2024.

Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, et al. One-shot design of functional protein binders with bindcraft. *Nature*, pp. 1–10, 2025.

Tiexin Qin, Mengxu Zhu, Chunyang Li, Terry Lyons, Hong Yan, and Haoliang Li. Deep signature: Characterization of large-scale molecular dynamics, 2025. URL https://arxiv.org/abs/2410.02847.

Rui Qing, Fei Tao, Pranam Chatterjee, Gaojie Yang, Qiuyi Han, Haeyoon Chung, Jun Ni, Bernhard P Suter, Jan Kubicek, Barbara Maertens, et al. Non-full-length water-soluble cxcr4qty and ccr5qty chemokine receptors: Implication for overlooked truncated but functional membrane receptors. *Iscience*, 23(12), 2020.

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

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167, 2024.

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 *Forty-first International Conference on Machine Learning*, 2024.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, Mateo Reveiz, Roman Bushuiev, Tomáš Pluskal, Josef Sivic, Karsten Kreis, Arash Vahdat, Shamayeeta Ray, Jonathan T. Goldstein, Andrew Savinov, Jacob A. Hambalek, Anshika Gupta, Diego A. Taquiri-Diaz, Yaotian Zhang, A. Katherine Hatstat, Angelika Arada, Nam Hyeong Kim, Ethel Tackie-Yarboi, Dylan Boselli, Lee Schnaider, Chang C. Liu, Gene-Wei Li, Denes Hnisz, David M. Sabatini, William F. DeGrado, Jeremy Wohlwend, Gabriele Corso, Regina Barzilay, and Tommi Jaakkola. Boltzgen: Toward universal binder design. *bioRxiv*, 2025a. doi: 10.1101/2025.11.20.689494. URL https://www.biorxiv.org/content/early/2025/11/24/2025.11.20.689494.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy John O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, et al. Boltzgen: Toward universal binder design. *bioRxiv*, pp. 2025–11, 2025b.

Sophia Tang, Yinuo Zhang, and Pranam Chatterjee. Peptune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=FQoy1Y1Hd8.

Sophia Tang, Yinuo Zhang, Alexander Tong, and Pranam Chatterjee. Gumbel-softmax score and flow matching for discrete biological sequence generation. In *ICLR 2025 Workshop on AI for Nucleic Acids*, 2025b. URL https://openreview.net/forum?id=ITpCmDhSfu.

Sophia Tang, Yuchen Zhu, Molei Tao, and Pranam Chatterjee. Tr2-d2: Tree search guided trajectory-aware fine-tuning for discrete diffusion. *arXiv preprint arXiv:2509.25171*, 2025c.

Vladimir N Uversky. How to drug a cloud? targeting intrinsically disordered proteins. *Pharmacological Reviews*, 77(2):100016, 2025.

Sophia Vincoff, Shrey Goel, Kseniia Kholina, Rishab Pulugurta, Pranay Vure, and Pranam Chatterjee. Fuson-plm: a fusion oncoprotein-specific language model via adjusted rate masking. *Nature Communications*, 16(1):1436, 2025.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Rio Watson, Kishan Patel, Tong Chen, and Pranam Chatterjee. Targeting aggregating proteins with language model-designed degraders. 2025.

Chengxin Zhang, Xi Zhang, Lydia Freddolino, and Yang Zhang. Biolip2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 52(D1): D404–D412, 2024.

Appendix

A Model Architectures and Training Details

A.1 PepDFM Base Model

For all dynamics-conditioned sequence generation experiments, we employ **PepDFM** as the underlying peptide generator. PepDFM is an unconditional discrete flow matching (DFM) model introduced in (Chen et al., 2025c), trained to generate diverse peptide sequences while preserving realistic amino acid statistics and length distributions. In LieFlower, PepDFM provides a pretrained token-level velocity field over peptide sequences, which we treat as a fixed base sampler and guide using surrogate-predicted log-signatures during inference.

PepDFM operates in the discrete flow matching framework (Gat et al., 2024), modeling peptide generation as a continuous-time Markov chain (CTMC) over sequences $x \in \mathcal{A}^L$. The model learns a time-dependent, factorized velocity field $u_{\theta,t}$ that transports probability mass from a simple prior distribution (typically a fully masked or uniform sequence) to the data distribution along a mixture probability path. This formulation provides an explicit notion of token-level transition rates, enabling principled external guidance at each sampling step.

A.1.1 Model Architecture

The PepDFM backbone is a time-conditioned U-Net-style convolutional architecture. Amino acid identities are embedded into a continuous representation, and time $t \in [0, 1]$ is embedded separately and injected through dense conditioning layers at multiple stages of the network. The architecture consists of five convolutional blocks with varying dilation rates to capture both local and longer-range sequence dependencies. Skip connections are used to preserve fine-grained token information across resolution scales. The final output layer produces token-level logits corresponding to the factorized velocity components required by discrete flow matching.

PepDFM uses a polynomial convex schedule with exponent 2.0 for the mixture probability path, which defines how probability mass interpolates between source and target distributions over time. This schedule was selected to balance early exploration with stable convergence near the data manifold.

A.1.2 Training Configuration

PepDFM was trained on a curated peptide dataset assembled from PepNN, BioLip2, and PPIRef (Abdin et al., 2022; Zhang et al., 2024; Bushuiev et al., 2024). The dataset includes peptide sequences ranging from 6 to 50 amino acids in length and was split into training, validation, and test sets using an 80/10/10 ratio. Training was performed using the generalized KL divergence loss for discrete flow matching, which provides both a principled training objective and an evaluation metric for the learned velocity field.

The model was trained for 200 epochs with a batch size of 512 using the Adam optimizer and a cosine learning rate schedule with warmup. Dynamic batching was employed to efficiently handle variable-length peptide sequences. The final checkpoint was selected based on the lowest validation loss. As reported in (Chen et al., 2025c), PepDFM achieves low generalized KL loss on held-out data and generates peptides with high diversity, as measured by Hamming distance and Shannon entropy relative to the training distribution.

A.1.3 Use in LieFlower

In this work, PepDFM is used without additional fine-tuning. During sampling, the pretrained PepDFM velocity field defines the base discrete-time dynamics of the CTMC over peptide sequences. LieFlower introduces dynamics-level control by conditioning the sampling process on a target log-signature ℓ_{target}, modifying the effective vector field through surrogate-guided conditioning while preserving the underlying DFM structure.

This separation between the base generator (PepDFM) and the dynamics-conditioning mechanism allows LieFlower to leverage a high-quality, biologically grounded peptide prior while introducing explicit control over induced protein motion. Importantly, this design avoids retraining the base model and ensures that guided samples remain close to the peptide manifold learned by PepDFM.

B ALGORITHMS

Here, we provide pseudocode for training a LieFlower model in Algorithm 1 and for dynamics-conditioned sampling in Algorithm 2. The training loop jointly optimizes the conditional flow field, the surrogate dynamics predictor, and the holonomy-based control objective. The inference procedure integrates the learned reverse-time vector field to generate peptides that satisfy a specified target log-signature.

Algorithm 1 Training LieFlower

Input: dataset $\mathcal{D} = \{(y^{(i)}, \ell^{(i)})\}_{i=1}^{N}$ ▷ peptides and their log-signatures
  surrogate model Φ_ψ ▷ predicts induced log-signature from sequence
  conditional flow field v_θ
  target flow field u_t for flow matching
  noising kernels $q_t(x_t \mid x_1)$
Hyperparameters: $\lambda_{\text{Hol}}, \lambda_{\text{dyn}}, \eta$
while not converged **do**
  for each minibatch $\{(y^{(i)}, \ell^{(i)})\}$ **do**
    Initialize minibatch losses $\mathcal{L}_{\text{FM}}, \mathcal{L}_{\text{Hol}}, \mathcal{L}_{\text{dyn}} \leftarrow 0$
    for each sequence $y^{(i)}$ **do**
      $x_1^{(i)} \leftarrow \text{Encode}(y^{(i)})$ ▷ discrete or relaxed sequence representation
      Sample $t \sim \text{Uniform}(0, 1)$ ▷ random time for flow matching
      Sample corrupted sequence $x_t \sim q_t(x_t \mid x_1^{(i)})$
      Compute target flow $u_t(x_t \mid x_1^{(i)})$
      Compute predicted field $v_\theta(x_t, t, \ell^{(i)})$
      $\mathcal{L}_{\text{FM}} \mathrel{+}= \|v_\theta(x_t, t, \ell^{(i)}) - u_t(x_t \mid x_1^{(i)})\|^2$
      $\widehat{\ell}^{(i)} \leftarrow \Phi_\psi(y^{(i)})$ ▷ predict induced dynamics
      $g_{\text{induced}} \leftarrow \exp(\widehat{\ell}^{(i)}), \quad g_{\text{target}} \leftarrow \exp(\ell^{(i)})$
      $\Delta^{(i)} \leftarrow \log\big(g_{\text{induced}}^{-1} g_{\text{target}}\big)$ ▷ Lie algebra discrepancy
      $\mathcal{L}_{\text{Hol}} \mathrel{+}= \|\Delta^{(i)}\|_{\mathfrak{g}}^2$
      $\mathcal{L}_{\text{dyn}} \mathrel{+}= \|\widehat{\ell}^{(i)} - \ell^{(i)}\|_{\mathfrak{g}}^2$ ▷ supervised surrogate loss
    end for
    $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{Hol}}\mathcal{L}_{\text{Hol}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}}$
    $(\theta, \psi) \leftarrow (\theta, \psi) - \eta \nabla_{(\theta,\psi)} \mathcal{L}_{\text{total}}$
  end for
end while
return (θ^*, ψ^*)

Algorithm 2 Inference with LieFlower

Input: trained models (v_θ, Φ_ψ), target log-signature ℓ_{target}, prior $p_1(x_1)$
Goal: generate peptide $\widehat{y}$ satisfying $\Phi_\psi(\widehat{y}) \approx \ell_{\text{target}}$
Sample initial state $x_1 \sim p_1(x_1)$ ▷ maximally corrupted or fully masked
for $t = 1, 1 - \Delta t, \ldots, 0$ **do**
  Compute conditional vector field $v_t \leftarrow v_\theta(x_t, t, \ell_{\text{target}})$
  $x_{t-\Delta t} \leftarrow x_t - \Delta t \cdot v_t$ ▷ reverse-time integration
end for
Decode x_0 into peptide $\widehat{y}$
Optional: compute predicted dynamics $\Phi_\psi(\widehat{y})$
return $\widehat{y}$

Algorithm 3 Computing Truncated Log-Signatures from MD Trajectories

Input: MD trajectory $X : [0, T] \to \mathbb{R}^d$ sampled at times $\{t_0, \ldots, t_K\}$
  truncation order m
  interpolation rule (*e.g.*, piecewise linear)
Output: truncated log-signature $\ell_m(X) \in \mathfrak{g}_{\leq m}$
Interpolate discrete samples to a continuous path $\widehat{X}$ ▷ ensures well-defined iterated integrals
Initialize signature tensors:

$$S^{(0)} \leftarrow 1, \quad S^{(k)} \leftarrow 0 \text{ for all } k = 1, \ldots, m$$

for $k = 1$ **to** m **do**
  for $t_0 < t_1 < \cdots < t_k < T$ **do**
    Compute differential increments $d\widehat{X}_{t_j}$
    Accumulate

$$S^{(k)} \mathrel{+}= d\widehat{X}_{t_1} \otimes \cdots \otimes d\widehat{X}_{t_k}$$

  end for
end for
Assemble truncated signature:

$$S_m(X) = \left(S^{(0)}, S^{(1)}, \ldots, S^{(m)}\right)$$

Compute truncated log-signature:

$$\ell_m(X) \leftarrow \log(S_m(X))$$

▷ via truncated Baker-Campbell-Hausdorff expansion
Project coefficients onto the free Lie algebra basis ▷ ensures representation in $\mathfrak{g}_{\leq m}$
return $\ell_m(X)$

C Theoretical Proofs

In this section we provide theoretical guarantees for LieFlower. We first establish the consistency of the conditional flow matching objective, then analyze the Lie algebraic holonomy loss, and finally study representational completeness, truncation stability, gradients, sampling convergence, and the optimality of the holonomy-regularized control objective.

Throughout, we adopt the notation from the main text. In particular:

- $\mathcal{A}$ is a finite alphabet and $\mathcal{A}^*$ is the set of all finite sequences.
- $y \in \mathcal{A}^*$ denotes a peptide sequence and x_1 its discrete representation.
- $q_t(x_t \mid x_1)$ is a forward noising process over discrete sequences for $t \in [0, 1]$.
- $u_t(x_t \mid x_1)$ is the target vector field specified by the flow matching construction.
- $v_\theta(x_t, t, \ell)$ is the parametric vector field used in LieFlower, conditioned on a log-signature $\ell \in \mathfrak{g}_{\leq p}$.
- $\Phi_\psi : \mathcal{A}^* \to \mathfrak{g}_{\leq p}$ is the surrogate dynamics predictor that approximates induced log-signatures.
- $\exp$ and $\log$ denote the Lie exponential and logarithm between the truncated free Lie algebra $\mathfrak{g}_{\leq p}$ and its associated nilpotent Lie group $G_{\leq p}$.

C.1 Conditional Flow Matching Consistency

We first show that the conditional flow matching objective identifies the correct conditional flow field when the model class is sufficiently rich.

Theorem C.1 (Conditional Flow Matching Consistency)**.** *Fix a truncation order m and a log-signature $\ell \in \mathfrak{g}_{\leq p}$. Let $q_t(x_t \mid x_1)$ be a forward noising process on sequences with associated conditional path measure $\pi_t(x_t \mid \ell)$ induced by (x_1, ℓ) under the data distribution. Let $u_t(x_t \mid x_1)$ be the flow matching target field that transports $\pi_t(\cdot \mid \ell)$ to the conditional data distribution at $t = 0$.*

Consider the conditional flow matching loss

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_\ell \mathbb{E}_t \mathbb{E}_{x_1|\ell} \mathbb{E}_{x_t \sim q_t(\cdot|x_1)} \left[\| v_\theta(x_t, t, \ell) - u_t(x_t \mid x_1) \|^2 \right]. \tag{13}$$

If the model class $\{v_\theta\}$ is rich enough to represent $u_t(x_t \mid x_1)$ as a function of (x_t, t, ℓ) and if a minimizer θ^* of $\mathcal{L}_{\mathrm{FM}}$ exists, then

$$\mathcal{L}_{\mathrm{FM}}(\theta^*) = 0 \quad \Longrightarrow \quad v_{\theta^*}(x_t, t, \ell) = u_t(x_t \mid x_1) \tag{14}$$

almost everywhere under the joint law of (ℓ, t, x_1, x_t). Furthermore, the reverse-time flow induced by $v_{\theta^*}(\cdot, \cdot, \ell)$ transports the prior at $t = 1$ to the conditional data distribution $p(x_1 \mid \ell)$ at $t = 0$.

Proof. We first expand the loss in equation 13. For notational convenience, write

$$Z = (\ell, t, x_1, x_t)$$

and let $\mathbb{P}$ be the joint distribution of Z induced by the data distribution over (x_1, ℓ), the time sampling distribution over t, and the forward noising kernels $q_t(\cdot \mid x_1)$.

Then the loss can be written as

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \int \| v_\theta(x_t, t, \ell) - u_t(x_t \mid x_1) \|^2 \, d\mathbb{P}(\ell, t, x_1, x_t). \tag{15}$$

The integrand is nonnegative everywhere. If $\mathcal{L}_{\mathrm{FM}}(\theta^*) = 0$ for some θ^*, then by nonnegativity of the integrand we must have

$$\| v_{\theta^*}(x_t, t, \ell) - u_t(x_t \mid x_1) \|^2 = 0 \tag{16}$$

for $\mathbb{P}$-almost every (ℓ, t, x_1, x_t). This implies

$$v_{\theta^*}(x_t, t, \ell) = u_t(x_t \mid x_1) \tag{17}$$

$\mathbb{P}$-almost surely. In other words, the learned vector field coincides almost everywhere with the target flow under the joint sampling procedure used to define the loss.

Next we recall the standard flow matching construction in continuous time for generative modeling. For each fixed ℓ, the field $u_t(\cdot \mid x_1)$ is known to be a solution of the continuity equation

$$\partial_t \pi_t(x \mid \ell) + \nabla_x \cdot (\pi_t(x \mid \ell) \, u_t(x \mid \ell)) = 0, \tag{18}$$

with terminal condition at $t = 1$ given by the prior distribution and initial condition at $t = 0$ given by the conditional data distribution $p(x_1 \mid \ell)$. Under regularity assumptions on u_t (such as local Lipschitz continuity and linear growth), the continuity equation admits a unique weak solution, and the characteristic flow defined by

$$\frac{d}{dt} x_t = u_t(x_t \mid \ell) \tag{19}$$

transports the terminal law at $t = 1$ to the initial law at $t = 0$.

Since v_{θ^*} agrees with u_t almost everywhere under $\pi_t(\cdot \mid \ell)$ and is assumed to share the same regularity, the continuity equation with v_{θ^*} in place of u_t has the same weak solution. As a consequence, the reverse-time flow defined by

$$\frac{d}{dt} x_t = v_{\theta^*}(x_t, t, \ell) \tag{20}$$

transports the prior at $t = 1$ to the conditional data distribution at $t = 0$. This establishes the stated properties. □

C.2 Lie Algebraic Holonomy Consistency

We next analyze the holonomy loss and show that it is minimized when the surrogate dynamics predictor matches the target log-signature in the truncated free Lie algebra.

Theorem C.2 (Holonomy Consistency)**.** *Let $\mathfrak{g}_{\leq p}$ be a finite-dimensional free Lie algebra of step m and $G_{\leq p}$ its associated simply connected nilpotent Lie group. The truncated exponential map*

$$\exp : \mathfrak{g}_{\leq p} \to G_{\leq p} \tag{21}$$

is a global diffeomorphism with inverse $\log : G_{\leq p} \to \mathfrak{g}_{\leq p}$.

For any target log-signature $\ell_{\text{target}} \in \mathfrak{g}_{\leq p}$ and surrogate prediction $\widehat{\ell} = \Phi_\psi(y)$, define the group elements

$$g_{\text{induced}} = \exp(\widehat{\ell}), \quad g_{\text{target}} = \exp(\ell_{\text{target}}), \tag{22}$$

and the holonomy discrepancy

$$\Delta(y; \ell_{\text{target}}) = \log\big(g_{\text{induced}}^{-1} g_{\text{target}}\big) \in \mathfrak{g}_{\leq p}. \tag{23}$$

Then, for any norm $\|\cdot\|_{\mathfrak{g}}$ on $\mathfrak{g}_{\leq p}$,

$$\Delta(y; \ell_{\text{target}}) = 0 \quad \Longleftrightarrow \quad \widehat{\ell} = \ell_{\text{target}}. \tag{24}$$

In particular, the holonomy loss

$$\mathcal{L}_{\text{Hol}}(\psi) = \mathbb{E}_{\ell_{\text{target}}} \mathbb{E}_{y \sim p_\theta(\cdot|\ell_{\text{target}})} \left[\|\Delta(y; \ell_{\text{target}})\|_{\mathfrak{g}}^2 \right] \tag{25}$$

is minimized at zero if and only if $\Phi_\psi(y) = \ell_{\text{target}}$ almost surely under the conditional law.

Proof. Since $\mathfrak{g}_{\leq p}$ is nilpotent of step m, the Baker-Campbell-Hausdorff (BCH) series truncates at finite order. The Lie group $G_{\leq p}$ is connected and simply connected, and the exponential map

$$\exp : \mathfrak{g}_{\leq p} \to G_{\leq p} \tag{26}$$

is a global diffeomorphism with inverse $\log$.

We first show the implication $\Delta = 0 \Rightarrow \widehat{\ell} = \ell_{\text{target}}$. Suppose

$$\Delta = \log\Big(\exp(\widehat{\ell})^{-1} \exp(\ell_{\text{target}})\Big) = 0. \tag{27}$$

Applying the exponential map to both sides yields

$$\exp(\Delta) = \exp(0) = e_G, \tag{28}$$

where e_G is the identity in $G_{\leq p}$. On the other hand, by definition of Δ,

$$\exp(\Delta) = \exp(\widehat{\ell})^{-1} \exp(\ell_{\text{target}}). \tag{29}$$

Thus

$$\exp(\widehat{\ell})^{-1} \exp(\ell_{\text{target}}) = e_G, \tag{30}$$

which implies

$$\exp(\widehat{\ell}) = \exp(\ell_{\text{target}}). \tag{31}$$

Since $\exp$ is a bijection with inverse $\log$, we conclude

$$\widehat{\ell} = \ell_{\text{target}}. \tag{32}$$

For the converse implication, suppose $\widehat{\ell} = \ell_{\text{target}}$. Then $\exp(\widehat{\ell}) = \exp(\ell_{\text{target}})$ and

$$g_{\text{induced}}^{-1} g_{\text{target}} = \exp(\widehat{\ell})^{-1} \exp(\widehat{\ell}) = e_G. \tag{33}$$

Taking the logarithm,

$$\Delta = \log(e_G) = 0. \tag{34}$$

The holonomy loss is an expectation of a nonnegative quantity,

$$\|\Delta(y; \ell_{\text{target}})\|_{\mathfrak{g}}^2 \geq 0. \tag{35}$$

If $\mathcal{L}_{\text{Hol}}(\psi) = 0$, then this squared norm must be zero for almost every $(y, \ell_{\text{target}})$ under the joint law. By the norm property, this is equivalent to $\Delta(y; \ell_{\text{target}}) = 0$ almost surely. By the previous equivalence, this in turn is equivalent to $\Phi_\psi(y) = \ell_{\text{target}}$ almost surely. □

C.3 Representational Completeness of Conditional LieFlower

We now show that LieFlower is expressive enough to represent any conditional sequence distribution indexed by a log-signature, assuming sufficient capacity in the dynamics surrogate and the flow field.

Theorem C.3 (Representational Completeness). *Let μ be an arbitrary joint distribution on $\mathcal{A}^* \times \mathfrak{g}_{\leq p}$ with finite support in $\mathcal{A}^*$. For each $\ell \in \mathfrak{g}_{\leq p}$, let $\mu(\cdot \mid \ell)$ denote the conditional law on $\mathcal{A}^*$.*

Assume:

1. *The surrogate class $\{\Phi_\psi\}$ is dense in the space of measurable functions from $\mathcal{A}^*$ to $\mathfrak{g}_{\leq p}$ under $L^2(\mu)$.*
2. *The conditional flow class $\{v_\theta\}$ is dense in the space of measurable vector fields on (x_t, t, ℓ) relevant for flow matching.*
3. *The forward noising process $q_t(x_t \mid x_1)$ and corresponding flow matching target u_t are chosen such that for every conditional law $\mu(\cdot \mid \ell)$ there exists a corresponding flow u_t^ℓ transporting the prior to $\mu(\cdot \mid \ell)$.*

Then for every $\epsilon > 0$ there exist parameters (θ, ψ) such that:

$$\mathbb{E}_\ell \left[D_{\mathrm{TV}}(p_\theta(\cdot \mid \ell), \mu(\cdot \mid \ell)) \right] < \epsilon \tag{36}$$

and

$$\mathbb{E}_{(y,\ell)\sim\mu} \left[\|\Phi_\psi(y) - \ell\|_{\mathfrak{g}}^2 \right] < \epsilon. \tag{37}$$

Proof. We first approximate the dynamics map, then the conditional sequence distribution.

Step 1: Approximate ℓ by $\Phi_\psi(y)$. By assumption (1), the surrogate class $\{\Phi_\psi\}$ is dense in the space of measurable functions from $\mathcal{A}^*$ to $\mathfrak{g}_{\leq p}$ under $L^2(\mu)$. In particular, the function $F(y, \ell) := \ell$ (viewed as a function of y through the joint law) can be approximated arbitrarily well. More precisely, for any $\epsilon > 0$ there exists ψ such that

$$\mathbb{E}_{(y,\ell)\sim\mu} \left[\|\Phi_\psi(y) - \ell\|_{\mathfrak{g}}^2 \right] < \epsilon. \tag{38}$$

This establishes the second inequality in the theorem.

Step 2: Approximate conditional flows. For each fixed ℓ, consider the conditional law $\mu(\cdot \mid \ell)$ on $\mathcal{A}^*$. By assumption (3), the flow matching construction yields a target vector field $u_t^\ell(x_t)$ that transports the prior at $t = 1$ to $\mu(\cdot \mid \ell)$ at $t = 0$. By assumption (2), the model class $\{v_\theta\}$ is dense in the set of such vector fields, when viewed as functions of (x_t, t, ℓ). Hence, for any $\delta > 0$ there exists θ such that

$$\mathbb{E}_\ell \mathbb{E}_t \mathbb{E}_{x_1\sim\mu(\cdot|\ell)} \mathbb{E}_{x_t\sim q_t(\cdot|x_1)} \left[\|v_\theta(x_t, t, \ell) - u_t^\ell(x_t)\|^2 \right] < \delta. \tag{39}$$

By the same argument as in Theorem C.1, a small L^2 error in the flow field implies a small perturbation of the resulting probability path, under appropriate regularity of the dynamics. More concretely, standard stability results for continuity equations with respect to perturbations in the vector field imply that the marginal distributions at $t = 0$ induced by v_θ can be made arbitrarily close in total variation distance to those induced by u_t^ℓ. Thus, we can choose θ such that

$$\mathbb{E}_\ell \left[D_{\mathrm{TV}}(p_\theta(\cdot \mid \ell), \mu(\cdot \mid \ell)) \right] < \epsilon, \tag{40}$$

for any prescribed $\epsilon > 0$. Combining both steps yields the result. □

C.4 Stability under Log-Signature Truncation

We now study the impact of truncating the log-signature at order m. The following result is a direct specialization of classical rough path stability results to our setting.

Theorem C.4 (Truncation Stability). *Let $X : [0, T] \to \mathbb{R}^d$ be a path of finite p-variation for some $1 \leq p < \infty$. Let $\ell(X)$ denote its full log-signature and $\ell_m(X)$ its truncation in $\mathfrak{g}_{\leq p}$. Then for any compact subset $\mathcal{K}$ of paths in the p-variation topology and any order m, there exists a constant $C_{m,p,T}$ such that*

$$\|\ell(X) - \ell_m(X)\|_{\mathfrak{g}} \leq C_{m,p,T} \cdot \omega_X^{m+1}, \tag{41}$$

for all $X \in \mathcal{K}$, where ω_X is a control on the p-variation of X on $[0, T]$. In particular, $\ell_m(X)$ converges to $\ell(X)$ uniformly on $\mathcal{K}$ as $m \to \infty$.

Proof sketch. The signature $S(X)$ of a path of finite p-variation defines a group-like element in the tensor algebra. The log-signature $\ell(X)$ is obtained by applying the logarithm in the free Lie algebra. The truncation at order m discards all Lie brackets of degree greater than m.

Classical results in the theory of rough paths (for example, Lyons' universal limit theorem) show that iterated integrals of order k scale at most like ω_X^k for a suitable control ω_X on the p-variation of X. The BCH expansion for the logarithm writes $\ell(X)$ as a finite linear combination of iterated integrals at each degree. As a result, the contribution of all terms of degree greater than m can be bounded by a constant multiple of ω_X^{m+1}, uniformly over X in a compact subset $\mathcal{K}$ of the p-variation topology. This yields the stated inequality. Taking $m \to \infty$ then gives uniform convergence on $\mathcal{K}$. □

C.5 Gradient of the Holonomy Loss

We next derive the gradient of the holonomy loss with respect to the surrogate parameters ψ. The result is the Lie algebra analogue of hazard gradient computation in SBDG.

Theorem C.5 (Gradient of the Holonomy Loss)**.** *For a fixed target log-signature ℓ_{target} and a sample y, define $\widehat{\ell} = \Phi_\psi(y)$, $g_{\mathrm{induced}} = \exp(\widehat{\ell})$, $g_{\mathrm{target}} = \exp(\ell_{\mathrm{target}})$, and*

$$\Delta(\psi; y, \ell_{\mathrm{target}}) = \log\big(g_{\mathrm{induced}}^{-1} g_{\mathrm{target}}\big)\,. \tag{42}$$

Assume that exp *and* log *are differentiable, and adopt the squared norm holonomy loss*

$$\mathcal{L}_{\mathrm{Hol}}(\psi; y, \ell_{\mathrm{target}}) = \|\Delta(\psi; y, \ell_{\mathrm{target}})\|_{\mathfrak{g}}^2. \tag{43}$$

Then the gradient with respect to ψ is

$$\nabla_\psi \mathcal{L}_{\mathrm{Hol}}(\psi; y, \ell_{\mathrm{target}}) = 2\,\langle \Delta, \nabla_\psi \Delta \rangle_{\mathfrak{g}}\,, \tag{44}$$

where

$$\nabla_\psi \Delta = D\log\big(g_{\mathrm{induced}}^{-1} g_{\mathrm{target}}\big) \circ D\big(g_{\mathrm{induced}}^{-1} g_{\mathrm{target}}\big) \circ D\exp(\widehat{\ell}) \circ \nabla_\psi \Phi_\psi(y), \tag{45}$$

and D denotes the Fréchet derivative.

Proof. We proceed by the chain rule.

First, treat Δ as a function of $z = g_{\mathrm{induced}}^{-1} g_{\mathrm{target}} \in G_{\le p}$:

$$\Delta = \log(z). \tag{46}$$

The derivative of this mapping at z is the linear map

$$D\log(z) : T_z G_{\le p} \to \mathfrak{g}_{\le p}. \tag{47}$$

Second, view z as a function of g_{induced} with g_{target} fixed:

$$z = g_{\mathrm{induced}}^{-1} g_{\mathrm{target}}. \tag{48}$$

Infinitesimal variations $\delta g_{\mathrm{induced}}$ induce variations

$$\delta z = -g_{\mathrm{induced}}^{-1} \delta g_{\mathrm{induced}}\, g_{\mathrm{induced}}^{-1} g_{\mathrm{target}}, \tag{49}$$

which defines the derivative

$$D\big(g_{\mathrm{induced}}^{-1} g_{\mathrm{target}}\big) : T_{g_{\mathrm{induced}}} G_{\le p} \to T_z G_{\le p}. \tag{50}$$

Third, g_{induced} is a function of $\widehat{\ell}$ through the exponential:

$$g_{\mathrm{induced}} = \exp(\widehat{\ell}). \tag{51}$$

The derivative of exp at $\widehat{\ell}$ is the linear map

$$D\exp(\widehat{\ell}) : \mathfrak{g}_{\le p} \to T_{g_{\mathrm{induced}}} G_{\le p}. \tag{52}$$

Finally, $\widehat{\ell}$ depends on ψ through the surrogate $\Phi_\psi(y)$:

$$\widehat{\ell} = \Phi_\psi(y). \tag{53}$$

Its derivative with respect to ψ is

$$\nabla_\psi \Phi_\psi(y) : T_\psi \to \mathfrak{g}_{\leq p}. \tag{54}$$

Composing these maps, the derivative of Δ with respect to ψ is

$$\nabla_\psi \Delta = D\log(z) \circ D\big(g_{\text{induced}}^{-1} g_{\text{target}}\big) \circ D\exp(\widehat{\ell}) \circ \nabla_\psi \Phi_\psi(y), \tag{55}$$

with $z = g_{\text{induced}}^{-1} g_{\text{target}}$.

Now differentiate the scalar loss

$$\mathcal{L}_{\text{Hol}} = \langle \Delta, \Delta \rangle_{\mathfrak{g}}, \tag{56}$$

where $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$ denotes the inner product associated with $\| \cdot \|_{\mathfrak{g}}$. The chain rule gives

$$\nabla_\psi \mathcal{L}_{\text{Hol}} = 2 \langle \Delta, \nabla_\psi \Delta \rangle_{\mathfrak{g}}. \tag{57}$$

Substituting the expression for $\nabla_\psi \Delta$ yields the claimed formula. □

C.6 Convergence of the Reverse-Time LieFlower Sampler

We now analyze the reverse-time sampler defined by the learned vector field v_θ. The following result states that, under the regularity and consistency assumptions of Theorem C.1, the reverse-time integration converges to the desired conditional distribution.

Theorem C.6 (Sampler Convergence). *Fix a target log-signature ℓ and suppose that $v_{\theta^*}(\cdot, \cdot, \ell)$ satisfies the conditions of Theorem C.1, with corresponding conditional distribution $p(\cdot \mid \ell)$ at $t = 0$. Assume that v_{θ^*} is globally Lipschitz in x and continuous in t.*

Consider the reverse-time ordinary differential equation

$$\frac{d}{dt} x_t = v_{\theta^*}(x_t, t, \ell), \quad t \in [0, 1], \tag{58}$$

integrated backward from a random initial condition $x_1 \sim p_1(\cdot \mid \ell)$ at time $t = 1$. Then the solution x_0 at time $t = 0$ is distributed according to $p(\cdot \mid \ell)$. Furthermore, if we approximate the ODE with an explicit Euler scheme of step size Δt, the distribution of the discrete-time solution converges weakly to $p(\cdot \mid \ell)$ as $\Delta t \to 0$.

Proof. Under the Lipschitz and continuity assumptions on v_{θ^*}, the ODE admits a unique solution x_t for each initial condition x_1. The continuity equation associated with the flow

$$\partial_t \rho_t(x) + \nabla_x \cdot (\rho_t(x) v_{\theta^*}(x, t, \ell)) = 0 \tag{59}$$

has a unique weak solution given by the pushforward of the terminal law $p_1(\cdot \mid \ell)$ through the characteristic flow. By Theorem C.1, this weak solution satisfies $\rho_0 = p(\cdot \mid \ell)$, the desired conditional distribution.

The discrete-time explicit Euler scheme

$$x_{t_{k+1}} = x_{t_k} - \Delta t \, v_{\theta^*}(x_{t_k}, t_k, \ell), \quad t_k = 1 - k\Delta t, \tag{60}$$

converges uniformly on compact time intervals to the continuous solution as $\Delta t \to 0$ by standard ODE numerical analysis under the Lipschitz condition. By continuity of the mapping from paths to endpoint distributions, the law of the discrete-time solution at $t = 0$ converges weakly to the law of the continuous-time solution at $t = 0$, which is $p(\cdot \mid \ell)$. □

C.7 Optimality of Holonomy-Regularized Control

We finally analyze the holonomy-regularized control objective in the idealized setting where the surrogate is injective on the support of the sequence distribution.

Theorem C.7 (Holonomy-Regularized Control Optimality). *Let $\Phi : \mathcal{A}^* \to \mathfrak{g}_{\leq p}$ be a fixed injective map on a subset $\mathcal{Y} \subset \mathcal{A}^*$. Fix a target log-signature ℓ_{target} and consider distributions p supported on $\mathcal{Y}$. Define the holonomy cost*

$$J(p) = \mathbb{E}_{y \sim p} \left[\| \log(\exp(\Phi(y))^{-1} \exp(\ell_{\text{target}})) \|_{\mathfrak{g}}^2 \right]. \tag{61}$$

Assume there exists at least one $y^* \in \mathcal{Y}$ such that $\Phi(y^*) = \ell_{\text{target}}$. Then:

1. *$J(p) \geq 0$ for all p, with equality if and only if p is supported on the set*
$$\{y \in \mathcal{Y} : \Phi(y) = \ell_{\text{target}}\}.$$
2. *If Φ is injective and there is a unique y^* such that $\Phi(y^*) = \ell_{\text{target}}$, then the unique minimizer of $J(p)$ over distributions supported on $\mathcal{Y}$ is the Dirac mass $p^* = \delta_{y^*}$.*

Proof. Since the integrand is a squared norm, we have $J(p) \geq 0$ for all p. If $J(p) = 0$, then
$$\| \log(\exp(\Phi(y))^{-1} \exp(\ell_{\text{target}}))\|_{\mathfrak{g}}^2 = 0 \tag{62}$$
for p-almost every y. By the norm property this implies
$$\log(\exp(\Phi(y))^{-1} \exp(\ell_{\text{target}})) = 0 \tag{63}$$
almost surely. By Theorem C.2, this is equivalent to $\Phi(y) = \ell_{\text{target}}$ almost surely. Hence the support of any zero-cost distribution must be contained in the set
$$\{y : \Phi(y) = \ell_{\text{target}}\}.$$

Conversely, any distribution supported on this set yields zero cost, since the holonomy discrepancy vanishes identically. This proves part (1).

For part (2), if Φ is injective and there exists a unique y^* such that $\Phi(y^*) = \ell_{\text{target}}$, then the set above reduces to $\{y^*\}$. Thus any distribution p with $J(p) = 0$ must be supported on $\{y^*\}$, which implies $p = \delta_{y^*}$. This shows that $p^* = \delta_{y^*}$ is the unique minimizer of $J(p)$ over all distributions supported on $\mathcal{Y}$. □