AI/ML-ASSISTED REDESIGN OF NATIVE PROTEINS

Dirichlet Flow Matching With Applications To Dna Sequence Design

Hannes Staerk Massachusetts Institute of Technology **Chenyu Wang** MIT **Bowen Jing** MIT **Gabriele Corso** MIT **Bonnie Berger** MIT **Regina Barzilay** MIT **Tommi Jaakkola** MIT

Discrete diffusion or flow models could enable faster and more controllable sequence generation than autoregressive models. We show that naïve linear flow matching on the simplex is insufficient toward this goal since it suffers from discontinuities in the training target and further pathologies. To overcome this, we develop Dirichlet flow matching on the simplex based on mixtures of Dirichlet distributions as probability paths. In this framework, we derive a connection between the mixtures' scores and the flow's vector field that allows for classifier and classifier-free guidance. Further, we provide distilled Dirichlet flow matching, which enables one-step sequence generation with minimal performance hits, resulting in O(L) speedups compared to autoregressive models. On complex DNA sequence generation tasks, we demonstrate superior performance compared to all baselines in distributional metrics and in achieving desired design targets for generated sequences. Finally, we show that our classifier-free guidance approach improves unconditional generation and is effective for generating DNA that satisfies design targets.

We thank Andrew Campbell, Yaron Lipman, Felix Faltings, Jason Yim, Ruochi Zhang, Rachel Wu, Jason Buenrostro, Bernardo Almeida, Gokcen Eraslan, Ibrahim I. Taskiran, and Pavel Avdeyev for helpful discussions.
This work was supported by the NSF Expeditions grant (award 1918839: Collaborative Research: Understanding the World Through Code), the Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS) consortium, the Abdul Latif Jameel Clinic for Machine Learning in Health, the DTRA Discovery of Medical Countermeasures Against New and Emerging (DOMANE) Threats program, the DARPA Accelerated Molecular Discovery program, the NSF AI Institute CCF-2112665, the NSF Award 2134795, and the GIST-MIT Research Collaboration grant.