Fine-tuning BioEmu for Accurate Protein Folding Stability Prediction

Zhaoyang Li [*] [1]

Abstract

Recent advances in protein structure prediction models such as AlphaFold have largely resolved static folding problems. However, accurately profiling dynamic protein conformations to estimate folding stabilities remains a huge challenge. Biomolecular Emulator (BioEmu) is a recent generative deep learning framework that employs a Property Prediction Fine-Tuning (PPFT) algorithm that integrates extensive MEGAScale experimental datasets with molecular dynamics (MD) simulations to infer folding free energies. Despite its innovative design, preliminary fine-tuning results revealed limitations in predictive accuracy. In this work, we propose a novel approach that efficiently fine-tunes a $\mathrm{SE}(3)$ equivariant diffusion model using experimental expectation values, while preserving the majority of the pretrained parameters to maintain the integrity of the underlying diffusion process. This work may mark a significant advance in the integration of experimental data with deep generative models, paving the way for more reliable computational assessments of protein folding energy landscapes.

1. Introduction

Predicting protein stability (e.g. folding free energy changes $\Delta\Delta G$) from sequence and limited structural information is a long-standing challenge. Biomolecular Emulator (BioEmu) is a recently proposed generative diffusion model that addresses this by emulating protein conformational ensembles and integrating experimental stability data (Lewis et al., 2025). In BioEmu's original framework, a property-prediction fine-tuning (PPFT) algorithm was used to incorporate experimental stability measurements without requiring known structures. PPFT works by generating a small ensemble of structures (using a fast 8-step diffusion sampling) and comparing an observable (fraction of folded structures) to experimental values, then backpropagating the error to adjust the model. This strategy enabled BioEmu to predict stability with high accuracy, outperforming black-box sequence-based models.

However, the PPFT approach has limitations: it introduces an approximation in sampling that may perturb the pretrained distribution, and it does not explicitly account for the geometrical symmetries of protein conformations. Recent advances in $\mathrm{SE}(3)$ equivariant diffusion modeling provide a more principled framework for generative processes on the manifold of rigid-body transformations (Yim et al., 2023). Our central idea is to reinterpret fine-tuning as a constrained optimization on the manifold of protein conformations, where we impose that certain expected observables match experimental values while minimally perturbing the pretrained ensemble distribution. This project will pursue that idea in four stages, each serving as an independent milestone:

2. Prototype $\mathrm{IGSO}(3)$ Diffusion on $\mathrm{SO}(3)$

The first milestone is to prototype an isotropic Gaussian $\mathrm{SO}(3)$ ($\mathrm{IGSO}(3)$) diffusion process using a toy problem (Leach et al., 2022). We begin with a simple synthetic distribution on $\mathrm{SO}(3)$ that contains K $\mathrm{IGSO}(3)$ components with different mean rotations, variances, and mixture weights. The goal is to train a simple neural network to learn the score function (Proposition B.3) and model the distribution through reverse diffusion sampling (Algorithm 1). Results is visualized by plotting the sampled marginal distribution of the rotation angle ω (Proposition B.1).

Here we (i) formulate the forward and reverse stochastic differential equations (SDEs) on $\mathrm{SO}(3)$ (Bortoli et al., 2022; Hsu, 2002) (ii) parameterizing $\mathrm{SO}(3)$ diffusion with the axis-angle representation of rotation (Solà et al., 2021) and (iii) train a neural network $s_\theta(\mathbf{R}_t, t) \in \mathbb{R}^3$ via denoising score matching (Song et al., 2021).

Figure 1 shows the learned marginal densities of the rotation angle ω for an $\mathrm{IGSO}(3)$ mixture of identity, rotation by $\frac{\pi}{2}$ along the y-axis, and rotation by π along the z-axis. At the start of the reverse process ($t = 1.00$), the forward-perturbed distribution is essentially the uniform

[*] Equal contribution [1] Department of Bioengineering, Stanford University, CA 94305, USA. Correspondence to: Zhaoyang Li <zhaoyangli@stanford.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

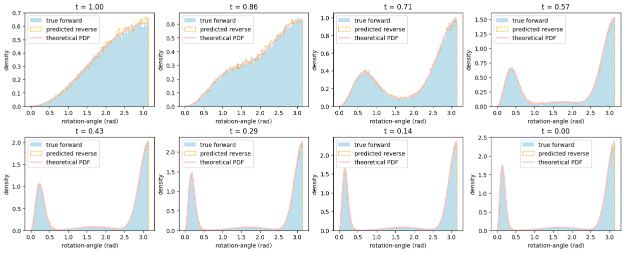

Figure 1. Learned marginal densities of the rotation angle ω for a three-component $\mathrm{IGSO}(3)$ mixture, evaluated at eight evenly spaced noise levels $t \in \{1.00, 0.86, \ldots, 0.00\}$. The orange dashed lines are the predicted reverse histograms, while the blue shading is the true forward samples. The solid pink line is the theoretical PDF given by Proposition B.1.

Haar measure on $\mathrm{SO}(3)$, and the network's predicted reverse histogram (orange dashed) overlaps almost perfectly with the true forward samples (blue shading).

As t decreases, the marginals collapse onto the two sharp peaks at $\omega \approx 0$ and $\omega \approx \pi$, as well as a third broad peak at $\omega \approx \frac{\pi}{2}$, which matches the original three-component mixture up to sampling noise. Across every time slice, the predicted reverse densities lie almost exactly on top of the theoretical PDF (solid red), which builds confidence in applying similar $\mathrm{SE}(3)$ equivariant diffusion principles to protein models.

3. Fine-tuning BioEmu Demonstration on a Single Protein

Next, we apply the new fine-tuning method (Proposition C.3) on the $\mathrm{IGSO}(3)$ mixture case. We use the same $\mathrm{IGSO}(3)$ mixture model as above, but add a constraint that the expected value of mixture weights $\mathbb{E}_{\mathbf{R}_0 \sim \mathrm{IGSO}(3)}[h_i(\mathbf{R}_0)]$ matches some slightly different values h_i^*, where $h_i(\mathbf{R}_0)$ is the i-th observable of the mixture model (Proposition C.2).

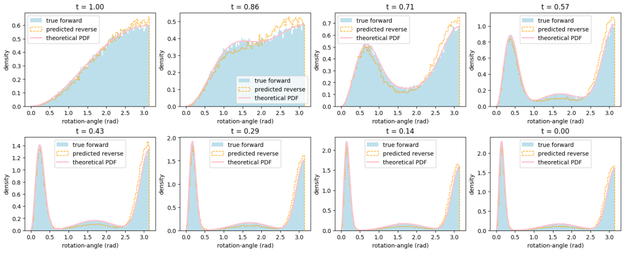

Figure 2. Fine-tuned marginal densities of the rotation angle ω for the $\mathrm{IGSO}(3)$ mixture. Legends remains the same as Figure 1.

Figure 2 shows the same set of marginal rotation-angle densities after fine-tuning the score network to satisfy the new mixture-weight constraints h_i^*. As we anneal down, The predicted reverse histogram overlaps the true forward samples better with their relative heights shifted in accordance with the target weights h_i^*, confirming that we realize the new mixture composition without retraining from scratch.

We will apply this approach on BioEmu for a single protein sequence to demonstrate a proof-of-concept in a real-world scenario. Prior to this, we will need to fine-tune the toy model of two $\mathrm{IGSO}(3)$ distributions to adjust their mixture weights.

Then we will select an example protein with an extreme stability phenotype. One candidate is an IDP from the CALVADOS dataset used in the BioEmu paper. Using this protein, our method will enforce this via the constrained optimization above. After fine-tuning on this single sequence, we expect that an IDP's generated conformations will be mostly unfolded, matching experimental observations.

4. Scaling to MEGAScale Dataset and MD Data

Having validated the approach on a single protein, we will extend it to a large-scale fine-tuning using the MEGAScale dataset of protein folding stabilities, which was also used in training the original BioEmu. For each protein or mutant in the training set, the model will generate an ensemble and compute an expected stability-related quantity. We will then compute the error between these model predictions and the experimental values, and update the model parameters to reduce this error. In addition to the experimental data, we will also attempt to incorporate the molecular dynamics (MD) simulation dataset to provide direct structural physics signals.

Milestone outcome: a fine-tuned version of the BioEmu model that integrates experimental stability data via our $\mathrm{SE}(3)$ equivariant constrained fine-tuning method. This model should have hopefully improved accuracy (in terms of predicted vs experimental $\Delta\Delta G$) while retaining physically plausible conformational sampling.

5. Benchmarking and Evaluation

The final stage focuses on rigorous evaluation of the fine-tuned model against benchmarks, and comparison to the original PPFT-based BioEmu. We will use the same evaluation protocols and datasets as the original BioEmu study to ensure a direct comparison. First, on the held-out stability dataset, we will assess the predictive $\Delta\Delta G$ accuracy of our model. Next, we will evaluate if our model has preserved the pretrained distribution aside from the intended shifts in stability-related aspects. For example, we will apply our fine-tuned model to sample ensembles for proteins with known conformational changes or binding events to verify that it still generates diverse, biologically relevant conformations. Finally, we will compare our model's predictions to other computational stability predictors such as single-point mutations with known experimental $\Delta\Delta G$ (ProTherm or the SKEMPI database) and see how well our model's predicted stability change correlates with experiments.

Milestone outcome: a comprehensive benchmark report. We expect to show that our SE(3) equivariant fine-tuning method achieves at least comparable accuracy to PPFT on stability prediction, and we will highlight any improvements.

6. Related Work

7. Experiment

8. Discussion

Acknowledgements

References

Bortoli, V. D., Mathieu, E., Hutchinson, M. J., Thornton, J., Teh, Y. W., and Doucet, A. Riemannian score-based generative modelling. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=oDRQGo8I7P.

Hsu, E. P. *Stochastic Analysis on Manifolds*. Number 38. American Mathematical Soc., 2002.

Leach, A., Schmon, S. M., Degiacomi, M. T., and Willcocks, C. G. Denoising diffusion probabilistic models on SO(3) for rotational alignment. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022. URL https://openreview.net/forum?id=BY88eBbkpe5.

Lewis, S., Hempel, T., Jiménez-Luna, J., Gastegger, M., Xie, Y., Foong, A. Y. K., Satorras, V. G., Abdin, O., Veeling, B. S., Zaporozhets, I., Chen, Y., Yang, S., Schneuing, A., Nigam, J., Barbero, F., Stimper, V., Campbell, A., Yim, J., Lienen, M., Shi, Y., Zheng, S., Schulz, H., Munir, U., Tomioka, R., Clementi, C., and Noé, F. Scalable emulation of protein equilibrium ensembles with generative deep learning. *bioRxiv*, 2025. doi: 10.1101/2024.12.05.626885. URL https://www.biorxiv.org/content/early/2025/02/25/2024.12.05.626885.

Solà, J., Deray, J., and Atchuthan, D. A micro lie theory for state estimation in robotics, 2021. URL https://arxiv.org/abs/1812.01537.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yim, J., Trippe, B. L., De Bortoli, V., Mathieu, E., Doucet, A., Barzilay, R., and Jaakkola, T. Se(3) diffusion model with application to protein backbone generation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

A. Preliminaries and Notation

Throughout this paper we adopt the following notation and conventions.

Manifolds and Lie groups. Let $\mathcal{M}$ denote a smooth, d-dimensional Riemannian manifold with metric $\langle\cdot,\cdot\rangle_{\mathcal{M}}$ and associated volume form $\mathrm{d}V$. $\mathbf{X}\in\mathcal{M}$ is a point on the manifold. We write $\mathrm{SO}(3)$ for the group of 3×3 rotation matrices and $\mathfrak{so}(3)$ for its Lie algebra, and similarly $\mathrm{SE}(3)\cong\mathrm{SO}(3)\ltimes\mathbb{R}^3$ with Lie algebra $\mathfrak{se}(3)=\mathfrak{so}(3)\oplus\mathbb{R}^3$.

For any Lie group G and its Lie algebra $\mathfrak{g}$, $\exp:\mathfrak{g}\to G$ is the Riemannian exponential map, and $\log:G\to\mathfrak{g}$ its (local) inverse. The isomorphism hat $[\cdot]^{\wedge}:\mathbb{R}^d\to\mathfrak{g}$ and vee $[\cdot]^{\vee}:\mathfrak{g}\to\mathbb{R}^d$, i.e. the vectorization and de-vectorization maps, induce $\mathrm{Exp}:\mathbb{R}^d\to G$ and $\mathrm{Log}:G\to\mathbb{R}^d$ by the composition of mappings, respectively.

The left action of $g\in G$ on $h\in G$ is $L_g(h)=gh$ and its differential is $\mathrm{d}L_g:T_hG\to T_{gh}G$. The metric on $\mathrm{SE}(3)$ is given by the canonical left-invariant metric, which is induced by the standard inner product on $\mathfrak{so}(3)$ and $\mathbb{R}^3$, i.e. $\langle\mathbf{t}_1,\mathbf{t}_2\rangle_{\mathrm{SE}(3)}=\langle\mathbf{x}_1,\mathbf{x}_2\rangle_{\mathbb{R}^3}+\langle\mathbf{r}_1,\mathbf{r}_2\rangle_{\mathrm{SO}(3)}$ for $\mathbf{t}_1=(\mathbf{r}_1,\mathbf{x}_1)\in\mathfrak{se}(3)$ and $\mathbf{t}_2=(\mathbf{r}_2,\mathbf{x}_2)\in\mathfrak{se}(3)$. The bi-invariant Riemannian metric on $\mathrm{SO}(3)$ is given by its Killing form $\langle\mathbf{r}_1,\mathbf{r}_2\rangle_{\mathrm{SO}(3)}=\frac{1}{2}\mathrm{tr}(\mathbf{r}_1^{\mathrm{T}}\mathbf{r}_2)$, where $\mathbf{r}_1,\mathbf{r}_2\in\mathfrak{so}(3)$.

Protein backbone frames. A protein backbone of N residues is represented by a sequence of rigid frames

$$\mathbf{T}=(\mathbf{T}_1,\ldots,\mathbf{T}_N)\in\mathrm{SE}(3)^N,\quad \mathbf{T}_n=(\mathbf{R}_n,\mathbf{x}_n)$$

where $\mathbf{R}_n\in\mathrm{SO}(3)$ is the rotation matrix and $\mathbf{x}_n\in\mathbb{R}^3$ is the translation vector for residue n. Each frame acts on the idealized residue coordinates $(\mathrm{N}^*,\mathrm{C}^*,\mathrm{C}_\alpha^*)\subset\mathbb{R}^3$ via $\mathbf{T}_n(v)=\mathbf{R}_nv+\mathbf{x}_n$, so that the atomic positions for residue n are

$$(\mathrm{N}_n,\mathrm{C}_n,\mathrm{C}_{\alpha n})=\mathbf{T}_n(\mathrm{N}^*,\mathrm{C}^*,\mathrm{C}_\alpha^*),$$

and the O atom is placed by an additional torsion angle ψ_n around the $\mathrm{C}_\alpha-\mathrm{C}$ bond.

Distributions on Lie groups. We denote the isotropic Gaussian distribution on $\mathrm{SO}(3)$ as $\mathrm{IGSO}(3)(\mathbf{R}_0,\sigma^2)$, where $\mathbf{R}_0$ is the mean rotation and σ^2 is the variance. The probability density function (PDF) of $\mathbf{R}_t\sim\mathrm{IGSO}(3)(\mathbf{R}_0,\sigma^2)$ when $t=\sigma^2$ is given by

$$p(\mathbf{R}_t;\mathbf{R}_0,\sigma)=\frac{1}{8\pi^2}\sum_{\ell=0}^{\infty}(2\ell+1)e^{-\frac{\sigma^2}{2}\ell(\ell+1)}\chi_\ell(\mathbf{R}_0^{\mathrm{T}}\mathbf{R}_t)\tag{1}$$

with respect to the canonical Haar measure on $\mathrm{SO}(3)$, $\mu_{\mathrm{SO}(3)}=4\sin^2\frac{\omega}{2}\mathrm{d}\omega\wedge\mathrm{d}\mathit{\Omega}$. Here χ_ℓ is the ℓ-th irreducible unitary representation of dimension $2\ell+1$ and $\mathit{\Omega}$ is the solid angle on $\mathbb{S}^2$. The axis-angle representation $\mathbf{q}=\mathrm{Log}(\mathbf{R})$ and $\omega=\|\mathbf{q}\|_2$ is used to describe the rotation for score matching. A random variable $\mathbf{R}_t\sim\mathrm{IGSO}(3)(\mathbf{R}_0,\sigma^2)$ is sampled from $\mathbf{R}_0\,\mathrm{IGSO}(3)(\mathbf{I},\sigma^2)$, where $\mathbf{I}$ is the identity matrix.

Diffusion on manifolds. Let $\mathbf{X}_t$ be a diffusion process on manifold $\mathcal{M}$ with drift $b(\mathbf{X}_t,t)$ and diffusion $g(t)$. We consider time $t\in[0,1]$ and a forward and reverse Itô SDE on $\mathcal{M}$:

$$\mathrm{d}\mathbf{X}_t=b(\mathbf{X}_t,t)\mathrm{d}t+g(t)\mathrm{d}\mathbf{W}_t^{\mathcal{M}}\tag{2}$$

$$\mathrm{d}\mathbf{X}_t=\big(b(\mathbf{X}_t,t)-g(t)^2\nabla_{\mathbf{X}_t}\log p_t(\mathbf{X}_t)\big)\mathrm{d}t+g(t)\mathrm{d}\mathbf{W}_t^{\mathcal{M}}\tag{3}$$

where $\mathbf{W}_t^{\mathcal{M}}$ is Brownian motion on $\mathcal{M}$. This gives the same form as the standard Stratonovich SDE typically used on manifolds, as the diffusion term does not depend on the state $\mathbf{X}_t$. The time-reversed process requires the Stein score $\nabla_{\mathbf{X}_t}\log p_t(\mathbf{X}_t)$, which is the Riemannian gradient of the log-density.

Fine-tuning diffusion models. Suppose we introduce an additional drift term $u(\tilde{\mathbf{X}}_t,t)$ into the original reverse diffusion process:

$$\begin{aligned}\mathrm{d}\tilde{\mathbf{X}}_t=&\Big(b(\tilde{\mathbf{X}}_t,t)-g(t)^2\nabla_{\tilde{\mathbf{X}}_t}\log p_t(\mathbf{X}_t)\Big)\mathrm{d}t\\&+g(t)u(\tilde{\mathbf{X}}_t,t)\mathrm{d}t+g(t)\mathrm{d}\mathbf{W}_t^{\mathcal{M}}\end{aligned}\tag{4}$$

and solve the following constrained optimization problem:

$$\begin{aligned}&\arg\min_{u} D_{\mathrm{KL}}(\tilde{\mathbf{X}}_0 \parallel \mathbf{X}_0)\\ \text{s.t.}\quad &\mathbb{E}_{\tilde{\mathbf{X}}_0}[h_i(\tilde{\mathbf{X}}_0)] = h_i^*, \quad i = 1, \dots, N\,.\end{aligned} \tag{5}$$

We choose to minimize the Kullback-Leibler (KL) divergence between the original and perturbed distributions:

$$\begin{aligned}D_{\mathrm{KL}}(\tilde{\mathbf{X}}_0 \parallel \mathbf{X}_0) &\leq D_{\mathrm{KL}}(\tilde{\mathbb{P}} \parallel \mathbb{P})\\ &= \frac{1}{2}\mathbb{E}_{\tilde{\mathbb{P}}}\left[\int_0^1 \left\| u(\tilde{\mathbf{X}}_t, t)\right\|_{\mathcal{M}}^2 \mathrm{d}t\right]\end{aligned} \tag{6}$$

hence the fine-tuning loss of the diffusion model is given by

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{X} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\mathrm{sg}(\theta)}}\left[\hat{L}_\theta^{\mathrm{EV}}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(M)}) + \lambda \hat{L}_\theta^{\mathrm{KL}}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(M)})\right] \tag{7}$$

where $\hat{L}_\theta^{\mathrm{EV}}$ is the empirical variance loss and $\hat{L}_\theta^{\mathrm{KL}}$ is the empirical KL divergence loss. We have also applied the leave-one-out estimator for the empirical KL divergence loss.

B. Brownian Motion on Lie Groups

B.1. Marginal distribution and score function of $\mathrm{IGSO}(3)$

Proposition B.1 (Marginal Distribution of $\mathrm{IGSO}(3)$)**.** *Let* $\mathbf{R}_t \sim \mathrm{IGSO}(3)(\mathbf{R}_0, \sigma^2)$ *be a random rotation matrix. The marginal distribution of its rotation angle* $\omega_t = \|\mathrm{Log}(\mathbf{R}_t)\|_2$ *is given by its PDF* $\frac{1-\cos\omega_t}{\pi} f(\omega_t; \mathbf{R}_0, \sigma)$*, where* $f(\omega_t; \mathbf{R}_0, \sigma)$ *is defined as*

$$f(\omega_t; \mathbf{R}_0, \sigma) = \sum_{\ell=0}^{\infty} e^{-\frac{\sigma^2}{2}\ell(\ell+1)} \frac{\sin\left(\ell + \frac{1}{2}\right)\omega_0}{\sin\frac{\omega_0}{2}} \frac{\sin\left(\ell + \frac{1}{2}\right)\omega_t}{\sin\frac{\omega_t}{2}} \tag{8}$$

Here $\omega_0 = \|\mathrm{Log}(\mathbf{R}_0)\|_2$ *is the rotation angle of* $\mathbf{R}_0$*.*

Proof. The proof follows from the fact that the marginal distribution of ω_t is given by integrating the joint density against the Haar measure $\mathrm{d}\mu_{\mathrm{SO}(3)} = 4\sin^2\frac{\omega}{2}\mathrm{d}\omega \wedge \mathrm{d}\Omega$, constrained to rotations of fixed angle on $\mathbb{S}^2$:

$$\begin{aligned}\frac{1-\cos\omega_t}{\pi} f(\omega_t; \mathbf{R}_0, \sigma) &= 4\sin^2\frac{\omega_t}{2}\int_{\mathbb{S}^2} p(\mathbf{R}_t; \mathbf{R}_0, \sigma)\mathrm{d}\Omega\\ &= \frac{1-\cos\omega_t}{4\pi^2}\sum_{\ell=0}^{\infty}(2\ell+1)e^{-\frac{\sigma^2}{2}\ell(\ell+1)}\int_{\mathbb{S}^2}\chi_\ell(\mathbf{R}_0^{\mathrm{T}}\mathbf{R}_t)\mathrm{d}\Omega\end{aligned} \tag{9}$$

where $\chi_\ell(\mathbf{R}_0^{\mathrm{T}}\mathbf{R}_t)$ denotes the ℓ-th irreducible unitary representation of $2\ell+1$ dimension. Writing the character in terms of Wigner D-matrices:

$$\begin{aligned}\chi_\ell(\mathbf{R}_0^{\mathrm{T}}\mathbf{R}_t) &= \sum_{m=-\ell}^{\ell} D_{mm}^{(\ell)}(\mathbf{R}_0^{\mathrm{T}}\mathbf{R}_t)\\ &= \sum_{m=-\ell}^{\ell}\sum_{n=-\ell}^{\ell} D_{mn}^{(\ell)}(\mathbf{R}_0) D_{nm}^{(\ell)}(\mathbf{R}_t)\end{aligned} \tag{10}$$

Now consider integrating $D^{(\ell)}(\mathbf{R}_t)$ on $\mathbb{S}^2$. For any $\mathbf{R} \in \mathrm{SO}(3)$, the integral over the class of rotations sharing a given rotation angle is invariant under conjugation, which implies

$$\begin{aligned}\int_{\mathbb{S}^2} D^{(\ell)}(\mathbf{R}_t)\mathrm{d}\Omega &= \int_{\mathbb{S}^2} D^{(\ell)}(\mathbf{R}\mathbf{R}_t\mathbf{R}^{-1})\mathrm{d}\Omega\\ &= D^{(\ell)}(\mathbf{R})\left(\int_{\mathbb{S}^2} D^{(\ell)}(\mathbf{R}_t)\mathrm{d}\Omega\right)D^{(\ell)}(\mathbf{R})^{-1}\end{aligned} \tag{11}$$

According to Schur's lemma, this integral must be proportional to the identity matrix, so we can write

$$\begin{aligned}\int_{\mathbb{S}^2} D^{(\ell)}(\mathbf{R}_t)\mathrm{d}\varOmega &= \frac{1}{2\ell+1}\operatorname{tr}\left(\int_{\mathbb{S}^2} D^{(\ell)}(\mathbf{R}_t)\mathrm{d}\varOmega\right)\mathbf{I} \\ &= \frac{4\pi}{2\ell+1}\chi_\ell(\mathbf{R}_t)\mathbf{I}\end{aligned} \tag{12}$$

where $\mathbf{I}$ is the identity matrix. Thus, we can express the integral of $\chi_\ell(\mathbf{R}_0^{\mathrm{T}}\mathbf{R}_t)$ as

$$\begin{aligned}\int_{\mathbb{S}^2} \chi_\ell(\mathbf{R}_0^{\mathrm{T}}\mathbf{R}_t)\mathrm{d}\varOmega &= \sum_{m=-\ell}^{\ell}\sum_{n=-\ell}^{\ell} D_{mn}^{(\ell)}(\mathbf{R}_0)\frac{4\pi}{2\ell+1}\chi_\ell(\mathbf{R}_t)\delta_{mn} \\ &= \frac{4\pi}{2\ell+1}\chi_\ell(\mathbf{R}_t)\sum_{m=-\ell}^{\ell} D_{mm}^{(\ell)}(\mathbf{R}_0) \\ &= \frac{4\pi}{2\ell+1}\chi_\ell(\mathbf{R}_0)\chi_\ell(\mathbf{R}_t)\end{aligned} \tag{13}$$

Substituting this back into the marginal distribution finally gives

$$\begin{aligned}f(\omega_t;\mathbf{R}_0,\sigma) &= \frac{1}{4\pi}\sum_{\ell=0}^{\infty}(2\ell+1)e^{-\frac{\sigma^2}{2}\ell(\ell+1)}\int_{\mathbb{S}^2}\chi_\ell(\mathbf{R}_0^{\mathrm{T}}\mathbf{R}_t)\mathrm{d}\varOmega \\ &= \sum_{\ell=0}^{\infty} e^{-\frac{\sigma^2}{2}\ell(\ell+1)}\frac{\sin\left(\ell+\frac{1}{2}\right)\omega_0}{\sin\frac{\omega_0}{2}}\frac{\sin\left(\ell+\frac{1}{2}\right)\omega_t}{\sin\frac{\omega_t}{2}}\end{aligned} \tag{14}$$

□

Proposition B.2 (Axis-Angle Decomposition of IGSO(3))**.** *Let* $\mathbf{R} = \mathbf{R}_0^{\mathrm{T}}\mathbf{R}_t$ *be a random rotation matrix sampled from* $\mathrm{IGSO}(3)(\mathbf{I},\sigma^2)$ *(so that* $\mathbf{R}_t \sim \mathrm{IGSO}(3)(\mathbf{R}_0,\sigma^2)$*). Let* $\mathbf{q} = \mathrm{Log}(\mathbf{R})$ *and* $\omega = \|\mathbf{q}\|_2$ *be the axis-angle representation. Then the axis-angle decomposition of* $\mathbf{R}$ *is given by*

$$\frac{\mathbf{q}}{\omega} \sim \mathcal{U}(\mathbb{S}^2) \tag{15}$$

$$\omega \sim \frac{1-\cos\omega}{\pi} f(\omega;\mathbf{I},\sigma) \tag{16}$$

where $\mathcal{U}(\mathbb{S}^2)$ *is the uniform distribution on the unit sphere* $\mathbb{S}^2$*.*

Proposition B.3 (Score Function on $\mathrm{SO}(3)$)**.** *Let* $\mathbf{R}_0, \mathbf{R}_t \in \mathrm{SO}(3)$ *and write their relative rotation as* $\mathbf{R} = \mathbf{R}_0^{\mathrm{T}}\mathbf{R}_t$ *with axis-angle representation* $\mathbf{q}$ *and* ω*. Then the score function* $s^*(\mathbf{q},t) \in \mathbb{R}^3$ *at time* $t = \sigma^2$ *of the reverse diffusion process satisfies*

$$\begin{aligned}s^*(\mathbf{q},t) &= \left[\mathbf{R}_t^{\mathrm{T}}\nabla_{\mathbf{R}_t}\log p_t(\mathbf{R}_t \mid \mathbf{R}_0)\right]^{\vee} \\ &= \frac{\mathbf{q}}{\omega}\frac{\partial}{\partial\omega}\log f(\omega;\mathbf{I},\sigma)\end{aligned} \tag{17}$$

where $p_t(\mathbf{R}_t \mid \mathbf{R}_0)$ *is the conditional distribution of* $\mathbf{R}_t$ *given* $\mathbf{R}_0$*.*

B.2. Training and sampling SDEs on $\mathrm{SO}(3)$

We aim to train a neural network to approximate the score function $s_\theta(\mathbf{R}_t, t)$ of the diffusion process on $\mathrm{SO}(3)$. The training objective is to minimize the denoising score matching loss:

$$\arg\min_{\theta}\mathbb{E}_{\mathbf{R}_0,\mathbf{R}_t|\mathbf{R}_0,t}\left[\|\lambda(t)s_\theta(\mathbf{R}_t,t) - \lambda(t)s^*(\mathbf{q},t)\|_2^2\right] \tag{18}$$

where $\lambda(t)$ is a time-dependent weighting function defined as the inverse of the standard deviation of the score function:

$$\frac{1}{\lambda(t)} = \sqrt{\frac{1}{3}\,\mathbb{E}_{\mathbf{R}\sim\mathrm{IGSO(3)}(\mathbf{I},\sigma)}\left[\|s^*(\mathbf{q},t)\|_2^2\right]} \tag{19}$$

The reverse sampling process is performed using the Euler-Maruyama method on $\mathrm{SO}(3)$. The algorithm is as follows:

Algorithm 1 Euler-Maruyama Predictor on $\mathrm{SO}(3)$

```
Require: SDE on $\mathrm{SO}(3)$ $\texttt{SO3SDE}$, score network $\texttt{ScoreNet}$, number of steps $N_{\text{steps}}$, noise weight $\lambda(t)$
Ensure: Sample $\mathbf{R}_0$
$\{t_i\}_{i=0}^{N_{\text{steps}}} \leftarrow \mathrm{linspace}(1, 0, N_{\text{steps}} + 1)$
$\Delta t \leftarrow \frac{1}{N_{\text{steps}}}$
$\mathbf{R}_1 \sim \mathcal{U}(\mathrm{SO}(3))$
$\mathbf{R} \leftarrow \mathbf{R}_1$
for $i = 0$ to $N_{\text{steps}} - 1$ do
    $t \leftarrow t_i$
    $s_\theta(\mathbf{R}, t) \leftarrow \texttt{ScoreNet}(\mathbf{R}, t) \cdot \frac{1}{\lambda(t)}$
    $(b(\mathbf{R}, t), g(t)) \leftarrow \texttt{SO3SDE}(\mathbf{R}, t)$
    $b(\mathbf{R}, t) \leftarrow b(\mathbf{R}, t) - g(t)^2 s_\theta(\mathbf{R}, t)$
    $z \sim \mathcal{N}(0, \mathbf{I})$
    $\mathbf{R} \leftarrow \mathbf{R}\,\mathrm{Exp}\,(b(\mathbf{R}, t)\Delta t)$
    $\mathbf{R} \leftarrow \mathbf{R}\,\mathrm{Exp}\left(g(t) z \sqrt{\Delta t}\right)$
end for
$\mathbf{R}_0 \leftarrow \mathbf{R}$
return $\mathbf{R}_0$
```

C. Fine-tuning diffusion models on Riemannian manifolds

C.1. Unbiased estimators for loss and gradients

Lemma C.1 (Measure Transformation)**.** *Let $h^{(1)}, h^{(2)}, \ldots, h^{(M)} \in \mathbb{R}$ be a set of independent samples from the distribution $\mathbb{Q}_\theta$. Consider another distribution $\mathbb{Q}'_\theta$ which has a Radon-Nikodym derivative $w_\theta(h^{(i)}) = \frac{\mathrm{d}\,\mathbb{Q}_\theta}{\mathrm{d}\,\mathbb{Q}'_\theta}(h^{(i)})$. Then a loss function estimator $\hat{l}\big(h^{(1)}, h^{(2)}, \ldots, h^{(M)}\big)$ with continuous second-order partial derivatives satisfies*

$$\mathbb{E}_{h \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_\theta}\left[\hat{l}\Big(h^{(1)}, h^{(2)}, \ldots, h^{(M)}\Big)\right] = \mathbb{E}_{h \overset{\text{i.i.d.}}{\sim} \mathbb{Q}'_\theta}\left[\hat{l}\Big(w_\theta(h^{(1)})h^{(1)}, w_\theta(h^{(2)})h^{(2)}, \ldots, w_\theta(h^{(M)})h^{(M)}\Big)\right] \tag{20}$$

if for all $i = 1, \ldots, M$,

$$\frac{\partial^2 \hat{l}}{\partial {h^{(i)}}^2} = 0 \tag{21}$$

Proof. Assume $\frac{\partial^2 \hat{l}}{\partial {h^{(i)}}^2} = 0$. The loss function should be able to be written as

$$\hat{l}\Big(h^{(1)}, h^{(2)}, \ldots, h^{(M)}\Big) = C_1^{(i)}\Big(h^{(1)}, \ldots, h^{(i-1)}, h^{(i+1)}, \ldots, h^{(M)}\Big)h^{(i)} + C_2^{(i)}\Big(h^{(1)}, \ldots, h^{(i-1)}, h^{(i+1)}, \ldots, h^{(M)}\Big) \tag{22}$$

where $C_1^{(i)}$ and $C_2^{(i)}$ are functions independent of $h^{(i)}$. Therefore

$$\begin{aligned}
\mathbb{E}_{h \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_\theta}\left[\hat{l}\Big(h^{(1)}, h^{(2)}, \ldots, h^{(M)}\Big)\right] &= \mathbb{E}_{h \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_\theta}\left[C_1^{(i)} h^{(i)} + C_2^{(i)}\right] \\
&= \mathbb{E}_{h \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_\theta}\left[C_1^{(i)}\right] \mathbb{E}_{h_i \sim \mathbb{Q}'_\theta}\left[w_\theta(h^{(i)})h^{(i)}\right] + \mathbb{E}_{h \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_\theta}\left[C_2^{(i)}\right] \\
&= \mathbb{E}_{h_{j\neq i} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_\theta, h^{(i)} \sim \mathbb{Q}'_\theta}\left[C_1^{(i)} w_\theta(h^{(i)})h^{(i)} + C_2^{(i)}\right] \\
&= \mathbb{E}_{h_{j\neq i} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_\theta, h^{(i)} \sim \mathbb{Q}'_\theta}\left[\hat{l}\Big(h^{(1)}, \ldots, h^{(i-1)}, w_\theta(h^{(i)})h^{(i)}, h^{(i+1)}, \ldots, h^{(M)}\Big)\right]
\end{aligned} \tag{23}$$

Using the same argument for every $i = 1, \ldots, M$, we can show that Eq. (20) holds. □

Proposition C.2 (Unbiasedness of the Gradient Estimator)**.** *Let $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)}$ be a set of independent samples from the distribution $\mathbb{P}_\theta$. Suppose*

$$\hat{l}\Big(h^{(1)}, h^{(2)}, \ldots, h^{(M)}\Big) = \big(\overline{h} - h^*\big)^2 - \frac{1}{M(M-1)} \sum_{j=1}^{M} \Big(h^{(j)} - \overline{h}\Big)^2 \tag{24}$$

where $\overline{h} = \frac{1}{M}\sum_{j=1}^{M} h^{(j)}$*. Then the estimator for the loss function* $\mathcal{L}(\theta) = \left(\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\theta}[h(\mathbf{X})] - h^*\right)^2$ *is given by*

$$\hat{L}_\theta(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)}) = \hat{l}\Big(w_\theta(\mathbf{X}^{(1)})h(\mathbf{X}^{(1)}), w_\theta(\mathbf{X}^{(2)})h(\mathbf{X}^{(2)}), \ldots, w_\theta(\mathbf{X}^{(M)})h(\mathbf{X}^{(M)})\Big) \tag{25}$$

which satisfies both

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{X}\overset{\text{i.i.d.}}{\sim}\mathbb{P}_{\text{sg}(\theta)}}\left[\hat{L}_\theta(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)})\right] \tag{26}$$

$$\nabla_\theta\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{X}\overset{\text{i.i.d.}}{\sim}\mathbb{P}_{\text{sg}(\theta)}}\left[\nabla_\theta\hat{L}_\theta(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)})\right] \tag{27}$$

Here $w_\theta(\mathbf{X}^{(i)}) = \frac{\mathrm{d}\,\mathbb{P}_\theta}{\mathrm{d}\,\mathbb{P}_{\text{sg}(\theta)}}(\mathbf{X}^{(i)})$ *is the Radon-Nikodym derivative of the measure* $\mathbb{P}_\theta$ *with respect to* $\mathbb{P}_{\text{sg}(\theta)}$*.*

Proof. One can easily verify that $\hat{l}\big(h(\mathbf{X}^{(1)}), h(\mathbf{X}^{(2)}), \ldots, h(\mathbf{X}^{(M)})\big)$ is the unbiased estimator of the loss function $\mathcal{L}(\theta)$. Recalling $w_\theta(\mathbf{X}^{(i)}) = 1$ when no gradient is applied, we provide the proof for Eq. (26).

Observe that for every $i = 1, \ldots, M$,

$$\frac{\partial^2\hat{l}}{\partial {h^{(i)}}^2} = \frac{2}{M^2} - \frac{2}{M(M-1)}\left(\frac{M-1}{M^2} + \left(1 - \frac{1}{M}\right)^2\right) = 0 \tag{28}$$

actually holds. Hence we may apply Lemma C.1, which yields

$$\mathbb{E}_{\mathbf{X}\overset{\text{i.i.d.}}{\sim}\mathbb{P}_\theta}\left[\hat{l}\Big(h(\mathbf{X}^{(1)}), h(\mathbf{X}^{(2)}), \ldots, h(\mathbf{X}^{(M)})\Big)\right] = \mathbb{E}_{\mathbf{X}\overset{\text{i.i.d.}}{\sim}\mathbb{P}_{\text{sg}(\theta)}}\left[\hat{L}_\theta(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)})\right] \tag{29}$$

By definition the left-hand side is exactly $\mathcal{L}(\theta)$. Thus, we obtain

$$\nabla_\theta\mathcal{L}(\theta) = \nabla_\theta\,\mathbb{E}_{\mathbf{X}\overset{\text{i.i.d.}}{\sim}\mathbb{P}_{\text{sg}(\theta)}}\left[\hat{L}_\theta(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)})\right] = \mathbb{E}_{\mathbf{X}\overset{\text{i.i.d.}}{\sim}\mathbb{P}_{\text{sg}(\theta)}}\left[\nabla_\theta\hat{L}_\theta(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)})\right] \tag{30}$$

completing the proof of Eq. (27). □

C.2. Fine-tuning loss function for diffusion models on manifolds

Proposition C.3 (Fine-tuning Diffusion Models on Manifolds)**.** *We adopt the* $\hat{l}$ *function and importance sampling weights defined in Proposition C.2. Here for the path measure* $\tilde{\mathbb{P}}_\theta$ *and* $\tilde{\mathbb{P}}_{\text{sg}(\theta)}$ *on the manifold* $\mathcal{M}$*, we have*

$$w_\theta(\mathbf{X}) = \exp\left(\int_0^1 \left\langle u_\theta(\mathbf{X}_t, t) - u_{\text{sg}(\theta)}(\mathbf{X}_t, t), \mathrm{d}\mathbf{W}_t^{\mathcal{M}}\right\rangle_{\mathcal{M}} - \frac{1}{2}\int_0^1 \left\|u_\theta(\mathbf{X}_t, t) - u_{\text{sg}(\theta)}(\mathbf{X}_t, t)\right\|_{\mathcal{M}}^2 \mathrm{d}t\right) \tag{31}$$

which has the first-order derivative

$$\nabla_\theta w_\theta(\mathbf{X}) = \int_0^1 \left\langle \nabla_\theta u_\theta(\mathbf{X}_t, t), \mathrm{d}\mathbf{W}_t^{\mathcal{M}}\right\rangle_{\mathcal{M}} \tag{32}$$

Then the empirical variance loss for different $\{h_i\}_{i=1}^N$ *is given by*

$$\hat{L}_\theta^{EV} = \sum_{i=1}^{N}\hat{l}\Big(w_\theta(\mathbf{X}_0^{(1)})h_i(\mathbf{X}_0^{(1)}), w_\theta(\mathbf{X}_0^{(2)})h_i(\mathbf{X}_0^{(2)}), \ldots, w_\theta(\mathbf{X}_0^{(M)})h_i(\mathbf{X}_0^{(M)})\Big) \tag{33}$$

and the empirical KL divergence loss is given by

$$\hat{L}_\theta^{KL} = \frac{1}{2M}\sum_{j=1}^{M}\left(w_\theta(\mathbf{X}^{(j)})\int_0^1 \left\|u_\theta(\mathbf{X}_t^{(j)}, t)\right\|_{\mathcal{M}}^2 \mathrm{d}t\right) \tag{34}$$