1141ML-Week 10 Final Project Conception

Reconstructing Ancient Life through Molecular Diffusion Models

分子擴散模型重建古生命

313652018 王宣瑋

September 2025

1 AI 的未來能力：重建古環境

未來二十年，隨著生成式模型特別是擴散模型（Diffusion Model）的演進，AI 可能具備從現存生物分子資料中「推回」過去的能力，重建古代生命的分子結構與蛋白質形態。這種技術不再只是模仿自然，而是嘗試以數據與物理法則爲基礎，生成曾經存在過、但已消逝的生命結構。故若此能力能實現，對人類與科學將具有深遠意義。

首先，它可使我們超越現有化石與古 DNA 的侷限，利用 AI 的推理能力探索「生命從何而來」這一核心問題，重現遠古酵素、細胞機制與代謝系統的可能樣貌。其次，這將爲合成生物學與藥物設計開啓新途徑，AI 所生成的古代或假想蛋白質可能擁有現代生物未具備的穩定性與功能，成爲新材料或新藥設計的重要來源。最後，這樣的突破也將迫使人類重新思考生命的定義與倫理界線——若 AI 能創造一種從未存在但可運作的生命形態，那麼「生命」是否仍僅限於自然演化的產物？

2 涉及的機器學習類型

本專題主要屬於「**非監督式學習（Unsupervised / Self-supervised Learning）**」，較精確地說是**自我監督式學習（self-supervised learning）**的一種分佈生成模型（score-based diffusion model），而非傳統的監督式學習或強化學習。

在本計畫中，模型的目標是學習三維點雲（或未來的分子／蛋白質結構）在空間中的**機率分佈**，也就是學習 $p(x)$ 以及其對數密度的梯度（score）$\nabla_x \log p_t(x)$，以便在反向 diffusion 過程中從雜訊逐步生成合理的三維結構。在這個設定下：

- 資料集中並沒有外加的「正確標籤」y，只有三維座標 x 本身；
- 損失函數爲 DSM，

$$\mathcal{L}_{\text{DSM}} = \mathbb{E}_{x,\epsilon,t}\left\| s_\theta(x_t, t) + \frac{\epsilon}{\sigma_t} \right\|^2,$$

  其中目標訊號 ϵ/σ_t 是由加噪過程機率模型推導出來，而不是人工標註的 label；
- 「資料分佈」爲檢測優劣的重點資訊。

因此這樣的機器學習被歸類爲非監督／自我監督式學習。

在 toy model 階段，資料 x 來自**合成的三維點雲**，例如兩團高斯分佈的混合：

$$x \sim 0.5\mathcal{N}(\mu_1, I) + 0.5\mathcal{N}(\mu_2, I),$$

未來可替換爲實際的分子資料集（如 QM9）或蛋白質骨架片段。對於每筆樣本 x，我們透過 forward diffusion 加上高斯雜訊得到

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

再利用 DSM 損失讓模型學習 score。此處的「目標訊號」是由噪音模型 ϵ/σ_t 所構造，屬於自我監督訊號，而非外部標註。

這個計畫中，模型只對**固定資料分佈**進行學習，沒有與動態環境互動，也沒有透過回饋訊號（reward）來更新策略；訓練全程在離線資料集上完成。因此本計畫**不屬於強化學習**，而是典型的離線生成式自我監督學習設定。

3 第一步的模型化

（1）此簡化問題如何代表最終能力？ SE(3)-equivariant 3D diffusion 是所有分子與蛋白質生成模型的核心運算基礎。只要模型能成功在三維空間中具備：

- SE(3) 等變性（旋轉、平移不影響生成結果），
- 以 score-based diffusion 生成三維形狀，
- 能從純噪音逐步重建合理的 3D 幾何結構 [2]，

則代表此學習模型已具備未來進行分子生成、蛋白質生成甚至生命分子重建所需的關鍵幾何能力。

（2）如何測試此模型是否成功？ 可以透過以下方式檢驗：

- 生成的三維點雲是否接近原始資料分佈（以 Chamfer distance 或 RMSD 衡量）；
- 模型輸入經旋轉或平移後，其輸出是否同步變換（檢查 SE(3) 等變性）；
- 反向 diffusion 是否能穩定地從雜訊生成合理形狀 [3]。

若以上條件皆滿足，即視爲 toy model 成功。

（3）需要哪些數學或工具來解決？ 此模型需要：

- 3D 幾何與群論基本概念（SO(3), SE(3)）；
- 機率模型：forward diffusion、score matching；
- SE(3)-equivariant 架構（例如 EGNN）[1]；
- Python 與 PyTorch 實作工具。

這些構成未來分子／蛋白質生成模型的基礎技術。

References

[1] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

[2] 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.

[3] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.