1141ML-Week 9 Final Project Conception

Reconstructing Ancient Life through AI-Driven 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

本研究構想主要涉及**非監督式學習**（Unsupervised Learning）與生成式模型中的**擴散模型**（Diffusion Model），並可結合**強化學習**（Reinforcement Learning）作爲後續優化機制。

- 非監督式學習能讓模型從大量現存蛋白質與分子資料中，自主學習結構分布與演化規律，而不需人工標註。
- 擴散模型則透過模擬加噪與去噪的過程，逐步生成穩定且具有生物合理性的分子結構。
- 強化學習可進一步引入能量函數或結構穩定性作爲回饋訊號，使模型在反覆試驗中學習生成具功能性的蛋白質。

在此任務中，資料來源爲**現有的蛋白質序列資料庫與分子結構數據**，目標訊號則是**生成結構的穩定度與生物功能性評估**。模型透過反覆生成與評估的過程，與模擬環境形成學習迴圈，逐步改進生成品質。此結合非監督與強化學習的方法，能讓 AI 在缺乏標籤資料的情況下，自主探索生命分子演化的可能軌跡。

在近年來，非監督式或自監督學習在蛋白質／分子設計的研究中已有實際進展。比如 2023 年的研究透過未標註蛋白質序列用深度模型識別功能族群 [3][4]；2024 年則有分子表面嵌入研究利用無標籤電子性質資料進行學習 [5]。這些例子說明選用非監督學習於本研究構想中是可行且具有前瞻性的策略。

3 第一步的模型化

此問題的設計目的是模擬整個生命重建過程中最核心的部分：**讓模型能從現有蛋白質的分佈中推測出其共同祖先分子的可能結構。**

若 AI 能夠在受限資料下重建出具有演化合理性的蛋白質，則意味著它具備理解與生成生命分子規律的雛形。模型的可測試性可藉由兩項指標判定：

1. 生成蛋白質的三維結構能否通過能量穩定度評估；

2. 其序列特徵是否與已知演化路徑中的中間型態相符。

若模型生成結果在結構穩定性與演化一致性上皆達標，即可視爲成功的初步驗證。

爲解決此問題，可運用高維機率分佈建模與變分推論方法；機器學習上，主要依賴擴散模型（Diffusion Model）進行分子生成，並輔以自監督表示學習（Self-Supervised Representation Learning）以提取序列與結構特徵 [1][2]。爲了進一步提升生成品質，可採用強化學習（Reinforcement Learning）將物理模擬或能量函數作爲獎勵信號，使模型在反覆生成與評估的過程中逐步學習到穩定且具生物意義的結構。

References

[1] Roman Bushuiev, Anton Bushuiev, Raman Samusevich, Corinna Brungs, Josef Sivic, and Tomáš Pluskal. Self-supervised learning of molecular representations from millions of tandem mass spectra using dreams. *Nature Biotechnology*, pages 1–11, 2025.

[2] Michail Chatzianastasis, Yang Zhang, George Dasoulas, and Michalis Vazirgiannis. Geometric self-supervised pretraining on 3d protein structures using subgraphs. *arXiv preprint arXiv:2406.14142*, 2024.

[3] Kyle T David and Kenneth M Halanych. Unsupervised deep learning can identify protein functional groups from unaligned sequences. *Genome Biology and Evolution*, 15(5):evad084, 05 2023.

[4] Cong Fu, Keqiang Yan, Limei Wang, Wing Yee Au, Michael Curtis McThrow, Tao Komikado, Koji Maruhashi, Kanji Uchino, Xiaoning Qian, and Shuiwang Ji. A latent diffusion model for protein structure generation. In *Learning on graphs conference*, pages 29–1. PMLR, 2024.

[5] Viet Thanh Duy Nguyen and Truong Son Hy. Multimodal pretraining for unsupervised protein representation learning. *Biology Methods and Protocols*, 9(1):bpae043, 2024.