利用 AI 结构预测工具探索焦磷酸激酶的隐藏口袋

王煜，赵安祈，王誉凯，胡锦睿，史永泰

摘要

本项目以焦磷酸激酶（Pyrophosphokinase）作为预测对象，使用 AlphaFold2 衍生工具 AF-Cluster 与基于深度生成模型的构象采样工具 BioEmu，根据其氨基酸序列大量预测其蛋白质结构的构象复合物。通过分析预测构象的聚类，我们确定了代表其基本动态状态的构象的特征，并找到了那些具有“隐藏口袋”的构象。通过将隐藏口袋“打开”和“封闭”的预测状态分别与实验晶体结构 Holo（3IP0）和 Apo（1HKa）进行比较，验证了预测构象的准确性。本项目还比较了两种预测方法的优缺点，并研究了识别构象复合物中主要状态、评估预测准确性以及判断配体结合潜力的方法。

关键词：焦磷酸激酶；AF-Cluster；BioEmu；构象系综；隐藏口袋；分子对接

1. 引言

蛋白质的功能性能取决于其三维结构。三维结构不是静态的，而是在溶液中复杂的动态变化。这种构象动力学对酶催化、信号转导和配体识别至关重要[1][2]。传统的晶体结构预测倾向于以最低的能量获取稳定的构型（如 Apo 状态下的 1HKA）。然而，当配体不存在时，一些功能或药物结合所需的 “隐藏口袋”（Cryptic Pocket）可能是封闭的或不可见的（如 Holo 状态下的 3IP0），它们只在特定的构象状态下开放[3]。这为药物开发提供了新的潜在靶点。然而，传统的实验方法，如 X 射线晶体学，很难系统地捕捉这些过渡构象。

近年来，人工智能彻底改变了结构生物学领域。Alphafold2 在预测静态结构方面取得了突破性的成功，本项目使用的 AF-Cluster 和 AF-Cluster 均为在此基础上进行开发。AF-Cluster 方法通过输入序列的多个预测并扰动来生成多样化的构象集合。深度生成模型 BioEmu 使用去噪扩散框架直接从序列中生成近似平衡分布的构象集合，为蛋白质动力学建模提供了新的有效计算方法。

本项目以焦磷酸激酶为研究对象[4]，结合 AF-Cluster 与 BioEmu 方法，系统地预测了其构象。本项目的目的是探究基于大规模预测构象来识别主要功能状态的方法；判断预测的构象，与实验观察的一致性；评估这些隐藏口袋的配体结合潜力。通过这些研究，我们展示了计算机模型揭示蛋白质动态功能位点的能力。

2. 材料与方法

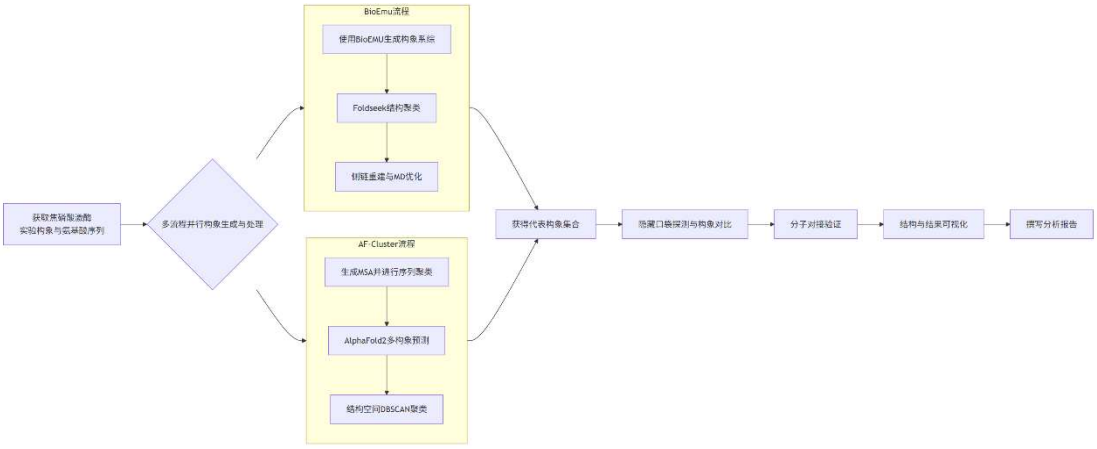

图 1 本项目的主要工作流程

图 1 展示了本项目的主要工作流程，整合了两种人工智能方法来研究具有隐藏口袋的焦磷酸激酶的动态构象。研究始于定义目标和准备数据。主链路采用并行策略：BioEmu 使用扩散发生器模型深入快速采样构象，然后通过 Foldseek 结构分组和分子动力学优化获得高质量的代表性构象；AF-Cluster 过程使 AlphaFold2 预测多种构象，并在结构空间中进行二次聚类筛选，组合序列以对齐多个序列。在合并从两条途径获得的一组代表性构象后，进行隐藏口袋的单一检测、构象状态的比对和分子结合的验证。最终，所有结果都被可视化并整合到一份完整的分析报告中。该过程结合了人工智能工具的研究思路，这些工具结合了不同的原理，以充分揭示蛋白质的构象景观。

2.1 目标序列获取

从 UniProt 数据库获取焦磷酸激酶 P26281 的完整氨基酸序列。从蛋白质结构数据库下载其 Apo 状态（1HKA）与 Holo 状态（3IP0）的晶体结构，分别作为无配体状态和配体结合状态的实验参考。

2.2 构象系综生成方法

2.2.1 AF-Cluster

AF-Cluster 是一种基于多序列比对扰动的生成方法。此方法的核心原理是通过扰动输入 AlphaFold2 的多序列比对，从而诱导其预测同一蛋白质序列的不同构象[5]。

首先，使用 ColabFold 生成目标序列的 MSA。随后，利用 AF-Cluster 脚本对 MSA 进行序列相似性聚类（DBSCAN, eps=11.00），将同源序列划分为 599 个具有不同进化特征的子集。每个子集作为一个独立的“进化背景”，被分别输入至 AlphaFold2 模型（model_3_ptm, recycles=3）中运行，最终获得 585 个初始预测构象。

此方法巧妙地利用了 AlphaFold2 框架本身，通过系统性地改变其共进化信号输入，使其能够采样到进化上允许的、不同于单一主态的其他稳定构象，同时保持了较高的预测效率。

2.2.2 BioEmu

BioEmu 是一种基于去噪扩散概率模型的深度生成模型，其目标是直接模拟蛋白质在溶液中的近似平衡构象分布[6]。

BioEmu 以目标蛋白序列为输入，配置生成参数（num_samples=165, model_name='bioemu-v1.1'）。模型利用 AlphaFold2 衍生的序列成对表示作为条件，通过一个逆向扩散过程直接生成三维坐标，快速产生包含 158 个构象的初始轨迹。

与依赖于现有 MSA 的方法不同，BioEmu 通过在海量分子动力学和实验数据上训练的生成模型，直接学习构象空间的概率分布，能够更高效地采样包括稀有态在内的多种构象，可能更好地捕获功能相关的构象变化。

2.3 构象后处理、聚类与代表构象选取

为从生成的大量初始构象中提取具有代表性的结构状态，我们对两种方法的结果分别进行了专门的后处理与聚类分析。

2.3.1 AF-Cluster 构象的聚类与代表选取

对 AF-Cluster 生成的构象，我们在结构空间进行二次聚类，以区分主要状态和稀有状态。使用 MDTraj 加载所有 PDB 结构，选择 Cα原子进行最优叠合，并计算两两之间的均方根偏差（RMSD），构建 585×585 的距离矩阵。采用 DBSCAN 算法（eps=2.0 Å, min_samples=2）对该矩阵进行聚类。此方法能自动确定簇的数量，并将稀疏点标记为“噪声”，有效分离出离散构象。

代表构象的选取采用双重策略。按簇大小降序排列，选取前 3 个最大的簇，计算每个簇内所有构象到其他构象的 RMSD 之和，选取总和最小的构象作为该簇的中心代表。所有被 DBSCAN 标记为“噪声”的构象以及其余小簇成员均被保留为稀有构象，作为后续寻找隐藏口袋的重点筛查对象。

2.3.2 BioEmu 构象的聚类与代表选取

对 BioEmu 生成的初始轨迹，首先，将轨迹文件提取为独立的 PDB 文件。随后，使用 Foldseek 工具进行高效结构聚类（参数：tmscore_threshold=0.6, coverage_threshold=0.7），依据三维结构相似性将构象归类。接着，对聚类得到的代表构象进行两步优化：先进行侧链重建以补全原子细节，再进行分子动力学局部能量最小化（LOCAL_MINIMIZATION），以消除空间冲突、优化侧链构象，提升模型的物理合理性。经过上述流程，最终从初始构象中提炼出 15 个经过优化、代表不同结构簇的高质量构象（命名为 frame1 至 frame15），用于后续分析。

2.4 隐藏口袋探测与分析

使用 fpocket 对所有代表构象进行口袋探测与评分。比较不同构象间口袋的存续与评分，识别构象特异性隐藏口袋。

使用 AutoDock Vina 进行分子对接验证。从 3IP0 结构中提取共结晶配体（HHS）作为对接小分子。分别以 AF-Cluster 和 BioEmu 方法筛选出的最优含口袋构象为受体，执行分子对接，评估结合亲和力（kcal/mol）与结合模式。

2.5 结构比对与验证

使用 PyMOL 的 align 命令，将预测的代表构象与相应的实验晶体结构（1HKA 或 3IP0）进行全 Cα原子叠合，计算 RMSD 值。

3. 结果

3.1 构象系综预测与聚类概览

我们分别采用 AF-Cluster 与 BioEmu 两种方法对焦磷酸激酶进行了大规模构象采样。AF-Cluster 方法通过对多序列比对进行序列聚类，引导 AlphaFold2 预测生成多个初始构象。随后，我们通过计算 Cα原子 RMSD 矩阵并进行结构空间聚类（DBSCAN, eps=2.0 Å），揭示了构象分布的高度不均一性（图 1）：识别出一个包含 314 个构象的绝对主导簇（簇 0），一个包含 12 个构象的次要簇（簇 1），以及多个小簇和 224 个离散的“噪声”构象。我们选取了前三大簇的中心构象作为主要代表：EX_548.pdb（簇 0）、EX_487.pdb（簇 1）和 EX_260.pdb（簇 5）。

BioEmu 通过深度扩散模型直接生成构象，并经由 Foldseek 结构聚类、侧链重建与分子动力学优化流程，最终获得了 15 个高质量的代表性构象（命名为 frame1 至 frame15）。

3.2 代表构象的识别与隐藏口袋的发现

使用 fpocket 对所有代表构象进行系统性的口袋探测与评分后，我们发现了具有显著构象特异性的隐藏口袋。在 AF-Cluster 的三个主要代表构象中，一个高分口袋仅存在于 EX_487.pdb（簇 1）和 EX_260.pdb（簇 5）中，而在主簇代表 EX_548.pdb 中完全缺失。其中，EX_487.pdb 的口袋评分最高，被选定为 AF-Cluster 的最优含口袋构象。

在 BioEmu 的 15 个代表构象中，同一隐藏口袋在 8 个构象（frame4, 6, 7, 8, 10, 12, 13, 14）中被检测到，进一步证实了该口袋的可及性。其中，frame8 的口袋评分最高，

被确定为 BioEmu 的最优含口袋构象。这一发现表明，该隐藏口袋并非蛋白最稳定主态的一部分，而是与特定的、可能能量稍高的构象亚态相关。

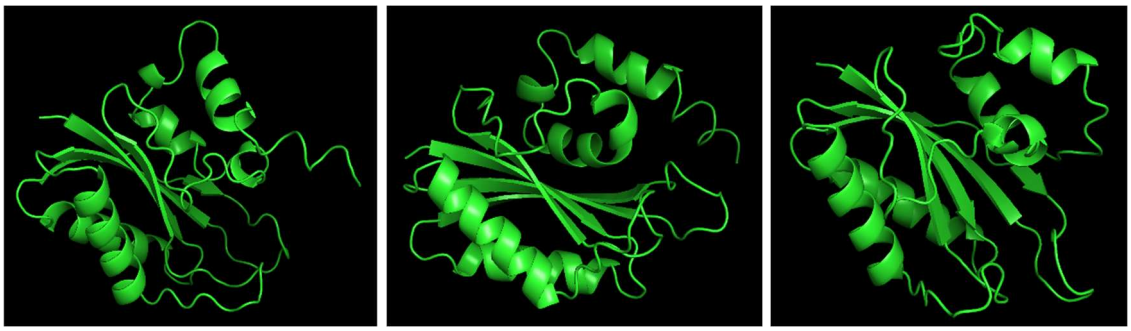

图 2 AF-Cluster 生成的代表构象

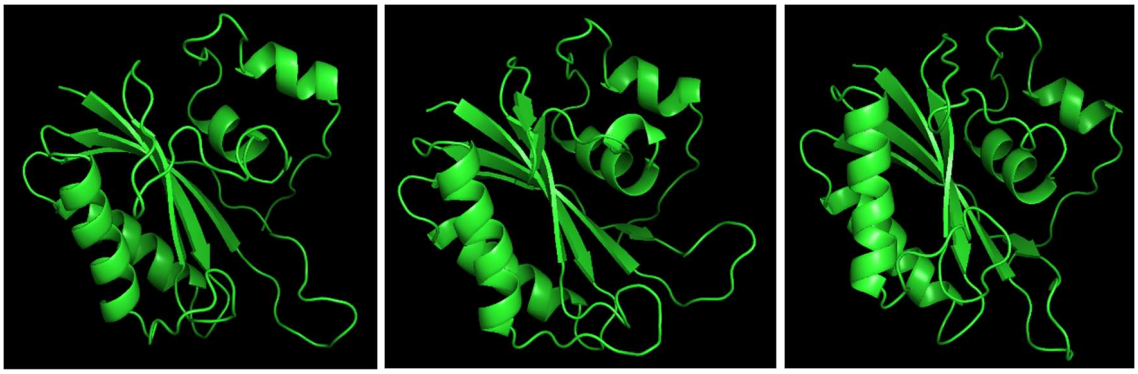

图 3 BioEmu 生成的代表构象

图 2 展示了 AF-Cluster 生成的三个代表构象，从左至右分别是 EX_487, EX_260, EX_548；图 3 展示了 BioEmu 生成的代表构象中的三个，从左至右分别是 frame8，frame14，frame5。生成式模型生成的构象彼此之间具有一定区别，这揭示了蛋白质分子的不同结构和功能状态。

3.3 分子对接验证口袋功能

为直接验证该隐藏口袋的功能潜力，我们以 3IP0 中的共结晶配体 HHS 作为探针分子，对 EX_487.pdb 和 frame8 进行了分子对接计算。

AF-Cluster (EX_487.pdb)：最优对接模式的结合亲和力为 -6.4 kcal/mol，表明该口袋具有良好的结合能力。将其与 EX_548 与该对接结果进行比对，从而验证了 EX_487 含有其余构象所不包含的隐藏口袋。

BioEmu (frame8)：最优对接模式的结合亲和力为 -6.6 kcal/mol，略优于 AF-Cluster 的结果。此外，其对接结果中高亲和力（≤-6.0 kcal/mol）的模式更多，暗示结合构象可能更稳定或更多样。

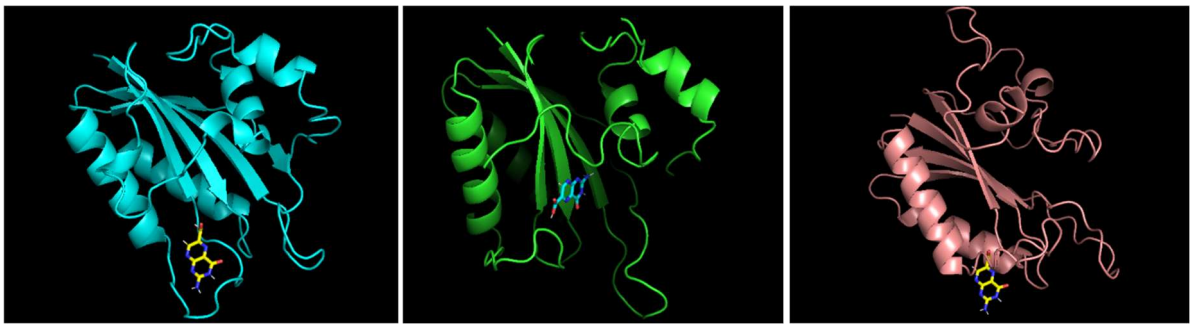

图 4 预测构象中发现的隐藏口袋

图 4 是预测构象中发现的隐藏口袋特写，从左至右分别是 EX_487，frame8，EX_548 的分子对接结果。从图中可见，EX_487，frame8 均具有显现的口袋，而 EX_548 所代表的构象不具有口袋。这进一步证明了我们成功地生成了具有隐藏口袋和不具有隐藏口袋的两种构象。

3.4 与实验晶体结构的比对验证

为评估预测构象的合理性，我们将其与已知的实验结构进行比对。AF-Cluster 的主态构象 EX_548.pdb（无口袋）与 Apo 状态结构（PDB: 1HKA） 叠合良好，Cα RMSD 为 1.2 Å，且结合区域均呈封闭状态。

更重要的是，两个最优含口袋构象与 Holo 状态结构（PDB: 3IP0） 的比对显示：AF-Cluster 的 EX_487.pdb 与 3IP0 的 RMSD 为 1.8 Å；而 BioEmu 的 frame8 与 3IP0 的 RMSD 更低，为 1.005 Å。叠合显示，frame8 中预测的隐藏口袋其形状和位置与 3IP0 中配体 HHS 的实际结合腔高度重合，提示其高度的生物学真实性。

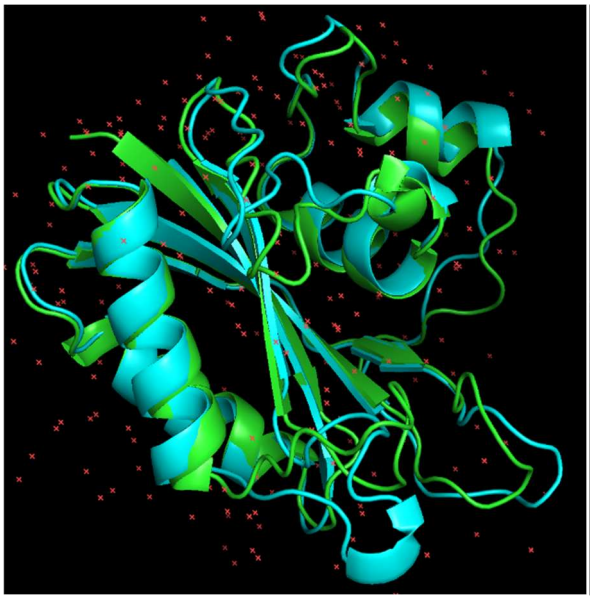

图 5 无隐藏口袋构象与 Apo 结构（1HKA）的叠合

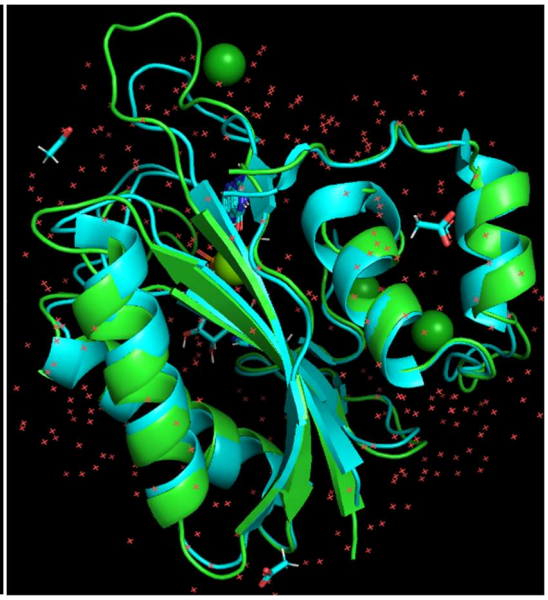

图 6 有隐藏口袋构象与 Holo 结构（3IP0）的叠合

图 5（左）以 EX_548 为例展示了无隐藏口袋构象与 Apo 结构（1HKA）的叠合，其中绿色构象是生成构象，蓝色构象是实验构象；图 6（右）以 frame8 为例展示了有隐藏口袋构象与 Holo 结构（3IP0）的叠合，其中绿色构象是生成构象，蓝色构象是实验构象。从图中可见，本项目生成的蛋白质构象与实验构象的相似度较好，展示了项目所采用方法的可靠性。

4.讨论

4.1 对关键问题的分析

1. 如何在预测出的构象系综中找出主要构象？

我们使用基于结构的不受控聚类分析。首先计算构象两两之间主链原子的 RMSD，建立距离矩阵；然后使用 DBSCAN 算法对该矩阵进行聚类。这种方法的优点是不需要预先指定簇的数量，分散的点可以自动识别为“噪声”，这可以有效地区分代表不定大小的稳定状态的高密度构象簇，以及代表稀有态的离散的构象。例如，在例如，对 AF-Cluster 的构象聚类后，我们清晰地识别出了一个包含 314 个构象的主体簇（簇 0）和包含 12 个构象的次要簇（簇 1）。

2. 如何将主要构象和晶体结构进行比对？

我们使用 PyMOL 的 align 命令进行叠合，并计算所有 Cα原子的 RMSD 值作为全局结构相似性的定量指标。这为口袋功能的相关性提供了直接证据。

3. 哪种方法效果最好？可能的原因是什么？

结合构象多样性、与实验结构的一致性和计算结合亲和力这三个指标，在本项目中，BioEmu 的整体表现优于 AF-Cluster。BioEmu 产生了更丰富的构象（15 种代表性构象），

最佳构象（frame8）与实验 Holo 结构（3IP0）的相似度更高（RMSD: 1.005 Å vs 1.8 Å），分子对接也显示其具有更优的配体结合潜力（-6.6 kcal/mol vs -6.4 kcal/mol）。

AF-Cluster 通过扰动 MSA 诱导 AlphaFold2 做出各种预测，仍然受到 AlphaFold2 预测的独特结构静态框架和进化信息的"平均效应"的限制，因此更适合预测最稳定的主要构象。而 BioEmu 是直接在大量分子动力学建模数据上训练的生成式模型，旨在研究和模拟蛋白质构象的平衡分布。因此，它可能更擅长采样到亚稳态或激发态，这些状态通常不是能量景观中的最低状态，但对于配体结合发挥作用至关重要，这正是揭开隐藏口袋的关键。

4. 针对预测出的构象，如何判断配体是否与之结合？

可以通过一系列计算，包括口袋检测和筛查，预测构象和实验构象比对，以及分子对接验证得出有利证据。使用 fpocket 等工具寻找物理化学尺寸、形状和性质合理的空腔，并进行评估筛查；如果预测的含口袋构象与 Holo 态的已知晶体结构非常相似（低 RMSD 值），并且口袋位置重合，则可以推断结论很可能成立；将已知活性的配体对接到预测口袋中。如果可以获得稳定、能量低、构象合理的结合方式，则证明这个口袋完全有能力与配体结合。

4.2 综合讨论

本项目通过整合两种先进的人工智能预测工具来识别焦磷酸激酶的一个重要隐藏口袋，并通过计算手段系统地评估了其其结构合理性与功能潜力。结果表明，仅用最弱能量分析基本态构象极有可能遗漏隐藏口袋，有必要对构象复合体的结合进行全面分析，并系统筛查所有代表性亚态。本项目采用的"大规模生成-结构聚类-全代表聚类验证"可以构成检测隐藏口袋的有效策略。

4.3 前景

本项目依赖于计算预测，所有构象均未经过实验验证，有待后续湿实验的验证。尽管与晶体结构比对结果良好，但采样可能仍未覆盖所有相关的生理构象。未来的工作可以围绕 frame8 构象的结构进行：进行更长的分子动力学建模来评估其稳定性；基于这个隐藏口袋寻找新的先导化合物；实验测试该口袋对酶活性的影响。

5.结论

本项目结合 AF-Cluster 和 BioEmu 两种人工智能策略，系统地研究了焦磷酸激酶的构象，并能够检测到具有重要功能的隐藏口袋。这种隐藏口袋具有明显的构象特异性，仅存在于特定的亚稳态。在研究中，BioEmu 方法表现出了更高的性能，其可预测的构象更接近实验配体结合态的结构，计算也显示出其更高的配体键合潜力。该项目展示了一个生物信息学研究框架，该框架整合了多种人工智能工具，进行构象预测，可用于基于动态结构发现药物。

参考文献

[1] Johnson, T. A., & Holyoak, T. (2010). Increasing the Conformational Entropy of the Ω-Loop Lid Domain in Phosphoenolpyruvate Carboxykinase Impairs Catalysis and Decreases Catalytic Fidelity. *Biochemistry*, 49(25), 5176-5187. https://doi.org/10.1021/bi100399e

[2] Haldane, A., Flynn, W. F., He, P., Vijayan, R. S. K., & Levy, R. M. (2016). Structural propensities of kinase family proteins from a Potts model of residue co-variation. *Protein Science*, 25(8), 1378-1384. https://doi.org/10.1002/pro.2954

[3] X. Luo et al. The Mad2 spindle checkpoint protein has two distinct natively folded states. 2004, Nat. Struct. Mol. Biol.
[4] Yun, M.-K., Hoagland, D., Kumar, G., Waddell, M. B., Rock, C. O., Lee, R. E., & White, S. W. (2014). The identification, analysis and structure-based development of novel inhibitors of 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase. *Bioorganic & Medicinal Chemistry*, 22(7), 2157-2165. https://doi.org/10.1016/j.bmc.2014.02.022
[5] Wayment-Steele HK, Ojoawo A, Otten R, et al. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature. 2023;625:832-839. doi:10.1038/s41586-023-06832-9. https://www.nature.com/articles/s41586-023-06832-9
[6] Lewis S, Hempel T, Jiménez-Luna J, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. Science. 2025;389(6761). https://www.science.org/doi/10.1126/science.adv9817

致谢

感谢《生物信息学》课程提供的学习与实践机会，感谢 UniProt、colab 等公共数据库与开源工具提供的数据与算法支持。

感谢小组成员的共同努力。

王煜：使用 BioEmu 生成目标蛋白构象并聚类

赵安祈：获取目标蛋白的实验构象和氨基酸序列，使用 AF-Cluster 生成目标蛋白构象

王誉凯：对 AF-Cluster 生成的构象进行聚类，绘制 RMSD 热图

胡锦睿：寻找隐藏口袋，进行分子对接验证

史永泰：获取目标蛋白的实验构象和氨基酸序列，部分结果可视化，撰写报告

项目文件由小组成员共同维护，已发布在：
https://github.com/kinggmars/Pyrophosphokinase-