arXiv:2601.06885v1 [cs.ET] 11 Jan 2026

理解图形处理器上端到端蛋白质设计流水线的性能行为

Jinwoo Hwang、Yeongmin Hwang、TadiwosMeaza、Hyeonbin Bae，
*IEEE*会员Jongse Park，*IEEE*高级会员

摘要—近期计算领域的进展使蛋白质设计流水线能够在图形处理器上端到端运行，但其异构计算行为在系统层面仍缺乏充分研究。我们实现并分析了一种具有代表性的流水线，在不同输入和超参数条件下，分别从组件粒度和完整流水线粒度进行评估。我们的特征分析发现，GPU 利用率普遍较低，且对序列长度和采样策略高度敏感。基于这些发现，我们概述了未来的研究方向，并发布了开源流水线和性能分析脚本，以促进进一步研究。

索引词—**蛋白质设计、蛋白质工程、特征分析、生物信息学**

I. 引言

PROTEINS是驱动生物系统的基础分子机器。蛋白质人工智能模型的近期进展 [1]–[8]显著降低了计算蛋白质结构预测的成本，使研究人员能够设计出具有增强功能或新颖功能的蛋白质，从而服务于下一代治疗药物、工业生物催化剂和生物传感器的开发 [9]。蛋白质设计流水线通常包含多个阶段，各阶段具有不同的计算行为并高度依赖图形处理器，但其组合后的系统级性能影响仍基本未得到探索。

我们通过在组件和流水线两个粒度上，对具有代表性的基于图形处理器的蛋白质设计流水线进行系统分析，考察序列长度、超参数和采样策略对性能的敏感性，从而弥补这一空白。我们的研究探讨了图形处理器利用率模式，以及共置场景下图形处理器调度的影响。我们还提出了未来研究方向，包括开发成本感知指标、扩展至多图形处理器环境，以及探索智能体辅助编排。为促进可复现性并加速未来研究，我们将流水线及其分析工具以开源软件形式发布。[1]

II. 背景与动机

A. 蛋白质与计算机模拟筛选

蛋白质工程。蛋白质功能的多样性源于其由氨基酸编码的三维结构
序列。改变氨基酸的顺序和组成，因而可以改变蛋白质的功能。从历史上看，发现和优化在很大程度上依赖物理筛选以及

稿件收到日期：2025年11月13日；接受日期：2025年12月16日。出版日期：2025年12月19日；当前版本日期：2026年1月6日。数字对象标识符10.1109/LCA.2025.3646250。本工作获得韩国政府（科学技术信息通信部）资助的韩国国家研究基金会（NRF）项目（RS-2024-00342148）支持。

作者供职于韩国科学技术院计算机学院，韩国大田，邮编34141。（电子邮箱：jwh- wang,ymhwang,tadiwos,hbbae,jspark@casys.kaist.ac.kr）。1

https://github.com/casys-kaist/protein-design-流水线.git

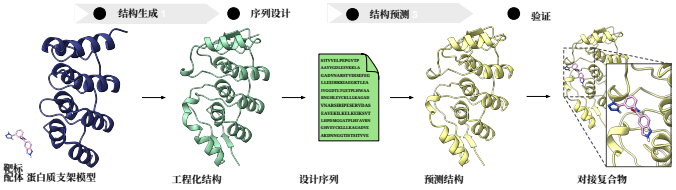


图1。蛋白质设计流程的四个主要阶段。

结构表征既昂贵又耗时，限制了对蛋白质设计空间的探索。

人工智能驱动的计算机模拟筛选。人工智能的最新进展通过大规模提供可靠的结构和功能预测，改变了这一工作流程。借助这些模型，与传统湿实验相比，以计算方式评估设计的边际成本显著降低。因此，过去需要数月的设计周期，如今使用规模适中的GPU集群只需数小时，从而为快速计算机模拟筛选这一新范式铺平了道路。

高通量筛选。这一范式涉及快速评估规模庞大的蛋白质候选物虚拟文库。早期计算筛选器根据功能性、可制造性和稳定性评估候选对象，及早淘汰不合适的设计。剩余的高优先级候选对象将受到重点实验关注，只有最优秀的设计才会进入实体验证。在此条件下，高通量计算机模拟筛选变得至关重要。大批量处理能够扩大设计空间的探索范围，并提高识别出优质候选对象的概率 [10]。

B. 蛋白质设计工作流程

蛋白质设计工作流程因具体目标而异，并由可互换的组件组成。不过，通常遵循四阶段模式，如图 1 所示。

- **结构生成**。结构生成利用 RFdiffusion 等模型，提出满足几何或功能约束的工程化蛋白质结构 [2]。结构既可以从头设计，也可以基于已知的支架蛋白构建，通常还会指定用于优化的靶向配体。
- **序列设计**。序列设计利用 ProteinMPNN 等逆折叠模型，生成与给定结构相匹配的氨基酸序列 [3]。包括 ESM-2 [4], 在内的蛋白质语言模型还能进一步优化并探索序列。
- **结构预测**。结构预测利用 AlphaFold 等预测器将序列折叠回三维结构 3 [1],ESMFold [4], 和 Protenix [6]。除 ESMFold 外，大多数模型都需要进化背景，会通过 MMseqs 进行同源性搜索和序列比对2 [5]。
- **验证**。评估通过形成对接结构来估算结合亲和力和结构稳定性等关键属性

表I 所选支架和配体。我们报告PDB编号及长度：氨基酸残基（氨基酸残基）和重原子（重原子）。

	短	中	Long
支架编号（氨基酸残基）	1SHG（62）	1N0R (126)	1TIM (247)
配体编号 (重原子)	3DX1 (9)	4DE1 (23)	3PRS (50)

复合物，并应用 AutoDock Vina 和 DiffDock 等评分方法 [7], [11], [12]。如需全面了解背景和工具比较，请参阅近期的一篇综述 [13]。

C. 组件间异质性

多样化的工作负载。各组件采用不同的模型架构和内核，涵盖扩散、 [1], [2], [6], [7],Transformer、 [1], [4], [6], 图神经网络、 [3], 用于序列比对的自定义 GPU 内核 [5], 以及评分方法 [12]。这种架构多样性带来了不同的计算模式、激活占用和内核行为。输入和超参数进一步加剧了这种多样性。有些组件对序列长度较为敏感，还有一些组件提供可调超参数，用于权衡吞吐量和准确率，例如扩散步数。

特征分析的必要性。这些见解促使我们开展两项互补分析。首先，我们针对各种输入和超参数，对各个组件进行特征分析。随后，我们在共享 GPU 上评估完整端到端流水线的性能，并分析系统层面的调度交互及其对性能的影响。

III. 蛋白质设计流水线

A. 蛋白质工程场景

我们以一个蛋白质工程场景为例：对一种稳定且经过充分特征分析的支架蛋白进行改造，使其能够结合特定的小分子靶标配体。

蛋白质与配体对。我们选取四种蛋白质支架和三种配体，以涵盖不同的序列长度和复杂度，详情见表I。对于蛋白质支架，我们采用了从短序列到长序列的成熟结构。对于配体，我们从 CASF-2016 [14] 数据集中抽取三种分子，分别选取长度处于第 1、50 和 99 百分位数的样例。

组件与超参数。我们为各流水线组件选取了广泛采用的模型，以体现当前行业标准，如第 II-B 节所述。

在超参数方面，我们设置了三种配置：吞吐量优化、均衡和准确率导向。均衡配置在可行时采用参考实现中的默认设置，而其他配置则调整计算资源，以兼顾吞吐量和质量。表 II 汇总了各组件的这些超参数。

范围与局限性。 这一合成场景旨在全面表征系统性能。它反映了尺寸驱动的影响以及流水线组件之间的异质性，并涵盖在受控改变蛋白质和配体尺寸的条件下具有代表性的计算条件。然而，它并未

表 II 组件与超参数。THR./BAL./ACC. 分别表示吞吐量优化、均衡和准确率导向。

组件	超参数	Thr.	Bal.	Acc.
RFdiffusion	num_步	20	50	100
ProteinMPNN	–	–	–	–
ESM-2	模型 _size	150M	650M	3B
MMseqs2	最大迭代次数	–	3	5
Protenix	–_步数_	100	200	300
Protenix	循环次数_	2	4	8
ESMFold	循环次数_	2	4	8
AlphaFold3	步数_	100	200	300
AlphaFold3	num_次循环	5	10	15
DiffDock	num_步	10	20	30
Vina-CPU	搜索强度	4	8	16
Vina-GPU	–	–	–	–

评估生化相容性，也不评估特定骨架-配体对在实验上的可行性。

B. 实现细节

软件配置。每个组件都使用 Docker 容器化，以确保可复现性和可移植性。工作流由 Nextflow [15], 编排；Nextflow 是一种广泛使用的生物信息学工作流管理器。任务通过 Kubernetes 调度，并借助阿里巴巴 GPUShare 插件将 GPU 内存作为可调度资源，从而防止并发操作期间发生内存不足错误。根据我们的工作负载特征，实际观察到这些编排层对性能的影响可以忽略不计。软件栈包括 Ubuntu 22.04、Nextflow 25.04、Docker 28.1 和 Kubernetes 1.33。

硬件配置。实验在一台服务器上进行，该服务器配备一块 RTX 3090 GPU、两颗 Intel Xeon Gold 6326 CPU、256 GB 动态随机存取存储器和一块 4 TB PCIe Gen 4 NVMe 固态硬盘。

IV. 性能表征

A. 组件级观察

图 2 展示了蛋白质设计流水线中各个组件在三个运行点和九组支架–配体配对下的性能分析，这些配对涵盖了不同的序列长度。我们通过 NVIDIA 管理库（NVML）测量墙钟时间、时间利用率和峰值设备内存使用量。为了评估空间利用率，我们首先使用 Nsight Systems 汇总执行时间至少占 90% 的内核，然后使用 Nsight Compute 计算运行时间加权平均 SM 占用率。

输入和超参数。输入长度会显著影响各组件的运行时间。RFdiffusion 的增幅尤其明显，支架长度从 1SHG 增至 1TIM 时，运行时间最高增加至 5.1 ×。其他组件的运行时间也有所增加，但幅度较小。MMseqs2 的运行时间会出现波动性峰值，其影响因素不仅是序列长度，还包括数据库交互等。调整超参数始终会影响运行时间，但总体而言，输入长度的影响更为显著。与吞吐量导向设置相比，准确率导向配置会延长执行时间。

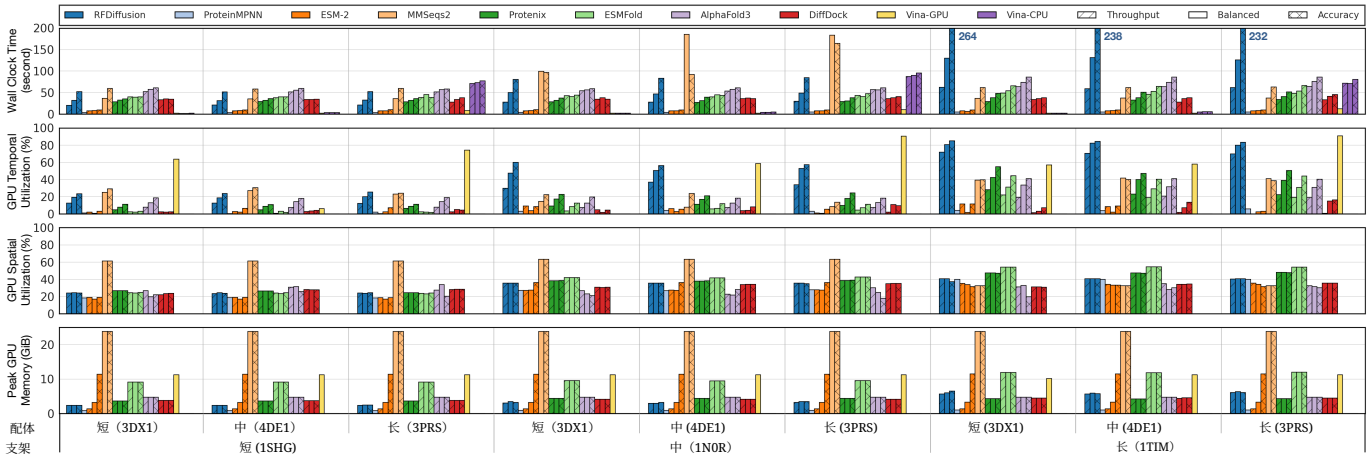


图2. 不同输入和超参数下各个流水线组件的性能分析。Vina-CPU 不使用 GPU 资源。Vina-GPU 基于 OpenCL 实现，而 OpenCL 不支持空间分析；因此，此处仅报告时间利用率。

GPU 利用率。各组件的 GPU 利用率随时间变化显著。RFdiffusion 和 Vina-GPU 的利用率相对较高。MMseqs2 的利用率适中，约为 26%。对于较大的配体，Vina-GPU 的利用率明显提高，可达 70% 至 91%。序列较长时，RFdiffusion 的利用率也会提高。输入较大时，ESMFold、Protenix 和 AlphaFold 3 偶尔也能达到中等利用率，但总体上仍以低利用率为主。

空间利用率总体较低，主要集中在24%至38%之间。值得注意的例外包括 ESMFold 的大型输入场景，其利用率约为54%；MMseqs2 的利用率则始终较高，为61%至63%。在 RTX 3090 上进行的实验表明，模型规模和执行参数未能充分发挥图形处理器的计算能力，说明仍有优化空间。

GPU峰值显存。除 MMseqs2 外，GPU峰值显存使用量始终低于设备的24 GB上限，表明仍有充足余量可同时运行多个进程。

B. 流水线级观察

我们将分析从独立组件扩展到完整的端到端流水线，并采用包含中等长度支架和配体的均衡配置。在结构预测阶段，我们选择 Protenix，在验证阶段则使用 Vina-GPU。

用于探索的采样。多个组件提供了采样超参数，可扩大搜索空间。RFdiffusion 能为每个提示生成多个骨架结构，ProteinMPNN 能为每个骨架生成多个序列，ESM模型可提出序列变体，AlphaFold 风格的预测器可返回多个结构候选项，而 AutoDock 会评估多个对接构象。当上游组件输出多个候选对象时，这些候选对象会传递到下游。

为说明采样的影响，我们采用统一的采样配置进行实验，将所有支持采样的组件的采样数量设为 1、2 或 3，分别生成 1、16 和 81 个样本。图 3 显示了

各案例的归一化运行时间分解。每个组件仅处理一个样本时，总运行时间与图 2 所示的各组件级运行时间之和一致。复用

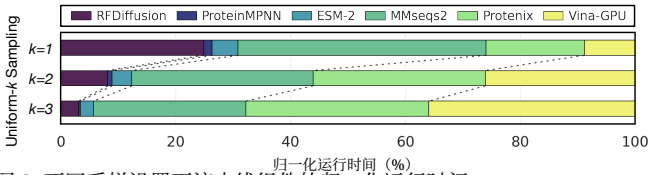


图 3. 不同采样设置下流水线组件的归一化运行时间。

早期阶段的共同输出可显著提高探索吞吐量，使每样本延迟从 4.32 分钟（uniform-1）降至 1.18 分钟（uniform-2），并进一步降至 0.83 分钟（uniform-3）。随着样本数量增加，运行时间逐渐转移至下游组件，这反映了工作负载的乘法式扩展。

GPU 共置。图 4(a) 和 (b) 分别展示了在单个图形处理器上，对于均匀采样次数为 2 次和 3 次时，有无共置情况下的 GPU 时间利用率。在多个进程分时共享 GPU 资源的共置模式下，时间利用率和吞吐量得到提升，从而缩短了总体执行时间。随着采样次数增加，这一优势进一步扩大，并提供更大的调度灵活性。具体而言，GPU 共置使均匀二阶采样的延迟降低了 11.9%，使均匀三阶采样的延迟降低了 27.7%。由于峰值内存需求较低，后期流水线组件尤其受益于共置，从而能够实现更多并发执行。

多 GPU 扩展。图 4(c) 和 (d) 评估了在启用共置的情况下，使用更高的四阶和五阶均匀采样数扩展至多 GPU 的效果。在早期阶段，可用样本数量有限，许多 GPU 处于空闲状态，因此平均时间利用率较低。随着可用样本逐渐增多，利用率也逐步提高。在当前的均匀采样级别下，GPU 总体利用不足，导致扩展效率不佳。提高采样数能带来更显著的扩展效果。例如，均匀五阶采样下，从一个 GPU 扩展到两个 GPU 可将延迟最多降低 42.8%；而均匀四阶采样下，改善幅度较小。当扩展到三个和四个 GPU 时，扩展效率明显下降，延迟分别仅降低 47.6% 和 50.2%。在执行初期提高 GPU 利用率，并在多个 GPU 上实现一致的扩展

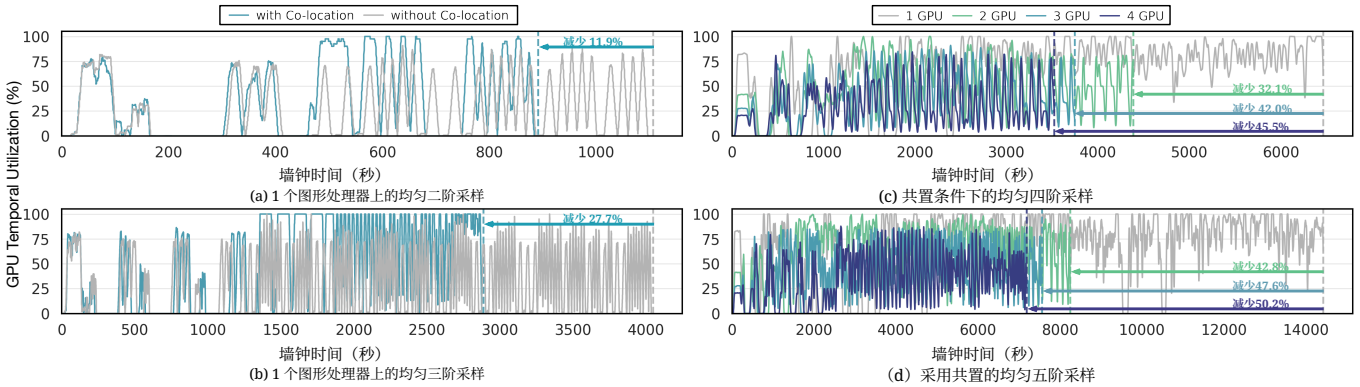


图4. GPU 时间利用率在执行时间内的变化。

能力，是未来工作的关键考量。

V. 未来方向

计算成本分析。 在模型系列和超参数之间进行选择，会改变准确率与吞吐量之间的帕累托前沿。诸如改用更大型的 ESM 变体、选择 ESMFold 或 Protenix 而非其他预测器，或调整对接种子策略等决策，都会影响质量和运行时间。早期探索可以优先采用快速近似方法，之后再提高保真度。然而，当前测量结果取决于输入，因此难以跨研究进行比较。我们建议采用标准化指标，例如每样本延迟，或在既定质量阈值下每个合格候选项的能耗。

大规模配置。 随着采样量增加，或执行多聚体预测等任务，计算需求会大幅增长，使流水线扩展变得更加复杂。在多GPU环境中进行大规模部署，会带来资源放置和调度方面的挑战。未来的工作应研究同时考虑计算资源和内存的资源打包策略，并支持多GPU调度。CPU 和存储等其他资源也可能变得更加重要。还需研究分区方法（例如 MIG、MPS）与长度分桶策略之间的相互作用，以减少碎片化。

智能体大语言模型。 候选对象集的扩大加剧了人工监督这一瓶颈。智能体语言模型（LLM）能够对候选对象排序、自适应地分配计算资源，并为升级处理或提前停止等决策提供理由。展示了智能体驱动的科学发现的 Coscientist [16]，，凸显了其在蛋白质设计流水线中类似的编排作用。将 LLM 集成为新的计算组件，需要分析其在现有流水线基础设施中的调度策略、资源管理策略和计算行为。

VI. 相关工作

据我们所知，在真实的共置场景下对完整蛋白质设计流水线进行特征分析的研究仍然有限。在组件层面，大多数工作聚焦于结构预测，包括对推理 [8]和训练 [18],的特征分析 [17]，系统级优化，以及支持更长序列的专用集成电路设计 [19] 。同源性搜索也已实现显著的 GPU加速 [5]。我们的研究通过强调完整流水线中的端到端行为、资源共享和利用模式。

VII. 结论

我们对蛋白质设计流水线进行了可复现的、以图形处理器为中心的特征分析，揭示了计算异质性和图形处理器利用不足。研究结果强调了图形处理器感知资源管理、成本感知评估指标，以及多图形处理器配置下可扩展性分析的重要性。通过发布开源产物，我们支持未来开展可比较研究并进行实用的流水线优化。

参考文献

[1] J. Abramson 等，“利用 AlphaFold 3 准确预测生物分子相互作用结构”，《自然》，卷 630，页 493–500，2024年。[2] J. L. Watson 等，“利用 RFdiffusion 从头设计蛋白质结构和功能”，《自然》，卷 620，页 1089–1100，2023年。[3] J. Dauparas 等，“使用 ProteinMPNN 进行稳健的深度学习蛋白质序列设计”，《科学》，卷 378，页 49–56，2022年。[4] Z. Lin 等，“利用语言模型进行进化尺度的原子级蛋白质结构预测”，《科学》，卷 379，1130，2023年，1123页。–。[5] F. Kallenborn 等，“利用 MMseqs2 进行 GPU 加速同源性搜索”，《自然·方法》，页 1–4，2025年。[6] 字节跳动 AI4Science 团队，“Protenix：通过全面复现 AlphaFold3 推进结构预测”，bioRxiv，2025年。[7] G. Corso 等，“DiffDock：用于分子对接的扩散步骤、扭转与转动”，载于国际学习表征会议，2023年。[8] M. Mirdita 等，“ColabFold：让所有人都能进行蛋白质折叠”，《自然·方法》，卷 19，页 679–682，2022年。[9] D. Listov 等，“蛋白质功能设计与优化中的机遇和挑战”，《自然》，卷 25，页 639–653，2024年。[10] N. Raouraoua 等，“MassiveFold：通过优化的大规模并行采样揭示 AlphaFold 的隐藏潜力”，《自然计算科学》，卷 4，期 11，页 824–828，2024年。[11] O. Trott 和 A. J. Olson，“AutoDock Vina：通过新的评分函数、高效优化和多线程提升对接速度与准确率”，《计算化学杂志》，卷 31，页 455–461，2010年。[12] S. Tang 等，“Vina-GPU 2.1：进一步优化 AutoDock Vina 及其衍生工具的对接速度和精度”，《IEEE/ACM计算生物学与生物信息学汇刊》，2024年。[13] K. I. Albanese 等，“计算蛋白质设计”，《自然综述：方法入门》，卷 5，页 13，2025年。[14] M. Su 等，“评分函数的比较评估：CASF-2016 更新”，《化学信息与建模杂志》，卷 59，页 895–913，2018年。[15] P. Di Tommaso 等，“Nextflow 支持可复现的计算工作流”，《自然》，卷 35，页 316–319，2017年。[16] J. Gottweis 等，“迈向人工智能协同科学家”，arXiv，2025年。[17] J. Kim 等，“AlphaFold 3 工作负载特征分析：瓶颈和性能扩展的综合分析”，载于国际工作负载特征分析研讨会，2025年。[18] H. La 等，“MegaFold：用于加速蛋白质结构预测模型的系统级优化”，arXiv，2025年。[19] S. Han 等，“LightNobel：通过自适应激活量化改善蛋白质结构预测模型的序列长度限制”，载于国际计算机体系结构研讨会，2025年，页 1940–1955。