arXiv:2611.28529v1 [q-bio.GN] 21 Nov 2026

通用语言模型是能够根据序列注释蛋白质的神经符号推理器

Christopher J. Mungall* Marcin P. Joachimiak

美国加利福尼亚州伯克利 94720，劳伦斯伯克利国家实验室环境基因组学与系统生物学部

摘要

确定生命之树中各种基因的功能，是生物学面临的主要挑战之一。针对这一问题的生成式人工智能方法通常分为两类：一类是基于生物序列训练的语言模型，能够根据序列中的信号预测结构与功能；另一类是基于文本训练的通用语言模型，能够以符号方式推理生物学功能。通常的假设是，蛋白质语言模型不擅长符号任务，而通用语言模型不擅长序列任务。然而，近期一份报告显示，前沿通用模型在下一个残基预测任务上的表现出乎意料地好。我们着手探索这些能力是否也适用于蛋白质注释。

我们为多个前沿通用模型分配了各种蛋白质注释任务，包括预测 Pfam 结构域是否存在及其坐标、功能性催化残基的缺失，以及预测结构与活性。在所有注释任务中，前沿模型的表现都远高于机会水平；我们没有将其与专用工具进行基准比较。这些模型还会产生具有特征性的幻觉错误，并且对这些错误的表述与正确判断一样自信。

尽管尚未与专门方法进行对比测试，但前沿模型初步展现出的仅凭序列进行蛋白质注释的能力颇为引人关注。在单个模型中对序列、结构和符号进行推理的能力，使其能够采用神经符号方法，将序列、实验数据与文献中的知识相结合，从整体上改进对蛋白质全集的从头功能注释。

提示词、测试集、模型输出和代码：
https://github.com/ai4curation/fable-sequence-to-function结果摘要：
https://ai4curation.io/fable-sequence-to-function/

1 引言

蛋白质语言模型（如 ESM-2 和 ProtGPT2）从蛋白质数据库中学习序列统计信息，并支持结构预测和设计 [1, 2]。通用语言模型则是为其他任务而构建的。即便如此，Tu和Chen [3, 4] 最近有报告指出，其中两个模型具备序列知识。对于研究较少的未知功能结构域家族中的蛋白质，Claude Fable 5.2 和 Claude Opus 5.5 预测真实序列的下一个氨基酸的准确率远高于打乱后的序列，水平与专门的蛋白质语言模型相当。该报告还发现

*通讯作者：cjmungall@lbl.gov

在其他厂商测试的模型中，这类信号很弱或根本不存在，而较早期的 Claude 模型中则有较弱的信号。我们测试了来自两家提供商的四个模型：上述两个模型、一个较早期的 Claude 模型和一个 Gemini 模型。

通用模型也已接受蛋白质变体排序方面的基准测试。PG-LLM 将野生型序列、测定描述以及待排序的变体序列列表提供给了十三个此类模型，涵盖 276 项任务，其中 217 项来自 ProteinGym，59 项来自近期发表的研究；整个过程中未使用工具、比对或结构。其中表现最佳的模型优于大多数仅序列预测器，但仍不及领先的专业预测器 [5]。该基准测试依据测得的适应度，对完整序列的排序进行评分。而这里的测试则要求在匹配诱饵的条件下，对残基和特征进行明确注释，并沿着两处与人工整理记录不一致的结果，继续追溯至相关证据。

下一残基准确率并不能说明模型是否能够完成注释员所需的任务。功能预测工具会产生一些人类注释员能够避免的典型错误 [6]。对于酶而言，一个重要问题是：将催化活性赋予某种蛋白质，而该蛋白质虽然保留了酶的折叠，却失去了这种活性。这类蛋白质称为假酶或失活酶 [7, 8]。一项覆盖整个生物界的分析估计，它们占酶家族中蛋白质的 5% 至 10% [9], ，而一篇关于其在 UniProtKB 中注释情况的综述给出的比例为 10% 至 15% [10]。许多假酶充当调节因子 [8], ，其中假激酶是研究最为深入的案例 [11]。只识别家族的模型会重复这一错误；能够读取残基的模型则不应如此。

仅凭残基进行判断也有局限。催化残基缺失可以预测酶不具活性，但不能证明这一点。对酶家族的调查发现，催化残基发生突变并不总能很好地表明酶缺乏活性 [12]。在某些蛋白质中，催化所需的残基已移至序列中的其他位置 [13]。例如，WNK1 缺少激酶亚结构域II中的催化赖氨酸，却仍是一种活性激酶；其活性需要亚结构域I中的 Lys233 [14]。

我们进行了四项小规模测试，并设置了旨在检验模型是否会响应序列本身的对照：对携带我们引入的单点替换的激酶及其他酶进行催化活性判定；进行结构域和膜特征注释，并以打乱序列作为对照；对近期发布的结构进行二级结构、接触和功能预测，并按先前的结构同源性分层；以及对下一残基结果进行子样本复现。

2 结果

2.1 研究设计

图 1 概述了研究设计。我们测试了来自两家提供商的四个模型（表 2），每个模型均参与全部四项测试：`claude-fable-5-2`，`claude-opus-5-5`，`claude-opus-4-6` 和 `gemini-4-neon`。所有模型使用相同的固定测试集和相同的提示词。努力程度设置采用各提供商自己的标签，因此不同模型和测试之间的设置并不相同。由于这些设置不能代表可比的推理预算，我们还报告了各模型记录的思考令牌数量（表 9）。在 特征、伪和STRFN 测试中，Opus 4.6 使用了两种设置；除非另有说明，我们报告的是其高努力程度运行结果。模型标识符采用端点返回的标识符。我们未确定任何模型的训练截止时间。其他 Gemini 模型（Gemini 3.1 Pro Preview 和 Gemini 3.1 Flash Lite）在部分测试中的运行结果，以及其他路由的不完整运行结果，均保存在代码仓库中，此处不予报告。

我们将 Fable 5.2、Opus 5.5 和 Gemini 4 Neon 称为前沿模型。Opus 4.6 是较早一代的 Claude 模型，我们将其用作对照。

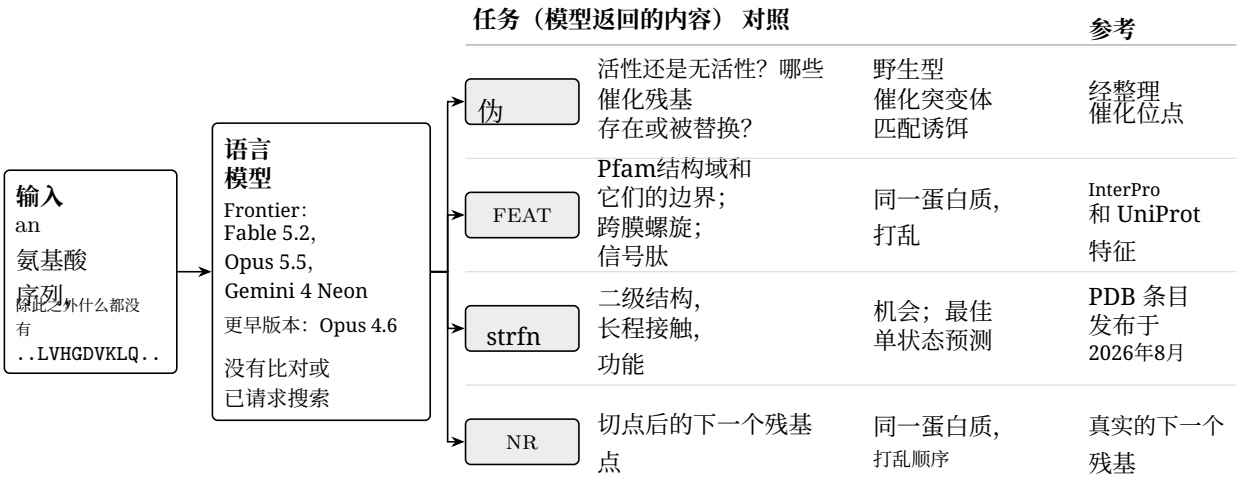

图1：研究概览。每个模型都会接收一条氨基酸序列和四项任务之一，任务名称采用全文使用的短代码。每项任务都设有一个对照，旨在区分读取序列与回忆家族标签。

各次运行在排除外部信息方面的严格程度有所不同，表1列出了三种路线。第一种 Claude 路线无法排除某些调用包含了额外上下文或连接器工具的可能性。Fable 5.2 和 Opus 5.5 的特征和 STRFN 结果仅来自这一路线；它们对 20 种命名为非活性的蛋白质所进行的调用也是如此。此外，除 Opus 5.5 的八个问题外，它们的NR 答案也仅来自这一路线。在这些结果中，只有伪野生型和突变体通过修订路线进行了重复测试。重试、输入大小和软件版本的详情见补充方法。

表1：模型的调用方式。我们在任何调用路径中都未提供工具或系统提示词 .

调用路径	排除的内容	Runs
Claude命令行工具，原始标志（18 11月18日）	内置工具 off, 空系统提示。额外上下文或连接器无法排除调用中使用工具的可能性。	Fable 5.2 和 Opus 5.5：四项测试全部通过。Opus 4.6：以低强度完成全部四项测试。
Claude命令行工具、修订后的标志（19 11月）	同上，用户设置和连接器工具不会纳入调用。	Opus 4.6 在高强度下（伪、特征、strfn）。诱饵对照，并重复伪野生型和突变体的问题。第一轮后还剩 8 个 Opus 5.5 nr 问题未提问。
Vertex AI API（19 11月19日）	请求正文不包含任何工具，也不包含系统指令。	Gemini 4 Neon：四项测试均通过。

向每个模型提供一条蛋白质序列，并要求其完成四项任务之一（详见第4节）。我们用括号中的简称指代这些测试。提示词、测试集以及每个模型的回复均存放在项目代码库中（https://github.com/ai4curation/fable-sequence-to-function）。

催化残基和假酶（伪）。判断一种酶是否具有催化活性，以及存在哪些催化残基或哪些残基发生了替换。序列包括20种活性激酶、催化性天冬氨酸或ATP结合赖氨酸被替换的同一批激酶、Swiss-Prot 标注为“非活性”的20种激酶折叠蛋白，以及20种其他酶（其中活性位点残基被丙氨酸替换或未被替换）。在针对 Claude 模型运行的诱饵对照中，同类

表2：每次报告的运行所使用的模型和努力程度设置。标签由各提供商自行定义，不同提供商之间并不匹配。

模型	提供商	伪	FEAT	strfn	NR
Claude Fable 5.2	Anthropic	high	high	high	低，高
Claude Opus 5.5	Anthropic	high	high	high	低，高
Claude Opus 4.6	Anthropic	低，高	低，高	低，高	low
Gemini 4 Neon	Google	low	low	中等	low

替换被放在一个没有功能注释的位置。该测试衡量的是敏感性：当带注释的催化残基被替换时，模型的判断是否会改变；而当匹配的其他位置残基被替换时，模型的判断是否不会改变。我们没有测量任何突变体或诱饵的活性。

结构域和膜特征（特征）。为45种抽样的细菌和古菌蛋白质，以及15种来自冷僻的未知功能结构域家族的蛋白质，标出 Pfam 结构域及其边界、跨膜螺旋和任何信号肽。每种蛋白质也都提供了打乱后的形式。

结构与功能（strfn）。为2026年8月发布的30个 PDB 结构预测二级结构、残基接触和功能。若某条目在2026年6月之前没有发布过序列一致性达到或超过30%的 PDB 实体，我们就称其为“未匹配”；若至少有20个这样的实体，则称其为“匹配”。这些标签只描述这项检索，不代表其他含义：未匹配蛋白质不一定是科学界或模型中的新发现。一种未匹配蛋白质——人类 ASPRV1——被抽取了四次，计为一种，因此共有12种未匹配蛋白质。

下一个残基（nr）。从截断点继续补全序列。对照是同一蛋白质，其残基经过打乱。

2.2 一次 c 检测到催化替换，并忽略匹配诱饵

Fable 5.2、Opus 5.5 和 Neon 均将全部 20 种野生型激酶判定为活性，并将 20 种天冬氨酸突变为天冬酰胺的突变体全部判定为非活性（图 2）。在 60 条突变体序列中，它们分别报告称 57、47 和 47 条序列的替换位置发生了替换（表 3）。对于赖氨酸，Opus 5.5 和 Neon 改变判断的情况较少：它们分别将 20 种赖氨酸突变为甲硫氨酸的突变体中的 7 种和 4 种判定为活性。Neon 还将其他酶的 20 种丙氨酸突变体中的 8 种判定为活性。将突变体判定为“活性”，与我们关于该突变体为非活性的预测相悖。我们并未检验这一预测，因此这种判定并不构成已证实的错误。

对于赖氨酸和其他酶，Opus 4.6 改变判断的情况少得多。在高强度下，它将 20 种天冬氨酸突变为天冬酰胺的突变体中的 2 种判定为活性，并指出了其中 18 种的替换位置。但它将 20 种赖氨酸突变为甲硫氨酸的突变体中的 17 种，以及其他酶的 20 种突变体中的 14 种判定为活性，却仅分别指出了其中 1 种和 2 种的替换位置。在低强度下，它分别将 20 种突变体中的 18、18 和 16 种判定为活性，且没有指出任何替换位置，因此它对天冬氨酸突变体的结果取决于努力程度。

诱饵对照表明，Fable 5.2 和 Opus 5.5 会对催化位点作出响应，而不是对任何变化都作出响应（见表 4）。通过一条路线提出的每个问题中，两个模型都判定全部 58 个诱饵具有活性，而在同一蛋白质的 58 个催化位点突变体中，仅分别判定 3 个和 8 个具有活性。两者都没有将诱饵位点称为缺失残基。有一个案例结果相反：Opus 5.5 未判定一种甘露糖基转移酶（P26725）的野生型具有活性，却判定其诱饵具有活性。仅以低强度向 Opus 4.6 提供诱饵时，它判定诱饵、突变体和野生型具有活性的比例相近。

表 3：伪测试：模型报告了替换位置的突变体，每行 20 个。Claude 模型采用高强度；Neon 采用低强度。

突变体类别	Fable 5.2	Opus 5.5	Opus 4.6	Neon
激酶 Asp→Asn	20	19	18	20
激酶 Lys→Met	20	13	1	16
其他酶位点→丙氨酸	17	15	2	11

通过修订路线再次询问突变体和野生型，也重复了伪测试。对于 Fable 5.2，100 个问题中有 99 个的活性判定与首次运行一致；对于 Opus 5.5，则有 96 个一致。我们没有对 Neon 进行诱饵对照。它判定全部 20 个野生型激酶具有活性，这排除了笼统回答“非活性”的可能，但不能排除它倾向于拒绝任何改变后的序列。

表 4：诱饵对照，所有列均由一种方法查询：38 种蛋白质、58 个催化位点突变体，以及 58 个在未注释位置携带相同替换类型的诱饵。计数为被判定具有催化活性的序列数。

模型（努力程度）	野生型	催化位点突变体	诱饵
Fable 5.2（高）	38/38	3/58	58/58
Opus 5.5（高）	37/38	8/58	58/58
Opus 4.6（低）	34/38	45/58	50/58

Fable 5.2、Opus 5.5 和 Neon 分别将 Swiss-Prot 命名为非活性激酶的 20 个蛋白质中的 18、18 和 19 个判定为非活性。这表示它们与数据库名称一致，而不是衡量生物学准确性。在这三组答案中，分别有 17、17 和 19 个蛋白质的名称已经包含“非活性”或“假激酶”，因此该集合主要衡量的是对条目名称的召回率。名称本身也不能证明其非活性：我们依据名称选出了这 20 个蛋白质，并仅将其中三个标签与文献进行了核对（见下文）。高强度的 Opus 4.6 将 20 个中的 12 个判定为非活性。

至少有两个案例中，名称与已发表的测定结果相矛盾。拟南芥 RLK902 (Q9LVI6) 和 RKL1 (Q9LP77) 的激酶结构域经大肠杆菌 表达为融合蛋白并纯化后，在与放射性标记ATP孵育时均发生了磷酸化；作者据此得出结论，认为二者都是“具有自磷酸化能力的功能性蛋白激酶” [15]。Swiss-Prot 中 RLK902 的条目在预测其激酶结构域为非活性的同时，也注明了这一结果 [16]。Fable 5.2、Opus 5.5、Neon 和 Opus 4.6 在高强度下均判定这两种蛋白质为非活性。因此，认同名称并不等于判断正确。单项体外自磷酸化实验也不足以确定这些蛋白质在植物中的作用。

三个表现较强的模型都将秀丽隐杆线虫kin-32（Q95YD4）判定为活性激酶，尽管 Swiss-Prot 将其命名为非活性酪氨酸激酶。它们提到了完整的甘氨酸环、VAVK赖氨酸、HRD和DFG基序，这些在序列中确实都存在。Swiss-Prot 的命名依据是测定，而非序列：该条目记载，这种蛋白质“在哺乳动物细胞中表达时，体外似乎没有酪氨酸激酶活性”。在该研究中，从转染的小鼠细胞中免疫沉淀得到的 kin-32 对一般酪氨酸激酶底物没有可测得的活性[17]。作者指出，kin-32 缺少哺乳动物黏着斑激酶发生自磷酸化的那个酪氨酸，并提出它可能是在秀丽隐杆线虫体内发挥作用、但在小鼠细胞中未被激活的激酶。因此，典型基序完整并不能保证检测到活性，而仅凭对催化残基的解读，也无法将这种情况与活性激酶区分开来。一次测定也不足以确定 kin-32 在秀丽隐杆线虫体内是否非活性。

在被判定为活性的突变体中，反复出现两种模式。第一种模式中，模型将催化作用转移给了同类型的邻近残基。我们将葡萄糖醛酸基转移酶（Q9VLA1）的 Glu318 改为丙氨酸后，Fable 5.2 和 Opus 5.5 将 Glu317 指定为催化碱，并判定该酶有活性，尽管 Fable 5.2 在野生型中指定的是 Glu318。Neon 也将该突变体判定为有活性，但理由不同：它在野生型和突变体中都列出了 Asp233 和 Asp235，却没有列出催化谷氨酸。

在第二种模式中，模型报告了替换，随后围绕这一点展开论证。对于弓形虫 激酶 ROP17 (Q2PAY3)，Fable 5.2 指出赖氨酸位置上是蛋氨酸，并提出邻近赖氨酸会取而代之。催化赖氨酸移位的情况在 WNK 中已有先例1 [14], ，但在这里，模型将这一论据用在了我们对一种活性激酶所做的替换上。在这种模式中，模型确实检测到了替换；与我们的预测不同的是，它对活性的判断。

另外两处分歧源于存在争议的参考位置。对于核苷酸酶 YrfG (P64636) 和还原酶 ENV9 (Q08651)，Fable 5.2、Opus 5.5 和 Neon 对野生型和突变体中的催化残基都给出了不同于 UniProt 的位置，因此判定突变体具有活性。我们不将模型之间的一致视为某个位点的证据，以下分析也不会采用这种一致性。UniProt 标注的这两个位置均基于相似性，且这两种蛋白质都没有实验结构。

对于 YrfG，UniProt 将亲核体标注为 Asp9，而模型给出的是 Asp16。我们比对了 11 个经审阅的卤代酸脱卤酶超家族同源蛋白。其中 7 个具有实验支持的亲核体，且这 7 个均对应 Asp16。在两个结构模型中，Asp16 距镁离子不足 3.3 Å，距 Lys149 为 2.6 Å；而 Asp9 距二者均超过 20 Å。证据强烈支持 Asp16。

对于 ENV9，UniProt 将催化丝氨酸标注为 Ser172，而模型给出的是相邻的 Ser173。UniProt 引用的来源 O93868 中，催化丝氨酸位于 Ser149，该位点与 Ser173 对齐。此标注来自晶体结构 [18], ，UniProt 将其记录为注释员推断（ECO:0000305）。在两个结构模型中，Ser173 的侧链均指向催化 Tyr202 的羟基（距离分别为 3.2 和 4.5 Å），而 Ser172 的侧链则背向该羟基（距离分别为 7.0 和 6.8 Å）。ENV9 这一案例仍属暂定结论，因为候选残基彼此相邻，且来源中的标注是根据结构推断得出的。ENV9 已发表的活性测定对预测的活性位点 Asn146（N146L）和辅因子结合甘氨酸（G23–24A）进行了突变，每种突变都只留下残余活性。该论文将 Tyr202 和 Lys206 称为活性位点残基，但既未突变催化丝氨酸，也未提及它 [19]。

这两个案例令人质疑我们的替换是否移除了预期的催化残基。它们并不能证明这些突变体保留了活性。在排除这两种蛋白质的敏感性分析中，Fable 5.2 和 Opus 5.5 均将其他 18 个酶突变体中的 1 个判定为有活性，而 Neon 判定了 6 个；它们分别在 18 个突变体中的 17、15 和 11 个报告了替换位置。图 2保留了全部 20 个。

2.3 三种模型能够识别结构域和膜特征，且很少将结构域指派给打乱序列

Fable 5.2、Opus 5.5 和 Neon 分别为 45 个抽样蛋白质中的 39、35 和 20 个正确指出了至少一个 Pfam家族。对于这些蛋白质中正确命名的结构域，它们给出的边界与参考值相近，平均重叠度分别为 0.85、0.85 和 0.80（表 5汇总了全部 60 个）。对于来自不常见的未知功能结构域家族的 15 个蛋白质，Fable 5.2 为其中 8 个指出了确切家族，Opus 5.5 为 3 个指出了确切家族，而 Neon 一个也没有。Neon 的预测最为保守。它预测出 28 个结构域，

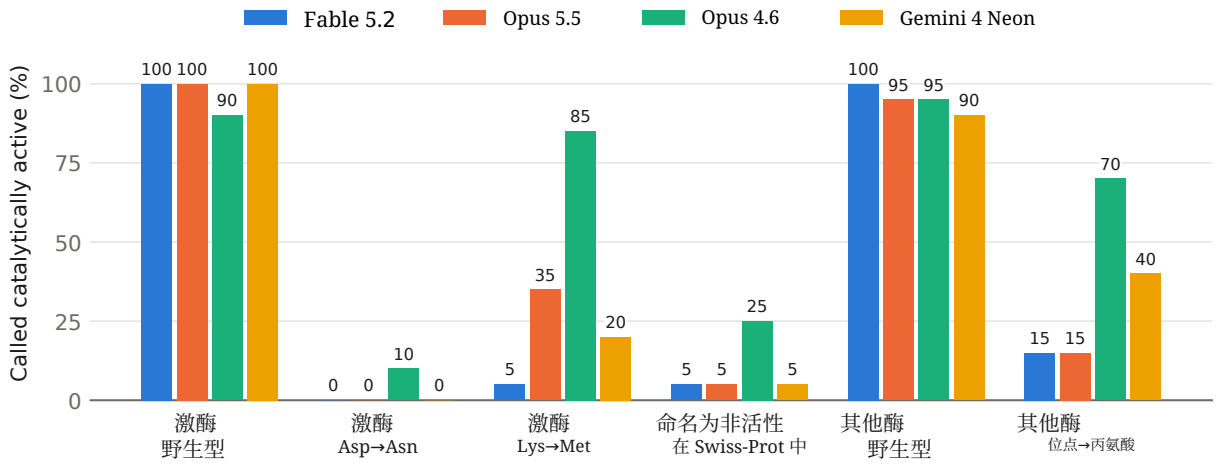

图2：被判定为具有催化活性的序列占比，每条柱代表20条序列。突变体带有一处为本研究引入的替换。Claude 模型采用高强度；Gemini 4 Neon 采用低强度。

其中 23 个的登录号正确；另外两个模型分别预测出 70 个和 51 个。对于打乱序列，三者几乎从未将其归入 Pfam 结构域（60 个中分别为 1、0 和 0 个）。不过，它们并非完全没有注释：三个模型分别在 60 个序列中的 12、15 和 15 个预测出了跨膜螺旋。这未必是错误，因为打乱序列保留了其组成，仍可能包含疏水片段。

高强度下，Opus 4.6 为45种抽样蛋白质中的7种正确命名了家族，并为60条打乱序列中的28条指派了结构域。此前一次运行的结果分别为45种中的5种和60条中的5条。根据作者的记录，此前的运行采用低强度；11月18日的 feat 行未记录该设置。因此，在修订路线上的高强度运行虽然召回率仅略有提升，却虚构了更多结构域。

模型在出错时没有给出任何迹象。Fable 5.2 将一种 DUF968 蛋白质归入了不同的家族。三个模型都拆分了多次跨膜转运蛋白中的螺旋，分别预测出 85、87 和 82 个螺旋，而注释结果为 74 个，因此找回每一条参考螺旋并不意味着膜注释就是正确的。两个测试蛋白质在 UniProt 中带有不合理的提交者名称：一个 128 个残基的 DUF3119 膜蛋白被标注为“甘油脱氢酶”，一个 DUF7147 蛋白质被标注为“甲硫核糖激酶”。Fable 5.2 和 Opus 5.5 都没有重复这些错误。

表5：仅根据序列进行注释。当家族判断正确时，交并比表示预测的结构域边界与参考结构域边界的重叠程度。前两行统计每个集合中的蛋白质数量；其余各行合并统计两个集合（60个蛋白质）。Claude 模型采用高强度，Neon采用低强度。存储的 Fable 5.2 和 Opus 5.5 行未记录设置（方法）。

	Fable 5.2	Opus 5.5	Opus 4.6	Neon
抽样蛋白质：正确的 Pfam家族	39/45	35/45	7/45	20/45
晦涩的功能未知结构域：正确的Pfam家族	8/15	3/15	0/15	0/15
预测结构域：正确的登录号	62/70	45/51	9/40	23/28
边界重叠（交并比）	0.86	0.86	0.64	0.80
恢复的跨膜螺旋	74/74	74/74	67/74	74/74
信号肽，真阳性 / 假阳性	9 / 0	9 / 0	9 / 4	9 / 0
给定任意结构域的打乱序列	1/60	0/60	28/60	0/60

2.4 对近期发布的蛋白质，结构预测结果高于机会水平；功能命名正确的比例不超过一半

对于 12 种未匹配蛋白质，Fable 5.2、Opus 5.5 和 Neon 的三状态准确率分别达到 76%、73% 和 69%，而最佳单状态猜测为 57%。它们的远程接触精确率分别为 37% 、 19% 和 28%，随机情况下为 2%（图 3，表 6）。这里的精确率是针对满足要求的配对计算的，即两个残基之间至少相隔 24 个残基，且两个残基都在结构中得到解析。统计所有提交的配对时，将不符合条件的配对计为错误，所得结果分别为 34%、18% 和 28%。对于匹配组，准确率为 78% 至 82%，接触精确率为 31% 至 48%。12 种蛋白质太少，无法根据数值差异对两个模型进行排名。

Opus 4.6 的表现差得多。在未匹配蛋白质上，采用高强度时，它的二级结构得分为 54%，低于单状态预测；接触精确率为 8%，高于机会水平，但远低于前三个模型。采用低强度时，这两个得分分别为 48% 和 4%。

我们没有对这些蛋白质运行 AlphaFold [20] 等专用结构预测器，因此不将结果与它们进行比较。我们仅声称，三个模型在未进行比对且关闭内置工具的情况下，表现显著高于机会水平。模型类别、努力程度和运行途径各不相同，因此这并非对不同代际或不同提供商进行的受控比较。

我们还询问了每个模型是否认出了该蛋白质。在 12种未匹配蛋白质中，Fable 5.2 表示认出了其中 2 种（ASPRV1 和 MrtR），Neon 表示认出了 1 种，Opus 5.5 表示一种也没认出，Opus 4.6 则表示认出了 4 种。这些自我报告与准确度并不一致：Opus 4.6 声称认出的最多，但结构得分最低。因此，认出数量无法证明召回率或训练数据暴露情况，而结构预测错误本身也不能证明认出声明是虚假的。

表 6：结构评分（百分比），ASPRV1 仅计一次：未匹配的蛋白质有 12 个，匹配的有 15 个；Opus 5.5 的匹配蛋白质有 14 个。Q3 是三态准确率；接触是对提交的蛋白质对计算的每蛋白质平均远程精确率，这些蛋白质对中两个残基均已解析，且相距至少 24 个残基。努力程度标签并不代表可比较的推理预算。Claude 行保留了列出多个实际提供模型的调用。

		未匹配		匹配	
模型	努力程度	Q3	接触	Q3	接触
Claude Fable 5.2	high	76.3	36.6	81.8	48.3
Claude Opus 5.5	high	73.4	19.0	78.3	31.4
Gemini 4 Neon	中等	69.0	28.4	78.6	48.0
Claude Opus 4.6	high	53.9	8.1	57.8	9.7
Claude Opus 4.6	low	48.1	4.1	45.1	4.0

两位评分者分别独立工作，在不知道答案由哪个模型生成的情况下，根据存档标题对三个前沿模型针对12种未匹配蛋白质的函数调用进行了评分。两位评分者都是 Claude 模型的实例，而非人类（方法）。在36项总评分中，两位评分者有32项达成一致。Fable 5.2 对5种蛋白质的判断正确、对2种部分正确；Opus 5.5 分别为4种和1种；Neon 分别为2种和1种。其余判断要么错误，要么属于弃答：根据评分者不同，Fable 5.2 和 Opus 5.5 各有2或3次错误调用，而 Neon 有4至6次。四项分歧全都涉及某个含糊回答究竟应视为弃答还是错误。没有任何模型对超过半数的蛋白质判断正确。大多数正确调用针对的都是大型、表征充分的家族成员。我们没有测试这些模型是通过预测结构得出结论，还是通过从序列中识别家族而得出结论。

表 7 给出了每种蛋白质的三项指标。它们并不同步变化。对于激酶结构域 9ZWQ，Fable 5.2 正确判断了19个接触中的18个，而对9TGX或9YP0一个也未判断正确；三者的二级结构准确率却相近。二级结构、三级接触和功能是彼此独立的能力。

表7：12种未匹配蛋白质，逐一列出。Q3以百分比表示，“基线”列给出最佳单状态猜测。接触表示正确配对数占符合条件的配对数的比例。功能为盲法总体评分：C表示正确，P表示部分正确，W表示错误，A表示弃答；评分者意见不同时列出两个字母。名称是从已提交条目中缩写而来。

		Fable 5.2			Opus 5.5		Gemini 4 Neon	
PDB	蛋白质	基线 Q3	接触	功能	Q3 连续	功能	Q3 连续	功能
37QB	酮泛酸还原酶 PanG	63	84 12/15	P	81 12/20	P	77 9/20	W
9TGX	胍基异戊烯基转移酶 NosH	86	82 0/19	W	84 0/7	A	76 0/20	A
9YP0	设计型半胱氨酸蛋白酶	43	79 0/20	W	76 0/16	W	72 0/20	A/W
10ZI	调节因子 MrtR	65	79 5/20	P	76 4/20	C	79 9/20	C
23AA	抗CRISPR AcrIIA17	58	78 2/19	A	73 1/19	A	72 0/20	A
9SA5	CBASS核苷酸环化酶	54	78 6/20	C	71 5/19	C	59 4/20	A/W
9ZWQ	N4BP2 多核苷酸激酶	53	76 18/19	C	72 8/20	C	70 14/20	C
27IK	ASPRV1 天冬氨酸蛋白酶	45	75 15/20	C	69 4/16	C	63 17/20	P
9T71	AtDUC1 去泛素化酶结构域	39	75 2/19	C	70 0/19	A	61 1/17	A
9WIA	昆虫效应蛋白HARP1	57	75 4/19	A	64 2/16	W	71 1/20	W
12WN	设计蛋白质 desE2D	60	72 11/20	C	80 6/18	A	75 9/20	W
10CW	RNase P 亚基 Rpp30	56	62 7/19	W/A	65 1/18	W/A	54 4/20	W

对于拟南芥 AtDUC1（PDB 9T71）的一个结构域，该结构以“EEIG1/EHBP1 蛋白氨基端结构域蛋白”之名提交，Fable 5.2 将其预测为一种具有半胱氨酸-组氨酸-天冬氨酸三联体的木瓜蛋白酶样半胱氨酸肽酶，“可能是一种去泛素化酶”。该条目展示的是其去泛素化酶结构域的结构。然而，这一判断不能归功于单凭序列作出的推断。描述该蛋白质家族为去泛素化酶的预印本于12月发布 2025 [21], ，比该结构公布早了八个月；此后训练的模型可能已经学到这一信息。我们没有搜索其他未匹配蛋白质的早期描述。

良好的二级结构预测并不意味着功能判断正确。对于胍基异戊烯基转移酶 NosH（9TGX），Fable 5.2 在一种几乎完全由螺旋构成的蛋白质上取得了82%的三态准确率，而将每个残基都判为螺旋即可得分87%。其19个符合条件的接触无一正确，而且它将该蛋白质命名为 TenA 家族硫胺素酶。没有任何模型识别出 RNase P 亚基 Rpp30（10CW）。对于一种设计型半胱氨酸蛋白酶（9YP0），Fable 5.2 和 Opus 5.5 均将其描述为一种天然细菌表面蛋白，且不具有催化功能。两位评分者都将这些判为错误声明，而非弃答。

Neon 的判断显示，在识别家族与注释蛋白质之间存在同样的差值。Neon 正确地将 9ZWQ 命名为多核苷酸激酶，并将 10ZI 命名为 AHL 响应型 LuxR 家族调节因子。对于保留的 ASPRV1 条目（27IK），它预测出了正确的家族——天冬氨酸蛋白酶，但将该蛋白质命名为 DDI2，并赋予其 DDI2 特有的底物生物学特征。它还将 PanG（37QB）归入 TyrA 脱氢酶家族，并将 Rpp30（10CW）归为一种代谢酶。

其中一个 ASPRV1 构建体（27IQ）意外地重复了伪测试，因为其催化性 DSG 基序被改为 NSG。尽管没有被问及活性，Fable 5.2 和 Opus 5.5 都指出，催化性天冬氨酸已被替换，该构建体可能没有活性。Neon 未识别出这一替换。

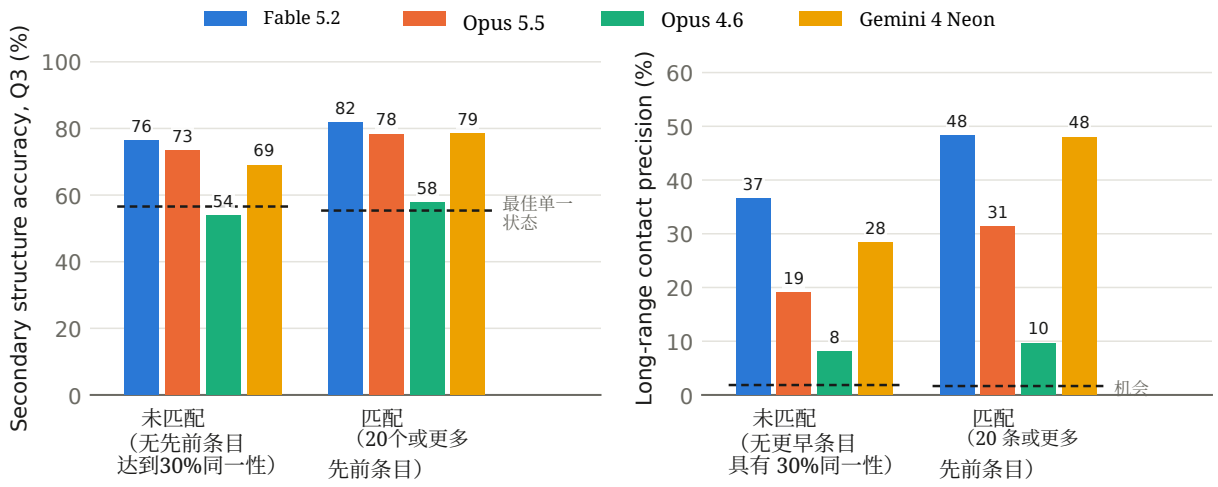

图 3：对 2026 年 8 月发布的 PDB 条目进行序列结构预测，ASPRV1 仅计一次。未匹配：2026 年 6 月之前发布的条目中，没有同一性达到 30% 或以上的 PDB 实体。匹配：至少有 20 个这样的实体。未匹配蛋白质有 12 个，匹配蛋白质有 15 个；对于 Opus 5.5，一个匹配条目的调用返回了 API 错误，而非答案。Claude 模型采用高强度设置；Gemini 4 Neon 采用中等强度设置。虚线分别表示最佳单状态猜测（左）和远程配对的机会率（右）。

2.5 在测试的四个模型中，有两个模型可检测到下一残基优势

引发本研究的报告测量了下一残基预测，我们在其部分问题上重复了这项测量。在两种努力程度设置下，Fable 5.2 对真实序列的预测都优于打乱序列；Opus 5.5 在高强度设置下也如此。在低强度设置下，Opus 5.5 的区间包含零。Opus 4.6 和 Gemini 4 Neon 均仅在低强度设置下运行，结果未显示差异（表 8）。Neon 在真实序列上的得分为 8.8%，在打乱序列上的得分为 10.4%，其十个残基的续写没有一个完全匹配。准确率远低于相同问题已发表的数值（Fable 5.2 为 58.8%，Opus 5.5 约为 40%）。我们怀疑原因在于答案经过了采样，以及我们的调用包含了额外上下文，但并未对此进行测试。仅在低强度设置下对 Claude 模型进行测试的反转序列，其得分与 Fable 5.2 的打乱序列相当；对 Opus 5.5 而言，得分介于打乱序列和真实序列之间。四种不使用模型的简单启发式方法（例如重复最后一个残基），在全部 3,340 个切分点上，对真实序列、打乱序列和反转序列的得分均为 7% 至 10%。

真实序列上的准确率随着努力程度设置的提高而上升。在相同的 136 个真实切分点上，Fable 5.2 在低强度下的得分为 25.0%，在高强度下为 41.2%；高强度下的每个答案都使用了思考令牌。在打乱序列上，其得分分别为 7.4% 和 11.8%。这是本研究中关于推理的唯一对照证据。这与原始报告不同，原报告中，不带思考过程的答案与带思考过程的答案准确率相同。

在一次运行中，使用思考过程的回答更常答对。不过，模型会自行选择何时进行思考，因此这一比较描述的是各次运行的情况，并不能衡量思考过程本身的作用。在低强度下，Fable 5.2 在 136 个真实切分点中的 38 个进行了思考，其中 57.9% 回答正确；相比之下，其余 98 个中的正确率为 12.2%。在打乱和反转的切分点中，它分别仅在 15 个和 17 个上进行了思考。在打乱序列上，思考过程没有带来帮助：使用思考过程时 15 个中有 0 个答对，而未使用时 121 个中有 10 个答对。在高

强度下，它对每个问题都进行了思考；在真实序列上的准确率为 41.2%，明显低于它自行选择进行思考的问题中的 57.9%。模型似乎会在识别出某些内容时进行思考，因此这种划分在一定程度上按照问题难度进行了分类。在高强度下，Opus 5.5 在它进行思考的 43 个真实切分点中有 37.2 % 答对，而在它未进行思考的 93 个切分点中，正确率为 10.8%。我们无法看到推理文本。

表8：68个家族的子样本上的下一残基准确率。差值为真实序列减去打乱序列，单位为百分点，并给出基于家族计算的95%自助法置信区间。Neon回答了136个真实序列问题和135个打乱的问题，且每个问题都进行了思考；Fable 5.2在低强度下，对136个真实切分点中的38个进行了思考。不同提供商的努力程度标签不可直接比较。

模型	努力程度	Real	打乱序列	反转序列	差值（95% 区间）
Fable 5.2	low	25.0%	7.4%	8.8%	+17.6 （+9.6至 +26.5）
Fable 5.2	high	41.2%	11.8%	–	+29.4 （+18.4至 +40.4）
Opus 5.5	low	16.9%	9.6%	12.5%	+7.4 （−1.5 至 +16.2）
Opus 5.5	high	19.1%	7.4%	–	+11.8（+3.7至 +20.6）
Opus 4.6	low	5.1%	7.4%	5.9%	−2.2 （−7.4至 +2.9）
Gemini 4 Neon	low	8.8%	10.4%	–	−1.5(−8.1 至 +5.1)

表9：每个评分答案记录的思考令牌中位数（包括零值）；所有评分行均有计数。nr 测试仅包括真实序列和打乱的问题，feat 测试则包括这两种条件。Opus 5.5 的结构中位数基于 29 个答案计算，未计入其 API 错误。

模型	伪	FEAT	strfn	NR
Gemini 4 Neon	2,474	1,049	17,881.5	634
Claude Fable 5.2	2,548	703.5	3,009	164
Claude Opus 5.5	1,275.5	418	1,342	0
Claude Opus 4.6	19,921	14,189.5	42,987	0

努力程度：Neon 在 nr、feat 和伪测试中为低，在 strfn 测试中为中；Fable 和 Opus 5.5 全程为高；Opus 4.6 在 nr 测试中为低，在其他测试中为高。对于 feat 测试，Fable 和 Opus 5.5 的设置依据作者对该数据集的记录，因为这些行中没有存储该信息。这些计数并不能说明各模型的预算相匹配。

3 讨论

四个模型中的三个——Fable 5.2、Opus 5.5 和 Gemini 4 Neon——所做的不只是识别家族。当一个催化残基发生变化时，它们也会改变答案，并指出该残基。对于在 30%同一性条件下没有更早 PDB条目的蛋白质，它们预测二级结构和接触的表现优于机会水平。它们也很少将 Pfam 结构域分配给打乱序列。这三个模型来自两家提供商，且通过不同途径获得，因此这种行为并非某一家供应商或某种调用模型方式所特有。我们尚未确定这种行为的适用范围。

Opus 4.6 表明这种行为并非普遍现象。它找到了全部 9 个信号肽、74 个跨膜螺旋中的 67 个，以及激酶天冬氨酸替换。但它漏掉了赖氨酸替换和其他酶家族中的活性位点变化，还为近一半打乱后的序列臆造了一个结构域。这种对比仅描述了我们观察到的情况。努力程度标签并不能给实际记录的思考过程排序（表9），访问途径各不相同，也没有在每种设置下都运行模型。

对于 Fable 5.2 和 Opus 5.5，诱饵对照表明，这些判断与催化位点相符。由于诱饵避开了其他保守位点，该对照无法说明模型如何处理这些位点上的变化，而且 Neon 未进行这项测试。在模型与整理后的位点不一致的两个案例中，比对和结构证据更支持模型提出的位置，其中 YrfG 的证据比 ENV9 更有力。因此，仅依据序列进行分析，也可能有助于标记那些因按相似性转移注释而被错误定位的位点。不过，两个案例不足以得出发生率。

在我们的下一残基测试中，其他能力均未表现出可检测到的优势。Neon 未显示出可检测到的下一残基优势，但它拒绝了全部 20 个天冬氨酸突变体，报告了 20 个赖氨酸替换中的 16 个，并达到了 28.4% 的接触精确率，而随机机会水平为 1.8%。这些任务彼此独立，但这并不能证明它们依赖不同的机制，也不能证明模型缺乏序列知识；而且，零结果只能提供微弱的证据。Neon 仅以低强度运行了 NR 测试。低强度下的 Opus 5.5 也未显示出可检测到的优势。Neon 在回答的每个问题上都使用了思考过程，而低强度下的 Opus 5.5 则一个问题也没有使用，因此我们无法判断 Neon 的零结果究竟源于模型本身还是设置。Claude 模型在完全相同的问题上的得分远低于已发表的数值（第 2.5 节），因此这一设置压低了所有模型在测试中的表现。在其主要测试集上，被引用的报告运行了 Gemini 4 Carbon，并通过原始续写方式运行了 Gemini 2.5 Pro，但没有运行 Neon。在条件允许时，报告采用温度 0，并关闭思考过程或将其设为最低档；结果发现这两个模型的得分都接近零 [3]。Neon 的结果与他们的结果一致，但并未复现该实验。

对于注释审查而言，实际用途有限。在接受某个催化术语之前，仅凭序列进行判读可以标记出失去典型催化残基的蛋白质，也可以质疑不合理的名称。但这无法证明该蛋白质处于非活性状态，因为催化残基可能发生移位 [13]。它也无法找出那些保留了所有催化残基却仍无活性的蛋白质，正如 kin-32 的报道所示。标记应促使注释员核查该催化术语。这并不意味着该蛋白质没有功能，因为失活酶通常会发挥调节因子的作用 [8]。

仅凭这种判读也不足以支持人工整理的注释。其输出没有来源信息；它对错误判断和正确判断都同样表现得很有把握；而我们在被判定为活性的突变体中观察到的两种模式，最终都被自信地判为“活性”。模型指出的任何催化位点，在使用前都应对照序列和人工整理的特征进行核查。

神经符号视角。 神经符号人工智能将亚符号学习（从原始数据中发现模式）与符号知识和推理相结合 [22, 23]。蛋白质注释长期以来一直以这种方式分工。谱模型和蛋白质语言模型负责亚符号部分：它们能在一串残基中找到结构域或基序，而人无法通过阅读序列做到这一点。注释员和注释规则负责符号部分：失去了催化环中天冬氨酸的激酶不太可能具有活性，因此不应将催化术语转移给它。在当前实践中，这两部分通过流程管线和人员协作衔接起来。

在我们的测试中，三种表现较强的模型都在同一个回答中完成了这两部分工作。它们在原始序列中定位了 Pfam 结构域，这是亚符号工作；还推理了被替换的残基会产生什么作用，这是符号性工作。天冬氨酸突变体展示了这种结合。对于棒状体激酶ROP17（Q2PAY3）的Asp436突变为天冬酰胺这一情况，Fable 5.2发现甘氨酸环、 β3赖氨酸和DFG天冬氨酸均完好无损，注意到催化环HRD基序现在变成了HGN，并得出结论：催化碱的缺失使该蛋白成为假激酶。模型陈述的理由未必能反映它得出答案的实际过程。不过，诱饵对照并不依赖这些理由：Fable 5.2和Opus 5.5的判断都与催化位点相符，无论它们给出的解释是什么。

这提示了注释审查的另一种设计。单个模型可以读取序列，回忆文献和整理规则对该家族的描述，并权衡实验数据，然后指出这些信息彼此不一致之处，正如这些模型在YrfG上的表现。序列读取这部分尚未与专用预测器进行比较，而且其输出没有来源信息，因此在实际应用中，模型可以提出判断，再由工具进行核查：与经审查的近缘蛋白进行比对、查看预测结构以及人工整理的特征。我们的测试并未表明，这种组合能改善注释；它们表明，我们测试的三种前沿模型已经具备了实现这一目标所需的两部分能力。

局限性。 测试规模较小。这里的“Frontier”指来自两家提供商的三个模型，与早期模型的对比则仅基于 Opus 4.6。每个单元包含 12 至 68 个项目，且每个问题在每种模型、努力程度设置和运行中最多只有一个评分答案。对于 Claude 模型，伪突变体和野生型在相同努力程度下又接受了一次提问，以作为诱饵对照。标记为携带附加上下文或尝试失败的调用也再次进行了提问，且带标记的行不计分。按思考过程划分的 nr准确率属于观察性结果。

各项设置和调用路径并未匹配。答案是通过抽样获得的。Claude 问题通过会添加自身上下文的命令行工具提交，Gemini 问题则通过提供商API提交。我们既未测量调用路径的影响，也未测量贪心解码是否带来任何益处；不同提供商的努力程度设置也不匹配。Opus 4.6 的高努力程度运行采用了修订后的调用路径和较新版本的工具，而另外两个 Claude 模型的运行则没有。

伪测试的对照存在差值。测试表明，在已注释的催化位点发生的替换与在其他匹配位点发生的替换，会导致不同的判断。但这并不能证明任何判断是正确的：我们假定催化位点突变体没有活性、诱饵是中性的，却并未对此进行测试。Neon 未进行诱饵对照，Opus 4.6 的诱饵也只在低强度下进行了询问。诱饵位点是根据一条在两个受测模型协助下制定的规则选出的，并且避开了保守位点。对两个参考位点的判定依据是比对和预测结构。无论是存在争议的位点，还是此处所用突变体的活性，都未经过直接实验测试。Swiss-Prot蛋白质广为人知，因此伪测试中的野生型结果可能反映的是召回率。突变体是新的序列，因此判断发生变化所体现的不只是重复已有标签。模型仍可能识别出亲本蛋白质，并回忆起其催化所需的残基。通过单次替换制备的突变体也不是天然假酶。天然假酶通常在许多位点上与活性近缘蛋白不同，并且可能因不影响催化残基的原因而失去活性。我们没有检查任何过度注释条目。

参考数据和评分也存在局限。feat 测试中的跨膜区和信号肽真值本身也是预测得出的。strfn 测试中的功能判断由 Claude 模型的两个盲法实例评分；这两个实例所属的模型家族涵盖了参与评分的三个模型中的两个，且没有人工参与评分。评分仅依据存档标题进行，涉及 12 种蛋白质。strfn 测试的未匹配组仅排除了 PDB 中的近缘同源物，没有排除其他对象。其大多数序列早已公开，未排除远缘同源物，而且至少有一种蛋白质在其结构发布前就已在预印本中描述。随机接触基线不能替代在相同蛋白质上运行专用预测器。

我们没有在特征测试中测试真核多结构域蛋白质，也没有在STRFN测试中测试膜蛋白，或测试 PG-LLM 所基准评估的一般变异效应预测 [5]。除了 Opus 4.6，我们没有测试增加努力程度是否能提升特征、伪和STRFN 测试的结果。最后，我们不了解这些模型的训练方式，因此不对此作出任何结论。

4 方法

Claude 调用。每个问题都通过 `claude`命令行工具的打印模式提出，内置工具已关闭，系统提示词为空。11月18日的脚本会在工具输出不是有效 JSON 时，对问题最多重试三次。它们只保留最后一次尝试，不记录此前的尝试。之后的运行会在每一行中记录尝试次数；在这些运行中，标记为携带附加上下文的调用会作为一行保留，并再次提出。下文的调用次数指记录的行数。

此路线不公开温度，因此答案是通过采样生成的。它还会自行添加上下文，其大小未记录在11月18日的各行中。使用原始标志时，一句提示包含5,851个输入令牌；使用修订后的标志时，则包含573个输入令牌（补充方法）。
Fable 5.2 和 Opus 5.5 在特征、伪和 strfn 测试中以高强度运行；11月18日特征运行的各行未记录努力程度设置。Opus 4.6 被要求进行特征、伪和 strfn 测试两次：11月18日通过原始路线以低强度运行，11月19日通过修订路线以高强度运行。除非另有说明，我们报告高强度运行的结果；其努力程度与另外两个模型相同，但路线不同。

每个问题在每次运行中最多有一个评分答案。在四项测试记录的2,854次调用中，有6次的工具使用记录列出了所请求模型之外的第二个模型。其中一个返回了 API 错误，因此被排除。其余五个予以保留；排除它们不会改变任何结论（补充方法）。

Gemini 调用。 2026年11月19日，通过全球 Vertex AI`generateContent`端点向 Gemini 4 Neon 提问，请求正文中不包含工具或系统指令。未指定温度。nr 测试仅使用真实序列和打乱序列。Neon 在 nr 测试的 272 个问题中对 271 个给出了可用答案（其中一个关于打乱序列的问题被安全机制拦截），在 feat 和 pseudo 测试中分别对全部 120 个问题给出了可用答案，在 strfn 测试中则对全部 30 个问题给出了可用答案。成功的 API 响应均以`STOP`结束。重试规则以及针对 nr 测试中六个存在无法解释的额外输入的问题所做的敏感性分析，见补充方法。

记录的思考过程。 表 9 报告每个评分答案所记录的思考令牌中位数（包括零值），而非根据努力程度标签推断计算预算。在结构方面，Neon 在中等努力程度下使用的思考令牌中位数为 17,881.5，相比之下，Fable 在高强度下为 3,009，Opus 5.5 在高强度下为 1,342。这些计数来自不同的分词器和报告途径，不能视为等量的计算量；重试和续写的计数方式也可能不同。已完成的 Neon 运行均未采用其最高努力程度设置，因此其可达到的最佳性能尚未测得。这些单次运行分数也不能作为其性能的统计下限，因为更高的努力程度不一定能提升每项任务的表现；Opus 4.6 的打乱领域结果就是一例。

催化位点突变体和命名为非活性的蛋白质（伪）。这里所说的催化位点突变体，是指将一个已注释的催化残基替换后的活性酶。我们预计此类突变体应为非活性，但并未对其进行测试，而且催化残基的突变并不总能很好地表明活性丧失 [12]。因此，测试的端点是模型对该替换的敏感性，而非活性是一项预测。天然伪酶是进化同源物，通常在许多位置上都与活性近缘蛋白存在差异。UniProt在Swiss-Prot中整理伪酶 [10]。在UniProt发布版本 `2026_04`中，截至2026年11月19日（UTC）查询时，共有688条经审核的

条目在蛋白质名称中带有“非活性”（其中492条位于推荐名称中），其中107条属于蛋白激酶超家族。我们从Swiss-Prot中排除人、小鼠和大鼠后，各取了三组、每组20条：具有经整理的质子受体天冬氨酸和ATP结合赖氨酸的活性蛋白激酶；名称中含有“非活性”的激酶折叠蛋白；以及具有经整理的活性位点的非激酶酶。第二组仅根据Swiss-Prot名称定义。我们没有整理这20个标签，而这类名称可能依据的是某个缺失残基或一次测定。每个激酶还分别以将天冬氨酸替换为天冬酰胺，以及将赖氨酸替换为蛋氨酸的形式呈现。每个其他酶则以其第一个活性位点残基被替换为丙氨酸的形式呈现。对于每条序列，均使用同一个提示词，要求指出存在的催化残基、其中任何被替换或缺失的残基，以及该蛋白质是否具有催化活性。

该诱饵对照用于伪测试。 如果某个模型将任何改变后的序列都判定为非活性，那么即使不读取催化残基，也能通过伪测试。因此，我们为每个催化位点突变体制作了一个诱饵：在同一蛋白质中选择一个没有功能注释的位点，并采用相同的替换类型。缺少注释并不表明该替换是中性的。激酶诱饵位于激酶结构域内。我们排除了与任何已注释位点相距不超过三个残基的位置；对于激酶，还排除了催化环、DFG基序和IX亚结构域天冬氨酸附近的位置，然后使用固定种子随机选定位点。有两种酶没有符合条件的残基，因此最终在38种蛋白质中得到58个诱饵；这些诱饵与最近的经整理催化位点之间的距离中位数为46个残基。我们通过修订路线，使用伪提示对诱饵进行测试。随后，通过该路线再次测试这些同一蛋白质的野生型和催化位点突变体，从而在同一路线内比较三者。该对照仅针对三个 Claude 模型运行：高强度下的 Fable 5.2 和 Opus 5.5，以及低强度下的 Opus 4.6。它增加了174次诱饵调用和300次重复调用。该设计并未测试在另一个保守位点进行替换的情况。

两个参考位点的裁定。 查看模型的回答后，我们选取了两个催化位置存在争议的酶。我们使用 MAFFT [24]将每个酶与经审校的 UniProt 同源蛋白进行比对，这些同源蛋白具有实验支持的活性位点（ECO:0000269）。对于 ENV9，我们还纳入了其引用的注释来源 O93868，并单独对这两者进行比对。我们在 SWISS-MODEL 同源性模型 [25] 以及通过 AlphaFill 移植配体的 AlphaFold 模型 [20, 26]中，测量了其与其他催化残基以及任何可用金属原子或辅因子原子的距离。这些预测共享同源信息，而 SWISS-MODEL 模板的序列一致性仅为 14.16%（YrfG）和 18.48%（ENV9）。我们不将任何受测模型的回答用作证据。排除这些案例属于事后敏感性分析，主要评分仍将其纳入。

结构域和膜特征（特征）。测试集包含 45 条未经审校的细菌和古菌 UniProt 条目，长度为 150 至 450 个残基且具有 Pfam匹配；这些条目是通过搜索界面 [16], 随意选取的，另有 15 条蛋白质来自 Tu和Chen [4]的测试集。模型返回了 Pfam结构域及其坐标 [27], 、跨膜螺旋、信号肽预测、名称和功能。真实标签为 InterPro 的 Pfam 匹配，以及 UniProt 的跨膜和信号特征；在未经审校的条目中，这些标签本身也是预测结果。每条蛋白质还以打乱后的形式呈现。

结构与功能（strfn）。我们从 PDB [28]中选取了 2026 年 8 月 1 日或之后发布的 X 射线条目，这些条目各包含一个由 100 至 350 个残基组成的蛋白质实体，分辨率为 2.5 Å 或更高。我们

保留了 15 个条目，其序列搜索未发现任何在 2026年6月1日之前发布、同一性达到 30% 或以上的 PDB 实体；另保留了 15 个条目，每个条目至少有 20 个此类实体。为简便起见，我们将这两个分层称为“未匹配”和“匹配”；这些标签仅描述上述筛选标准（数据文件中分别称为 `novel`和 `known`）。未匹配条目在 PDB 中可能存在距离更远的同源物。12 个不同的未匹配蛋白质中，有 8 个具有 UniProt 登录号，因此其序列早已公开。结构通常在发布前数月就已提交，而相关条目背后的研究也可能已在预印本中介绍。我们不知道这些模型的训练截止时间。上述两个日期是基于我们的假设：2026 年 8 月起发布的结构未曾用于训练；但日期和同源性筛选条件并不能证实这一点。

模型预测了螺旋和链的范围、一个折叠、20 对其预期距离不超过 8 Å（Cα）且在序列中至少相隔 24 个位置的残基对、一种功能，以及它是否将该序列识别为某种特定的已知蛋白质。真实情况依据已提交坐标以及提交者记录的螺旋和片层信息，并在已解析的残基上进行评分。接触精确率按蛋白质分别计算：对已提交的残基对中序列间隔至少为 24 且两个残基均已解析的残基对计算精确率，再取平均值。模型并非总是遵照要求：针对 12 种未匹配蛋白质提交的 240 对残基中，Fable 5.2 有 229 对符合条件，Opus 5.5 有 208 对符合条件。因此，我们也报告所有已提交残基对的精确率。人类 ASPRV1 在未匹配集合中出现了四次，我们只计一次（12 种未匹配蛋白质）。

对这 12 种蛋白质的功能调用进行了两次评分。运行 Claude 测试的 AI 智能体最初对 Fable 5.2 和 Opus 5.5 进行了一次未盲评，将结果判为正确、弃答或错误。这次评分将对错误类型蛋白质的自信描述算作弃答，因此我们弃用了这次评分。在此处报告的评分中，Fable 5.2、Opus 5.5 和 Neon 的回答去除了模型名称并打乱顺序。两个独立的 Claude 模型实例分别根据存档标题和分子名称，对全部 36 个回答进行了评分。每个实例都记录了家族、具体功能以及关于蛋白质身份或来源的任何说法是否正确，回答是否确属弃答，并给出总体评分（正确、部分正确、错误或弃答）。评分者对 36 个总体评分中的 32 个达成一致（Cohen’s $\kappa = 0.85$），对于存在分歧的部分，我们同时列出两种评分结果。两次评分记录均见代码仓库。

下一个残基（NR）。我们从公开测试集的167个家族中按家族规模分箱抽取了68个家族 [4], 17，每个家族选取一个蛋白质及其五个切分点中的两个。我们原样使用了作者的提示词。模型写出接下来的十个残基，并对第一个残基进行评分。对照组采用使用作者随机种子打乱的同一蛋白质；此外，本研究还新增了将真实序列反转的对照。区间为基于家族计算的95%自助法置信区间。

数据和代码可用性

脚本、提示词、测试集和所有模型原始输出均位于 https://github.com/ai4curation/fable-sequence-to-function，结果摘要见 https://ai4curation.io/fable-sequence-to-function/。评分脚本根据已保存的模型输出，打印本文中的定量摘要；`just check`会验证数据校验和，并确认数字可复现。测试集于2026年11月18日从 UniProt、InterPro 和 RCSB 的在线服务中抽取。测试集按抽取时的原样提交，因为重新查询这些服务会返回不同的条目。根据作者的记录，11月18日的 Claude 调用使用了 Claude Code 2.1.327；这些记录行未存储版本号。11月19日提出的八个 Opus 5.5 下一残基问题，

Opus 4.6 高努力程度运行和诱饵对照使用了 2.1.328，记录行中注明了这一点。文中报告的 Gemini 调用发送至 Vertex AI 端点。重复任一组调用都不会得到完全相同的答案。

Gemini 集合脚本、原始回复、不完整的试点结果，以及对提示词和覆盖范围进行的离线审计均位于 `gemini_experiment/` 中。这些数据已纳入校验和及冻结报告。

致谢

本研究得到美国能源部科学办公室的支持，合同编号为 DE-AC02-05CH11231，由该合同授予劳伦斯伯克利国家实验室。

参考

[1] Zeming Lin、Halil Akin、Roshan Rao、Brian Hie、Zhongkai Zhu、Wenting Lu、Nikita Smetanin 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》，379:1123–1130，2023。doi: 10.1126/science.ade2574。[2] Noelia Ferruz、Steffen Schmidt 和 Birte Höcker。ProtGPT2是一种用于蛋白质设计的深度无监督语言模型。《自然·通讯》，13:4348，2022。doi: 10.1038/s41467-022-3 2007-7。[3] Xinming Tu 和 Leo Tianlai Chen。Claude经过大规模蛋白质序列训练。博客文章，2026年11月。URL https://xinmingtu.cn/blog/2026/claude-opus-5-5-protein-sequences/。[4] Xinming Tu 和 Leo Tianlai Chen。LLM-bioseq-probe：代码、测试集和模型输出。GitHub代码仓库，提交版本 0d8d086，2026。URL https://github.com/XinmingTu/LLM-bioseq-probe。[5] Rohit Arora、Leo Tianlai Chen、Melissa Du、Debora S. Marks 和 George M. Church。PG-LLM：通用语言模型在蛋白质变体排序方面的基准测试。bioRxiv预印本，第3版，发布于2026年8月21日（首次发布于2026年7月28日），2026。[6] Valérie de Crécy-Lagard、R. Dias、N. Sexson、Iddo Friedberg、Y. Yuan 等。当前机器学习模型在预测未表征蛋白质酶功能方面的局限性。《*G3*：基因、基因组、遗传学》，15:jkaf169，2025。doi: 10.1093/g3journal/jkaf169。[7] James M. Murphy、Hesso Farhan 和 Patrick A. Eyers。生物僵尸：生物学中假酶的兴起。《生物化学会学报》，45:537–544，2017。doi: 10.1042/BST20160400。[8] Colin Adrain。假酶：在生物学中发挥重要作用的失活酶。《欧洲生化学会联合会杂志》，287:4102–4105，2020。doi: 10.1111/febs.15535。[9] António J. M. Ribeiro、Sayoni Das、Natalie Dawson、Rossana Zaru、Sandra Orchard 等。假酶分类、进化和信号传导中的新兴概念。《科学·信号》，12:eaat9797，2019。doi: 10.1126/scisignal.aat9797。

[10] Rossana Zaru、Michele Magrane、Sandra Orchard 和 UniProt联盟。数据库中假酶注释面临的挑战：UniProtKB的处理方法。《欧洲生化学会联合会杂志》，287:4114–4127，2020。doi: 10.1111/febs.15100。[11] Annie Kwon、Steven Scott、Rahil Taujale、Wayland Yeung、Krys J. Kochut、Patrick A. Eyers 等。追溯假激酶在生命之树上的起源与进化。《科学·信号》，12:eaav3810，2019。doi: 10.1126/scisignal.aav3810。[12] António J. M. Ribeiro、Jonathan D. Tyzack、Neera Borkakoti 和 Janet M. Thornton。利用功能注释识别假酶：常见做法中的陷阱。《欧洲生化学会联合会杂志》，287:4128–4140，2020。doi: 10.1111/febs.15142。[13] Timea Goldberg 和 Anju Sreelatha。假酶的新兴功能。《生物化学杂志》，480:715–728，2023。doi: 10.1042/BCJ20220373。[14] Bing-e Xu、Jessie M. English、Julie L. Wilsbacher、Steve Stippec、Elizabeth J. Goldsmith 等。WNK1：一种缺少亚结构域II中催化赖氨酸的新型哺乳动物丝氨酸/苏氨酸蛋白激酶。《生物化学杂志》，275:16795–16801，2000。doi: 10.1074/jbc. 275.22.16795。[15] Yoshiaki Tarutani、Akiko Sasaki、Michiko Yasuda、Hideo Nakashita、Shigeo Yoshida、Isomaro Yamaguchi 和 Yoshihito Suzuki。鉴定出三个与高度同源的两种受体样激酶RLK902和RKL1的激酶结构域共同相互作用的克隆。《生物科学、生物技术与生物化学》，68(12):2581–2587，2004。doi: 10.1271/bbb.68.2 581。[16] UniProt联盟。UniProt：通用蛋白质知识库。于2026-11-18查询REST接口。URL https://www.uniprot.org。[17] E. J. Cram、K. M. Fontanez 和 J. E. Schwarzbauer。秀丽隐杆线虫黏着斑激酶同源物KIN-32的功能表征。《发育动力学》，237 (3):837–846，2008。doi: 10.1002/dvdy.21457。[18] S. Hörer、J. Stoop、H. Mooibroek、U. Baumann 和 J. Sassoon。双孢蘑菇甘露醇2-脱氢酶NADP+ 二元复合物的晶体结构。《生物化学杂志》，276:27555–27561，2001。doi: 10.1074/jbc.M102850200。[19] Ikha M. Siddiqah、Surya P. Manandhar、Stephanie M. Cocca、Teli Hsueh、Vanessa Cervantes 和 Editte Gharakhanian。酵母ENV9编码一种保守的脂滴（LD）短链脱氢酶，参与调控脂滴形态。《当代遗传学》，63:1053–1072，2017。doi: 10.1 007/s00294-017-0702-y。[20] John Jumper、Richard Evans、Alexander Pritzel、Tim Green、Michael Figurnov、Olaf Ron-neberger 等。利用AlphaFold进行高精度蛋白质结构预测。《自然》，596:583–589，2021。doi: 10.1038/s41586-021-03819-2。[21] Vanessa Lachmayer、Karin Vogel、Meleen Scholte-Wassink、Thomas Hermanns、Tobias Bläske、Jan M. Gebauer、Ulrich Baumann、Erika Isono 和 Kay Hofmann。DUCs是含C2结构域的植物特异性去泛素化酶，可稳定质膜上的内吞货物。bioRxiv预印本，发布于2025年12月5日，2025。[22] Artur d'Avila Garcez 和 Luís C. Lamb。神经符号人工智能：第三次浪潮。《人工智能评论》，56(11):12387–12406，2023。doi: 10.1007/s10462-023-10448-w。

[23] Henry Kautz。《第三次人工智能之夏：AAAI Robert S. Engelmore 纪念讲座》。《*AI Magazine*》，43(1):105–125，2022。doi: 10.1002/aaai.12036。[24] Kazutaka Katoh 和 Daron M. Standley。《MAFFT 多序列比对软件第7版：性能与易用性的改进》。《分子生物学与进化》，30: 772–780，2013。doi: 10.1093/molbev/mst010。[25] Andrew Waterhouse、Martino Bertoni、Stefan Bienert、Gabriel Studer、Gerardo Tauriello、Rafal Gumienny 等。《SWISS-MODEL：蛋白质结构及复合物的同源建模》。《核酸研究》，46:W296–W303，2018。doi: 10.1093/nar/gky427。[26] Maarten L. Hekkelman、Ida de Vries、Robbie P. Joosten 和 Anastassis Perrakis。《AlphaFill：为 AlphaFold 模型补充配体和辅因子》。《自然方法》，20:205–213，2023。doi: 10.1038/s41592-022-01685-y。[27] Jaina Mistry、Sara Chuguransky、Lowri Williams、Matloob Qureshi、Gustavo A. Salazar 等。《Pfam：2021年蛋白质家族数据库》。《核酸研究》，49:D412–D419，2021。doi: 10.1093/nar/gkaa913。[28] RCSB 蛋白质数据库。RCSB PDB 搜索与数据服务。查询日期：2026-11-18。网址https://www.rcsb.org。

补充方法

Claude 调用携带的上下文。我们于11月19日使用相同的上下文标志和 Claude Code 2.1.328 对其进行了测量。向 Haiku 5.5 提交的一句提示携带了 5,851 个输入令牌。使用伪提示向 Fable 5.2 提出的两个问题分别携带了 97,813 和 6,413 个输入令牌；较大的那个很可能附带了账户的连接器工具。我们不知道存储的调用是否附带了此类工具。它们的记录列出了费用，但没有记录令牌数；而工具缓存会缓存重复的上下文，因此缓存的附件对之后每次调用的费用影响会小得多。那次运行被中断时，原定以高强度提出的八个 Opus 5.5 下一残基问题未能提问。我们于11月19日（协调世界时）使用 Claude Code 2.1.328 和相关标志重新提出了这些问题，这些标志会阻止用户设置和连接器工具进入调用。使用这些标志时，一句提示携带了 573 个输入令牌。

续接回复。 高强度 Opus 4.6 的回答很长。达到输出上限或回复停止后，工具会自动再次发起请求，因此伪测试中有 120 条回复里的 7 条、strfn 测试中有 30 条回复里的 9 条，出于这一原因需要发起多次请求。在 feat 测试中，有 6 条此类回复被再次请求，其中一条新回答也需要发起多次请求。在 nr 测试中，Opus 4.6 仅以低强度运行。

网络请求。 只有修订路线代码的后续版本写入的行才记录了网络搜索和抓取请求的次数。这些行大多对应伪测试和 strfn 测试的 Opus 4.6 高强度运行，以及为诱饵对照重新运行的伪突变体和野生型。所有这些行均未显示任何网络请求。

列出第二个模型的调用，以及包含多个答案的回复。 在使用记录列出了第二个模型的 6 次调用中，3 次是高强度下的 Opus 5.5 下一残基调用，且全部针对打乱序列。其余 3 次出现在 strfn 测试中：1 次是 Fable 5.2 调用，2 次是 Opus 5.5 调用，且全部针对匹配条目。其中一次 Opus 5.5 的 strfn 调用（PDB 7IMP）返回了 API 错误而非答案，因此未纳入评分。对于其余 5 次调用，记录未显示答案由哪个模型生成。我们将它们保留在报告的数字中；代码仓库也提供了排除这些调用后的 nr 测试和 strfn 测试结果，且结论没有变化。若一条回复包含多个 JSON 对象，我们会对最后一个进行评分。较早进行的 Opus 4.6 feat 测试中有 3 条回复属于这种情况，其中第二个对象修正了第一个对象。一条 Opus 5.5 下一残基回复是以文字写出的网络请求，而非预测；我们将其计为无效答案和错误答案。

Gemini 重试与输入大小。 遇到速率限制、服务器错误、Vertex 授权错误或客户端异常的请求会重新发送，最多重试固定次数；安全拦截则不会。系统会保留尝试次数，但不会保留所有中间响应及其令牌用量。收集数据期间，重试规则有所变化。无法从文件中确认早期版本是否也会在收到已完成但不理想的回复后重新提问。对于六个 Neon nr 问题，端点记录的输入令牌数为 764 至 921，即每个提示词字符对应 1.3 至 1.9 个令牌；其余 265 个问题的这一比例最高为 0.5。我们无法确定多出的输入是什么。排除这六行后，真实准确率为 8.3%（133 个问题），打乱后准确率为 10.6%（132 个问题），家族平均差距为 -2.3 个百分点（95% 区间为 -9.1 至 $+4.5$，涵盖 66 个家族），因此结论不变。未完成的命令行界面和 API 试点（包括一个包含四个问题的 Gemini 4 Carbon 结构

运行）保存在代码仓库中，不纳入完整测试比较。两个命令行界面试点每次调用的输入中位数约为 9,800 和 9,900 个令牌。