AntigenLM：面向流感的结构感知 DNA 语言建模

Yue Pei[1,4], Xuebin Chi[1,4*], Yu Kang[2,3,4*][1]中国科学院计算机网络信息中心[2]中国科学院北京基因组研究所[3]中国国家生物信息中心[4]中国科学院大学 ypei@cnic.cn, chi@sccas.cn, kangy@big.ac.cn

arXiv:2602.09067v1 [q-bio.GN] 9 Feb 2026

摘要

语言模型推动了序列分析的发展，但 DNA基础模型往往不如针对特定任务的方法，原因尚不明确。我们提出 AntigenLM，这是一种生成式 DNA 语言模型，使用功能单元完整且经过比对的流感基因组进行预训练。这种结构感知预训练使 AntigenLM 能够捕捉进化约束，并在不同任务间实现泛化。针对时间序列血凝素（HA）和神经氨酸酶（NA）序列进行微调后，AntigenLM 能够准确预测不同地区和亚型中未来的抗原变异体，包括训练期间未见过的变异体，表现优于系统发育模型和基于进化的模型。它还实现了近乎完美的亚型分类。消融实验表明，通过片段化或重排破坏基因组结构会严重降低模型性能，凸显了在 DNA语言建模中保持功能单元完整性的重要性。因此，AntigenLM 既为抗原进化预测提供了强大的框架，也为构建符合生物学基础的 DNA基础模型提供了一般性原则。

1 引言

流感病毒迅速进化以逃避宿主免疫，推动季节性流行病的发生，并因此需要频繁更新疫苗（Han 等人，2023）。因此，准确预测病毒进化对于疫苗株选择以及减轻全球公共卫生负担至关重要（Matz & Ellebedy，2025）。当前的疫苗建议综合了大规模基因组监测、抗原特征分析和流行病学监测，并由世界卫生组织（WHO）协调制定（Bucholc 等人，2025）。

传统预测方法依赖系统发育树动态和基于突变的预测模型（Hadfield 等人，2018；Huddleston 等人，2020；Łuksza & L¨assig，2014；Neher 等人，2014）。近年来，深度学习模型（Mehrotra 等人，2025；Lou 等人，2024）已被证明能够准确预测分支水平突变轨迹。*beth-1*显式建模位点级替换动态，与基于系统发育树的方法相比，提升了遗传匹配度，并凸显了机器学习在疫苗指导方面的潜力（Lou 等人，2024）。与此同时，基因组基础模型（例如 DNABERT（Zhou 等人，2024）、NT（Dalla-Torre 等人，2024）、GROVER（Sanabria 等人，2024）、HyenaDNA（Nguyen 等人，2023））和蛋白质基础模型（例如 ESM（Rives 等人，2021）、ProtGPT2（Ferruz 等人，2022）），以及流感专用蛋白质预测器（Hie 等人，2021；Ma 等人，2024；Ito 等人，2025），为抗原序列表示提供了另一种强有力的方法。

然而，病毒进化受到整个基因组中协同相互作用的影响（Cobey，2024年；Gouma 等，2020）。这些相互作用包括 RNA–RNA相互作用和共同包装（Bolte 等，2019；Yang 等，2024）、片段重配的约束（Holmes，2007），以及聚合酶片段（PB1/PB2/PA）与抗原蛋白血凝素和神经氨酸酶之间的共同适应（Vigeveno 等，2024；Noda，2020）。忽略这种结构背景的模型——例如位点级预测器，如

*通讯作者。

beth-1——会割裂生物学信号，限制泛化能力和可解释性。通用基础模型在异质的多物种基因组上进行训练，往往能够捕捉局部序列模式，却常常无法建模这些全基因组范围的高阶约束，因此其性能逊于物种感知且结构对齐的预训练方法（Karollus 等，2024；Benegas 等，2025）。

此外，尽管蛋白质层面预测能够捕捉抗原演化的某些方面，但病毒适应度的关键决定因素——包括同义突变、非编码调控元件、RNA 二级结构、包装信号、密码子使用介导的宿主适应以及聚合酶兼容性——对仅基于蛋白质的模型而言完全不可见。大量实验性证据凸显了这些核苷酸层面机制的重要性（Canale et al., 2018; Kryazhimskiy et al., 2008; Fujii et al., 2005; Liu et al., 2025; Gu et al., 2019），近期基准测试也表明，在相关预测任务中，基于 DNA 的模型优于基于蛋白质的方法（Boshar et al., 2024）。综合这些观察结果，我们认为有必要构建一种针对流感的全基因组核苷酸层面语言模型，以准确预测抗原序列。流感基因组较为紧凑（~13千个核苷酸），也使这种专用模型切实可行。

在此，我们介绍 AntigenLM，这是一种生成式 DNA 语言模型，可在预训练期间显式保留片段顺序和功能单元完整性。随后，我们对该自回归模型进行微调，利用按固定时间窗口收集的连续 HA–NA 序列，预测即将出现的优势毒株的抗原序列；这些序列隐式编码了演化轨迹。通过确保片段方向和抗原边界，AntigenLM 学到的表征能够整合局部序列依赖关系与全局基因组结构，从而实现跨分支、亚型和地理区域的泛化。

我们在多种流感预测任务中评估 AntigenLM，并将其与通用基础模型、世界卫生组织当前的筛选方法，以及最先进的演化预测器（如*beth-1*）进行基准比较。我们还通过仅抗原、截短基因组和片段打乱变体开展消融研究，以分离全基因组上下文和功能单元保留所带来的贡献。

我们的贡献有三方面：1. 功能单元感知预训练：我们提出一种 DNA 语言模型，在预训练期间确保功能单元的完整性及正确排列，并通过受控消融实验展示其优势。2. 改进流感预测：AntigenLM 的表现优于最先进的基于位点的演化模型，在抗原序列预测中实现了更低的氨基酸错配率。3. 可泛化的框架：我们提供了将生物结构融入生成模型的蓝图，这对预测基因组学、疫苗设计以及快速演化病原体建模均具有启示意义。

2 相关工作

2.1 经典进化预测方法

经典系统发育模型构成了当前世界卫生组织疫苗建议的基础。基于局部分支指数（LBI）的树动态方法（Neher 等人，2014；2016）和基于血凝抑制（HI）的抗原预测器（Du 等人，2017）等方法，能够对流行谱系进行排序，并估计它们未来占据主导地位的可能性。这些方法在整合血清学数据和序列数据方面效果显著，但假设位点动态相对均一，且未明确考虑高阶约束或全基因组协同演化。

2.2 基于深度学习的进化模型

近期的深度学习方法直接基于病毒序列对流感进化进行建模。*beth-1*（Lou 等，2024年）从基因组序列和人群血清阳性率推断位点级突变适应度，从而能够定量预测分支特异性适应度景观。尽管其性能出色，但该方法将突变视为相互独立的事件，限制了其捕捉跨越血凝素、神经氨酸酶及其他片段的协同变化的能力。EVE（Frazer 等，2021）是一种基于 H1N1 血凝素序列训练的深度生成模型，可预测分支水平突变轨迹（Mehrotra 等，2025），但其仅关注单一抗原，无法顺畅地推广到其他亚型抗原。

2.3 基因组语言模型

DNABERT（Ji 等，2021）、NT（Dalla-Torre 等，2024）和 HyenaDNA（Nguyen 等，2023）等核苷酸级语言模型能够学习 DNA序列的上下文表示。长上下文建模方面的进展——例如基因组规模 Transformer（Fishman 等，2025）和扩展上下文的 Hyena架构——使模型能够处理最长达兆碱基长度的序列，足以覆盖病毒基因组。然而，这些通用模型是在异质的多生物基因组语料库上训练的，而这些语料库中的基因组大小和结构差异极大，因此难以保留生物特异性结构组织或全基因组高阶约束。

2.4 通用及流感专用蛋白质语言模型

蛋白质语言模型，包括 ESM 等通用模型（Rives 等，2021；Lin 等，2023）和流感专用模型（Hie 等，2021；Ma 等，2024；Ito 等，2025），能够提供强大的抗原序列表征，并已证明其在抗原特征分析中的效用。然而，仅使用蛋白质的模型无法捕捉关键的核苷酸级进化机制，例如同义替换、密码子使用适应或非编码调控元件；这些机制会影响病毒适应度，却不会改变蛋白质序列。此外，除 ProtGPT2 （ Ferruz 等，2022 ）外，大多数蛋白质语言模型都不是自回归模型，因此无法生成全长抗原预测。

3 方法

3.1 模型概述

我们提出了 AntigenLM，这是一种基于 Transformer（Vaswani et al., 2017）的框架，专为甲型流感病毒基因组建模而设计（**图 1B**）。该模型源自 GPT-2（Radford et al., 2019）架构，但经过重新设计，以应对生物序列学习所面临的独特挑战：(i) 极长的输入上下文（每个基因组最多含 13k 个核苷酸（Van den Hoecke et al., 2015））；(ii) 结合生成任务与判别任务的多任务学习目标；(iii) 捕捉不同功能基因片段之间依赖关系的需求（Neverov et al., 2015）。骨干网络是一个仅解码器 Transformer，包含 6 层、384 个隐藏维度和 6 个注意力头。每个模块都采用内部维度为 1,536（即 4× 模型维度）且使用 GELU 激活函数的前馈子层，与 GPT-2 的实现一致。尽管与大型 NLP 模型相比，该架构较为紧凑，但其位置嵌入范围扩展至 13,000 个位置，从而无需截断即可进行全基因组建模，并避免使用 BPE 和 k-mer（Li et al., 2024）等手工设计的分词器。该设计兼顾覆盖范围与计算效率，使模型能够利用标准的多 GPU 环境，在数万个病毒基因组上进行训练。

在骨干网络之上，我们采用双头设计进行多任务学习：

1. 语言建模（LM）头：与嵌入矩阵绑定，以自回归方式预测下一个核苷酸标记，从而捕捉进化动态。2. 分类头：将哨兵位置处的隐藏状态映射为亚型对数几率，以支持监督式序列级判别。

该模型利用共享预训练，并在微调期间独立优化各项任务。这使模型能够通过自回归建模捕捉全基因组上下文，同时受益于监督式亚型分类，从而获得更丰富的生成式表征，并提升亚型识别性能。

3.2 功能单元编码

AntigenLM 采用**两阶段功能单元编码** 策略，同时保留全基因组上下文和片段级结构。在**预训练**期间，甲型流感病毒的全部八个片段（PB2、PB1、酸性聚合酶蛋白、血凝素、核蛋白、神经氨酸酶、基质蛋白、非结构蛋白）被拼接为一条连续的全基因组序列（Hoffmann 等，2000），并保持固定顺序和近似的位置对齐。这使 Transformer 能够捕捉长程共进化依赖关系，例如血凝素与神经氨酸酶之间的补偿性突变（Koel 等，2013），而无需人工标记或截断，并且

避免了繁重的结构标注（Huddleston 等，2020）。在**微调**期间，哨兵标记（例如，<血凝素>、 <神经氨酸酶>）会明确界定功能区域（Johnson 等，2017），引导亚型分类所需的注意力，并将自回归解码限制在各个抗原内部，以防止跨片段续写（Sennrich 等，2015）。这种将可扩展的无监督预训练与针对特定任务的结构监督相结合的方法，能够生成符合生物学特征且适应任务的表征。

3.3 复杂度与效率

可扩展性对基因组规模建模至关重要。因此，AntigenLM 兼具紧凑性和长上下文处理能力。其骨干网络仅使用 6 层、384 个隐藏维度和 6 个注意力头，却支持长达 13,000 个位置的序列，从而无需截断即可进行全基因组建模。生成任务和判别任务共用这一 Transformer 骨干网络，可降低内存使用量并促进表征的高效迁移（Avsec 等，2021）。实际应用中，在超过 54,000 个完整甲型流感病毒基因组上进行预训练，随后针对特定任务进行微调，都可在标准多 GPU 环境下完成（训练详情见第 4.5 节）。这些设计使 AntigenLM 适用于大规模流感监测数据，同时保持计算开销可控。

3.4 用于预测和分类的微调

我们使用相同的 GPT-2 风格骨干网络和不同的任务头，在两个下游任务上对 AntigenLM 进行微调。对于生成式预测，我们在提示上训练语言建模头；这些提示由三个历史 HA/NA 区块串接而成，后接一个未来区块。每个示例的结构如下：

$$\underbrace{\text{block}^{(1)}}_{\text{past}}\ \underbrace{\text{block}^{(2)}}_{\text{past}}\ \underbrace{\text{block}^{(3)}}_{\text{past}}\ \underbrace{\text{block}^{(\star)}}_{\text{future}},$$

其中

$$\text{block}^{(i)} = \texttt{<subtype>}\ \texttt{<HA>}\ \text{HA}^{(i)}\ \texttt{<NA>}\ \text{NA}^{(i)}\ \texttt{<sep>}$$

对于 $i = 1$，2，3。此处，$\text{HA}^{(i)}$ 和 $\text{NA}^{(i)}$ 是同一季节内三个过去时间点的 HA/NA 核苷酸序列。未来区块 $\text{block}^{(*)}$ 遵循相同的模式，但使用待预测未来时间点的 HA，NA，并省略开头的 <亚型> 标记。我们在完整的 token 序列上优化标准因果语言建模损失（不包括填充部分）；在推理时，仅输入三个历史区块，并从目标 <血凝素> 标记开始，以自回归方式生成未来区块，直到生成 <sep> 或达到最大长度。

对于亚型分类，我们使用相同的骨干网络和分类头，并采用更简单的输入，其中包含一对 HA/NA 和一个分隔符，

$$\texttt{<HA>}\ \text{HA}\ \texttt{<NA>}\ \text{NA}\ \texttt{<sep>}.$$

序列由 Transformer 进行编码，我们将最终词元位置的隐藏状态作为 HA/NA 对的紧凑表示。该表示经过线性层处理后生成亚型对数几率，并通过交叉熵损失进行训练。实际上，我们会对两个特定任务模型进行微调——一个是生成式模型，另一个是判别式模型——二者均从同一个预训练 AntigenLM 骨干网络初始化；此外，我们还支持通过对语言建模损失和分类损失进行简单加权求和，联合优化这两个目标。

4 实验

4.1 数据集

我们从 GISAID（Shu & Mc- Cauley, 2017）汇集了一个全面的甲型流感病毒基因组语料库，其中包括两个主要亚型 A/H3N2 和 A/H1N1，以及 10 种次要亚型（例如 H5N1、H7N9）。在预训练和微调中，我们均使用 2022年2月之前采集的序列，并通过去除重复序列、排除不完整基因组（少于八个片段），以及筛除相对于相应亚型参考序列平均片段覆盖度低于 $< 90\%$ 的序列，实施严格的质量控制。筛选后，数据集包含 32,758 个 H3N2、20,680 个 H1N1 和 1,074 个次要亚型基因组，总计 $\sim$600 百万个核苷酸，涵盖

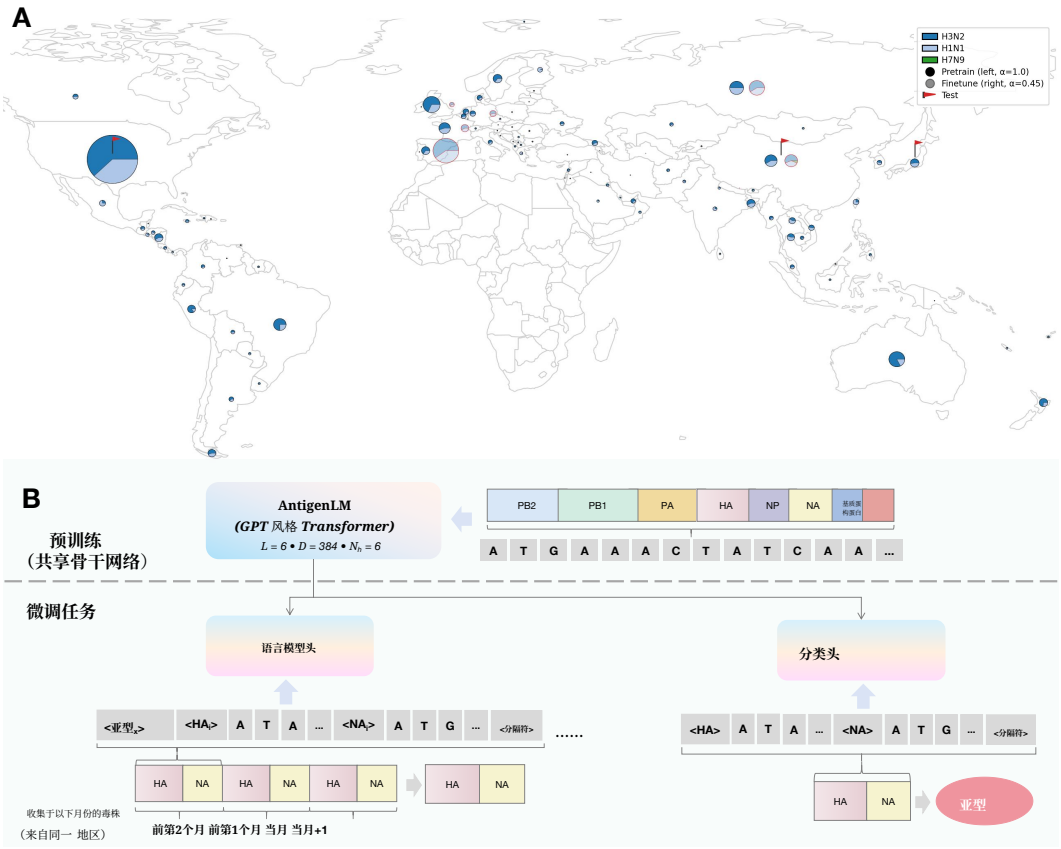

图 1：数据分布与 AntigenLM 架构。(A) 用于预训练、微调和测试的甲型流感病毒序列的全球分布。圆圈大小表示样本数；饼图扇区表示亚型组成。深色圆圈表示预训练数据，浅色圆圈表示微调数据，红色标记表示测试区域。(B) AntigenLM 架构。预训练和微调阶段的示意图。该模型采用 GPT 风格 Transformer 作为共享骨干网络（6 层，隐藏维度为 384，包含 6 个注意力头）。**上方（预训练）：** 骨干网络使用涵盖全部八个流感基因片段（PB2 至 NS）的核苷酸序列进行预训练。**下方（微调）：** 模型针对两项不同任务进行微调：病毒进化预测（左），使用语言模型头，根据同一地区的历史毒株（月份 $t-2$ 至 t）预测月份 $t+1$ 的序列；以及亚型分类（右），使用分类头，根据血凝素和神经氨酸酶片段识别病毒亚型。

54,512个基因组（**图 1A**）。对于每种病毒，我们按照固定顺序（片段从最长到最短）直接串联八个片段，形成全基因组序列，不进行多序列比对，也不添加间隙，然后进行分词（附录 A.2）。

2022 年之后的序列经过相同的质量控制流程，并按亚型和采集月份分层，留作回顾性测试集，以实现模拟真实世界预测的无数据泄漏评估。为评估地理泛化能力，我们使用 2022 年之前来自欧洲和亚洲的序列微调模型，并在 2022 年之后来自日本（分布内）和美国（分布外）的基因组上评估其性能。

4.2 预训练设置

为剖析跨片段依赖关系、核苷酸水平变异、基因组完整性和片段顺序对齐的影响，我们采用五种预训练变体开展了受控消融研究。每种变体的序列与片段组织方式各不相同；这些配置的示意图见**图 2**：

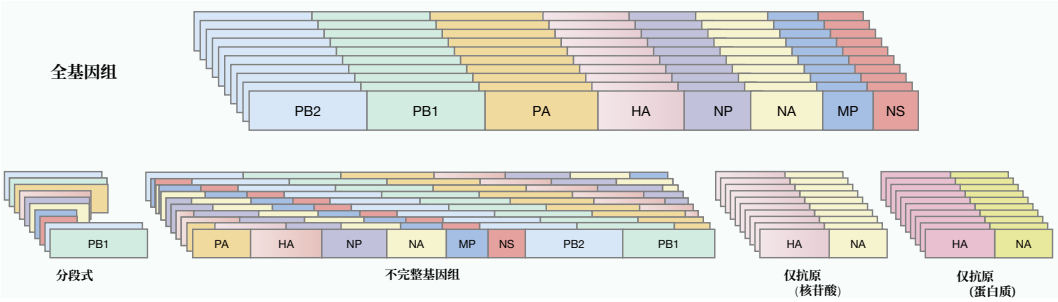

图 2：**AntigenLM 的预训练输入及消融变体**。标准的**全基因组**策略（上）按照固定顺序串接来自同一分离株的八个片段。**消融变体**（下）探索了其他输入格式：**分段式**使用彼此独立的分段序列；**不完整基因组**通过从长串接序列中随机裁剪固定长度的时间窗口生成，因此包含混合片段；而**仅抗原**模型仅使用 HA 和 NA 片段的核苷酸序列和蛋白质（氨基酸）序列作为输入。除**仅抗原（蛋白质）**外，所有配置均使用核苷酸序列。

1. **全基因组预训练：**将全部八个基因片段（PB2、PB1、酸性聚合酶蛋白、血凝素、核蛋白、神经氨酸酶、基质蛋白、非结构蛋白）连接成单条核苷酸序列（最多 $\sim$13,000 个词元，每个词元对应一个核苷酸），并采用自回归目标进行训练。这种方式保留了片段顺序和跨片段依赖关系，使模型能够学习基因组范围内的协同进化模式。2. **不完整基因组预训练：**将连接后的基因组随机裁剪成 12k 词元的时间窗口，同时保持语料库总规模不变。这会破坏部分基因边界，使我们能够量化失去完整功能上下文的影响。3. **分段式预训练：**每个训练样本都包含一个随机选取的基因片段。语料库规模与全基因组变体相同，但移除了所有跨片段上下文，从而单独考察片段内学习。4. **仅抗原（核苷酸）预训练：**仅纳入血凝素和神经氨酸酶片段。将血凝素和神经氨酸酶的核苷酸序列连接起来，并采用相同的自回归目标进行训练。这将模型限定于两种主要抗原决定簇，同时保留它们之间的协同进化信号。5. **仅抗原（蛋白质）预训练：**将血凝素和神经氨酸酶编码序列翻译为氨基酸序列，连接后按残基级别进行分词。直接使用血凝素和神经氨酸酶蛋白进行训练，该变体仅关注蛋白质层面的抗原变异，从而分离出同义核苷酸变化所带来的影响。

所有变体均使用相同的骨干网络、优化器、学习率调度策略、更新次数和有效词元预算；数据划分、初始化和重排均采用固定的随机种子，以确保比较条件受控。我们优化标准的因果语言建模损失

$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T-1} \log p(x_{t+1} \mid x_{\le t}),$$

其中 T 为序列长度。

4.3 微调与评估任务

我们设计了四项互补的下游任务，用于评估 AntigenLM 学习具有生物学意义的表征的能力，并据此对模型进行微调。这些任务评估短期和长期预测、地理区域与亚型泛化能力，以及用于分类的表征质量。

（1）下个月血凝素/神经氨酸酶序列预测。为评估短期预测能力，我们使用截至 t月份的连续三个月的血凝素和神经氨酸酶序列对 Anti-genLM 进行微调

并用于预测指定地区 $t+1$ 月份的序列。哨兵标记 <HA>和<NA>用于标记输入片段， <sep>用于分隔不同毒株的序列。我们检查输出中是否包含 <HA>和 <NA>，并将长度偏差超过 150 个核苷酸的预测标记为失败。对成功的输出进行以下评估：（i）**平均氨基酸错配数——**预测序列与最近观测到的毒株之间的平均氨基酸差异；（ii）**表位特异性错配——**仅针对已知表位位点计算的平均错配数；以及（iii）**词元级困惑度**。基线模型和消融模型同时进行训练和评估。

(2) 下一季 HA/NA 序列预测。 对于更长时间跨度的预测，AntigenLM 根据上一季（T 年 11 月至 $T+1$ 年 2 月）的连续序列，预测了 $T+1$ 季的序列。包括 *beth-1* 在内的基线方法，均使用相同的 2022 年前训练和 2022 年后评估数据划分进行训练和评估。

(3) 地理与跨亚型泛化。 为评估分布外性能，微调时留出所有美国序列，并仅将其用于评估。同时还测试了模型向次要亚型 H7N9 的迁移能力（占预训练语料库的 $< 5\%$，占微调数据的 $< 1\%$），结果表明 AntigenLM 能够泛化到罕见亚型。由于没有基线模型能够预测次要亚型序列，因此将其与 AntigenLM 的预训练变体进行比较。

(4) 亚型分类。 通过在 HA+NA 序列的哨兵标记嵌入上训练轻量级分类头，进一步评估了表征质量。性能采用微平均 F1 分数进行衡量。

4.4 基线比较

我们基准 测试了 AntigenLM，对照三类基 线:

（1）进化模型 — *beth-1*。 *beth-1* 是一种最先进的基于位点的动态模型，用于估计突变适应度，并将位点流行率向未来时间投影。它是我们的主要生物学基线。EVE 被排除在外，因为它产生的是分支层面而非全序列的预测。

（2）基于树的预测器 — LBI。 局部分支指数（LBI）基于由 HA/NA 序列构建的系统发育树运行，历史上为世界卫生组织疫苗株推荐提供了依据，从而为预测性能提供了有意义的下限。

(3) 通用 DNA/蛋白质语言模型。我们使用自回归模型 HyenaDNA 和 ProtGPT2，将 AntigenLM 与通用核苷酸和蛋白质语言模型进行基准测试。这些模型是在涵盖多个物种和不同基因组大小的大规模异质基因组或蛋白质组语料库上预训练的，使我们能够评估流感特异性、生物学信息引导的预训练所带来的额外价值。
所有语言模型均使用相同的训练/评估数据集，并在条件允许时匹配参数数量。*beth-1* 和 LBI 均使用公开可用的实现，并采用推荐设置运行。

4.5 实现

AntigenLM 使用 PyTorch 和 Hugging Face transformers库实现。除非另有说明，本节中的效率统计数据均指预训练阶段。
我们使用 8 块 NVIDIA A800 GPU（每块 80GB）进行预训练，每个设备的微批次大小为 1 个完整基因组，并在 4 个微步上进行梯度累积，使每次优化器更新的有效全局批次大小达到 32 个基因组，上下文长度上限为 13,000 个词元。在此配置下，预训练任务在全部 8 块 GPU 上每秒处理约 7×10^5 个词元；训练完整的 5.4 万基因组语料库所需的总计算量约为 10^{18} 浮点运算，同时每个设备的显存占用仍远低于 80 GB。
我们使用 AdamW 优化所有模型（峰值学习率为 1×10^{-4}，前 5% 的更新采用线性预热，随后进行余弦衰减，dropout 为 0.1，梯度裁剪阈值为 1.0）。针对特定任务的微调

用于预测和亚型分类的微调复用相同的预训练骨干网络，成本远低于预训练；因此，我们不单独报告这些阶段的效率数据。代码、预处理脚本和训练好的检查点将在论文录用后发布。

5 结果与分析

5.1 下个月 HA/NA 序列预测

我们首先评估了 AntigenLM 对下个月血凝素和神经氨酸酶序列进行预测的能力，以考察其捕捉短期进化动态的表现（**图3A**）。我们使用 2022 年以前的数据对 AntigenLM 进行微调，利用前面三个月（t-2、t-1、t）的序列预测月份 t+1 的 HA/NA 序列，并在 2022 年之后的观测数据上进行评估。模型生成的逐月预测平滑，血凝素的平均氨基酸（AA）错配数为 3–4 个（占 566 个氨基酸的 $< 1\%$），神经氨酸酶为 1–2 个（占 469 个氨基酸的 $< 0.5\%$）。表位区域内的错配同样罕见，且神经氨酸酶的错配数接近于零。

与通用语言模型和消融对照的比较凸显了生物学信息引导的预训练的优势。**(i)** HyenaDNA 和 ProtGPT2 生成的结果大多是有效的 HA/NA 序列，但长度偏差较大，且哨兵标记缺失或位置错误，因此无法计算错配数；**(ii)** 不完整基因组基线经常无法生成有效序列，且在成功生成时，AA 错配率显著更高；**(iii)** 分片段模型和仅抗原模型的表现与 AntigenLM 相当，但略有下降。总之，这些结果表明，保持各个片段（至少包括 HA 和 NA）的完整性对于生成有效预测至关重要；此外，纳入全基因组的跨片段上下文还能进一步提高预测准确性。
我们进一步使用词元级困惑度量化了所有消融模型的不确定性，但由于词元化方式不同，未包括仅抗原（蛋白质）变体。AntigenLM 的困惑度为 1.26，显著低于不完整基因组（3.55）、分段式（4.42）和仅抗原（核苷酸）（4.56），表明其对条件序列分布的建模能力更优。这些结果表明：（i）非 HA/NA 内部片段提供了有意义的预测信号（由仅抗原模型进行控制）；（ii）保留功能单元结构有助于提升语言模型的学习效果（由分段式模型进行控制）；（iii）序列完整性对于准确预测至关重要（由不完整基因组模型进行控制）。

5.2 预测下一季流行毒株

我们评估了 AntigenLM 对下一季优势毒株的预测能力，这与疫苗株选择直接相关。模型使用 2022 年之前来自欧洲和亚洲的序列进行微调，并在 2022 年之后来自日本的序列上进行测试，根据季T+1连续三个月的序列，预测T季的流行毒株。

图 3B展示了 H1N1 和 H3N2 的结果，涵盖全长基因和表位区域。AntigenLM 始终实现了最低的氨基酸错配数，在 H1N1-HA 和 H3N2-NA 上的提升最为显著：与世界卫生组织的疫苗建议（在**图 3**中标记为“当前系统”）相比，错配数减少了 70% 以上；与基于位点的模型*beth-1*相比，则减少了 50%。基于位点的模型表现出更高的方差，而且往往会过度预测单个位点扫荡，这反映了其独立性假设的局限性。无论是全长区域还是表位限制区域，改进都始终如一，H1N1-NA 除外，因为其错配数原本已低于 1，提升空间有限。这些结果凸显了 AntigenLM 在预测准确性方面的整体提升，也表明其有望为疫苗设计提供指导。
预测结果在季节间也更为稳定，避免了基于系统发育树的方法中观察到的突然进化枝转换。在日本进行的100次测试（每个亚型50次）中，H3N2毒株在下一季节转换到了新的进化枝，而H1N1大体上仍停留在现有进化枝内。AntigenLM准确预测了新出现的H3N2进化枝的序列，这与已知的抗原漂移模式一致。这些结果表明，AntigenLM能够提供生物学上连贯且可操作的预测，从而降低疫苗与未来流行毒株之间潜在的不匹配。

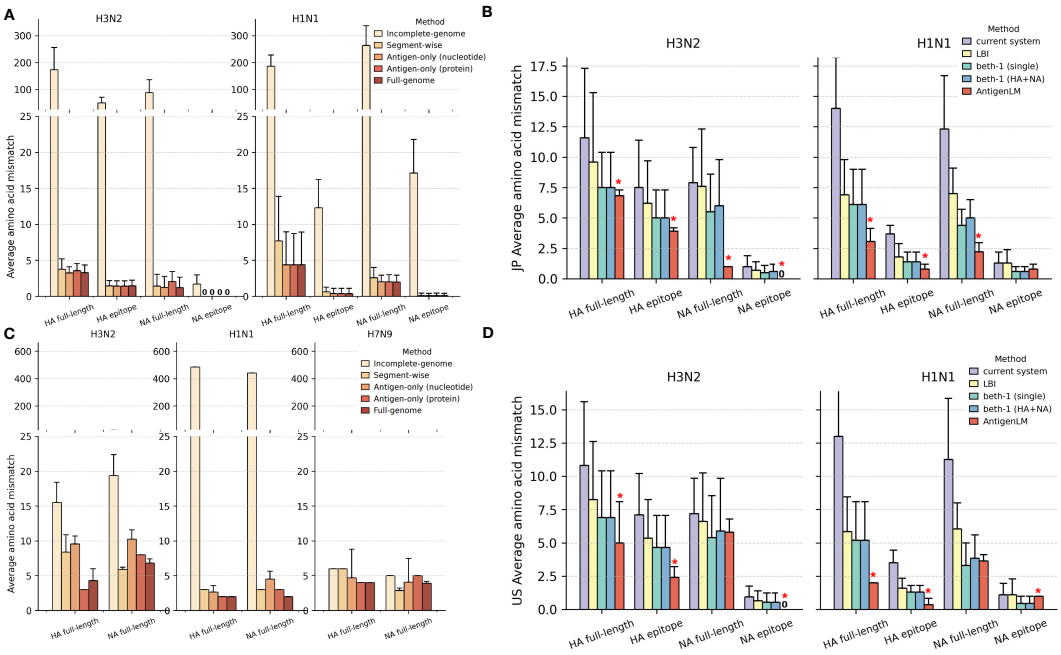

图 3：**AntigenLM 在所有预测任务中实现了最低的氨基酸错配**。（A）下月预测：将 AntigenLM（全基因组预训练）与消融对照进行比较。（B）基于 2022 年后日本数据的下一季预测（2022 年前的数据纳入微调）：将 AntigenLM 与基线模型进行比较。（C）下一季预测中的跨亚型泛化：将 AntigenLM（全基因组预训练）与用于 H7N9 预测的消融对照进行比较。（D）地理泛化：在微调期间未见过的美国数据上评估 AntigenLM，并与基线模型进行比较。星号表示 AntigenLM 与*beth-1*之间的统计显著性（t 检验）：* $p < 10^{-3}$。误差线表示标准差。

5.3 跨亚型与地理区域的泛化

当前的进化模型通常需要大量训练数据，并且往往难以预测次要亚型或历史采样有限的地区中新出现的毒株序列。因此，它们在预测采样不足的人群中新出现毒株的进化或大流行潜力方面效用有限。为评估AntigenLM超越其训练分布进行泛化的能力，我们开展了两项互补的分布外（OOD）实验。

跨亚型迁移。我们评估了 AntigenLM 在次要流感 A 亚型 H7N9 上的表现。该亚型仅占预训练语料库的 4.68%，占微调集的 0.3%（48 条序列）。尽管其代表性极为有限，AntigenLM 对 H7N9 生成的下一季预测，其测试数量和错配率仍与主要亚型 H1N1 和 H3N2 的结果相当。有趣的是，所有消融模型——包括 不完整基因组变体——在 H7N9 上的表现与在主要亚型上的表现相近，这可能是由于 H7N9 的序列保守性较高（图 3C）。总体而言，AntigenLM 实现了有效的跨亚型泛化，而这一能力仍是当前系统发育方法面临的挑战。

地理泛化。为评估地理稳健性，我们将美国作为未见地区留出，并仅使用来自欧洲和亚洲的序列对AntigenLM进行微调。该模型在H1N1和H3N2的血凝素上均保持了显著低于*beth-1* 的平均氨基酸错配，尽管神经氨酸酶的改进并不显著（**图3D**）。对100个测试案例的分析表明，所有案例均涉及向微调中未出现的进化枝转变。AntigenLM在其中90个案例中正确预测了进化枝变化（全部50个H1N1以及40/50个H3N2），不过缺乏暴露略微增加了总体错配，尤其是在神经氨酸酶中。神经氨酸酶的表位错配仍保持在极低水平（0–1个氨基酸），与分布内表现一致，并足以支持疫苗设计。这些结果表明，AntigenLM学习的是全球进化约束，而非记忆地区特异性突变，从而能够在历史上采样不足的人群中实现可靠预测。

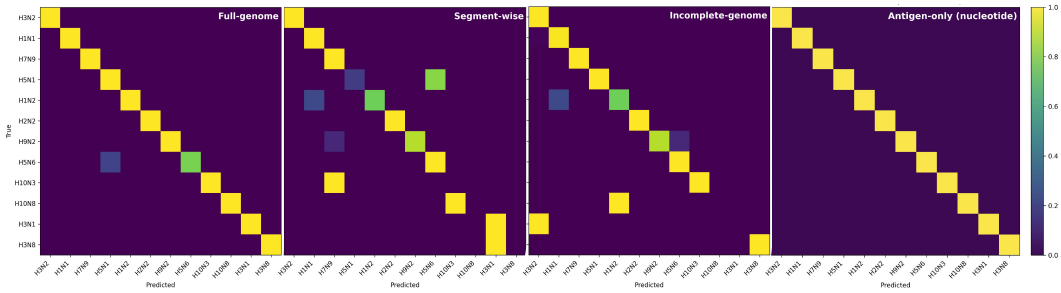

图4：AntigenLM 与消融模型的亚型分类性能。用于亚型分类的行归一化混淆矩阵。AntigenLM（左）和仅抗原（核苷酸）（右）显示出近乎完美的对角线主导，表明分类高度准确；而不完整基因组和分片模型（中）则显示出更多的非对角线误分类，尤其是对于罕见亚型。

综合来看，这些结果表明，AntigenLM并非仅仅拟合占主导地位的亚型，而是学习了可迁移的、具有生物学意义的表征，并能在不同亚型和地理区域之间泛化——这是当前基线方法难以实现的能力。这种稳健性对于现实应用至关重要，因为数据可得性并不均衡，新毒株也可能出现在采样充分的种群之外。

5.3.1 亚型分类

为进一步评估 AntigenLM 的多任务能力，我们将其作为一项独立的下游任务，测试了其在亚型分类上的性能。

亚型分类相对直接，因为各亚型之间存在明显的序列差异，因此 AntigenLM 及其所有变体均表现良好。仅抗原模型（核苷酸和蛋白质）达到了 100% 的准确率，这符合预期，因为亚型完全由抗原序列决定。值得注意的是，全基因组 AntigenLM 同样表现优异，微平均 F1 分数达到 99.81%；在 530 株测试毒株中，仅有一株 H5N6 毒株被误分类为 H5N1。尽管训练数据有限，所有其他次要亚型均被准确分类，表明 AntigenLM 的嵌入空间能够稳健地区分各亚型。相比之下，不完整基因组和分片变体表现出明显更多的亚型混淆（**图 4**）。

这些结果表明，AntigenLM 学到了结构良好且具备亚型感知能力的潜在空间，可支持自动化流感监测和实时毒株追踪等应用，且无需重新训练。

6 讨论与结论

AntigenLM 将功能单元保留融入语言建模，显著提升了流感预测的表现。在下个月、下一流行季和分布外任务中，它始终优于通用语言模型、进化方法和消融对照，实现了更少的氨基酸错配，并生成与亚型转换相吻合、符合生物学规律的预测。通过捕捉不同片段之间的高阶依赖关系，AntigenLM 能够泛化至不同亚型和地理区域，即使在采样不足的人群中也能作出准确预测。实时部署仍面临挑战，因为预测本质上具有概率性，应作为专家决策的补充（Ampofo 等，2011）。总体而言，AntigenLM 表明，将生物结构融入基因组语言模型，能够为病毒进化研究和公共卫生应用提供更准确、可解释且可用于行动的洞见。

数据与代码可用性

本研究使用的所有流感基因组序列均可从 GISAID 流感病毒资源公开获取。代码、数据和模型可在 AntigenLM 获取。

大语言模型的使用

我们仅使用大语言模型进行了少量语言润色。代码编写、数据处理、实验设计或分析均未使用大语言模型；所有报告的结果均由我们自行编写的代码生成，并根据运行日志进行了验证。

致谢

本研究获得国家重点研发计划（2024YFC3405704）和国家自然科学基金（32371537）的资助。

参考文献

William K. Ampofo、Norman Baylor、Sarah Cobey、Nancy J. Cox、Sharon Daves、Steven Edwards、Neil Ferguson、Gary Grohmann、Alan Hay、Jacqueline Katz、Kornnika Kullabutr、Linda Lam- bert、Roland Levandowski、A. C. Mishra、Arnold Monto、Marilda Siqueira、Masato Tashiro、An- thony L. Waddell、Niteen Wairagkar、John Wood、Maria Zambon 和 Wenqing Zhang。改进流感疫苗病毒选择：世界卫生组织非正式磋商报告，会议于2010年6月14日至16日在瑞士日内瓦世界卫生组织总部举行。《流感与其他呼吸道病毒》，6 (2):142–152，2011年8月。ISSN 1750-2659。doi: 10.1111/j.1750-2659.2011.00277.x。URL http://dx.doi.org/10.1111/j.1750-2659.2011.00277.x。

ˇZiga Avsec、Vikram Agarwal、Daniel Visentin 等人。通过整合长程相互作用，根据序列有效预测基因表达。《自然·方法》，18:1196–1203，2021年。doi: 10.1038/s41592-021-01252-x。

Gonzalo Benegas、Chengzhong Ye、Carlos Albors、Jianan Canal Li 和 Yun S. Song。基因组语言模型：机遇与挑战。遗传学趋势，41(4):286–302，2025年4月。ISSN 0168-9525。doi: 10.1016/j.tig.2024.11.013。URL http://dx.doi.org/10.1016/ j.tig.2024.11.013。

Hardin Bolte、Miruna E. Rosu、Elena Hagelauer、Adolfo Garc´ıa-Sastre 和 Martin Schwemmle。流感病毒基因组的包装受核糖核酸和核蛋白介导的相互作用构成的可塑性网络调控。病毒学杂志，93(4)，2019年2月。ISSN 1098-5514。doi: 10.1128/jvi.01861-18。URL http://dx.doi.org/10.1128/jvi.01861-18。

Sam Boshar、Evan Trop、Bernardo P de Almeida、Liviu Copoiu 和 Thomas Pierrot。基因组语言模型是否足以满足所有需求？探索基因组语言模型在蛋白质下游任务中的应用。*Bioinformatics*，40(9)，2024年8月。ISSN 1367-4811。doi: 10.1093/bioinformatics/btae529。URL http://dx.doi.org/10.1093/bioinformatics/btae529。

Magda Bucholc、Mark G. O'Doherty 和 Declan T. Bradley。2023/2024流感季的流感疫苗有效性：一项针对北爱尔兰因呼吸系统疾病急诊住院患者开展的检测阴性病例对照研究。《流感与其他呼吸道病毒》，19(9)，2025年9月。ISSN 1750-2659。doi: 10.1111/irv.70149。URL http://dx.doi.org/ 10.1111/irv.70149。

Aneth S. Canale、Sergey V. Venev、Troy W. Whitfield、Daniel R. Caffrey、Wayne A. Marasco、Celia A. Schiffer、Timothy F. Kowalik、Jeffrey D. Jensen、Robert W. Finberg、Konstantin B. Zel- dovich、Jennifer P. Wang 和 Daniel N.A. Bolon。甲型流感病毒血凝素基因起始处的同义突变会影响实验适应度。分子生物学杂志，430(8):1098–1115，2018年4月。ISSN 0022-2836。doi: 10.1016/j.jmb.2018.02.009。URL http://dx.doi.org/10.1016/j.jmb.2018.02.009。

Sarah Cobey。针对快速演化的病原体进行疫苗接种与记忆的纠缠。《自然·免疫学》，25(11):2015–2023，2024年十月。ISSN 1529-2916。doi: 10.1038/ s41590-024-01970-2。URL http://dx.doi.org/10.1038/s41590-024-01970-2。

Hugo Dalla-Torre、Liam Gonzalez、Javier Mendoza-Revilla、Nicolas Lopez Carranza、Adam Henryk Grzywaczewski、Francesco Oteri、Christian Dallago、Evan Trop、Bernardo P. de Almeida、Has- san Sirelkhatim、Guillaume Richard、Marcin Skwark、Karim Beguir、Marie Lopez 和 Thomas Pierrot。核苷酸转换器：构建和评估用于人类基因组学的稳健基础模型。《自然·方法》，22(2):287–297，2024年11月。ISSN 1548-7105。doi: 10.1038/ s41592-024-02523-z。URL http://dx.doi.org/10.1038/s41592-024-02523-z。

Xiangjun Du、Aaron A. King、Robert J. Woods 和 Mercedes Pascual。结合进化信息预测季节性甲型流感（H3N2）。《科学转化医学》，9(413):eaan5325，2017年。doi: 10.1126/scitranslmed.aan5325。

Noelia Ferruz、Steffen Schmidt 和 Birte H¨ocker。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(1)，2022 年 7 月。ISSN 2041-1723。doi：10.1038/s41467-022-32007-7。URL：http://dx.doi.org/10.1038/s41467-022-32007-7。

Veniamin Fishman、Yuri Kuratov、Aleksei Shmelev、Maxim Petrov、Dmitry Penzar、Denis Shepelin、Nikolay Chekanov、Olga Kardymon 和 Mikhail Burtsev。Gena-LM：面向长序列的一系列开源基础 DNA 语言模型。《核酸研究》，53(2)，2025 年 1 月。ISSN 1362-4962。doi：10.1093/nar/gkae1310。URL：http://dx.doi.org/10.1093/nar/gkae1310。

Jonathan Frazer、Pascal Notin、Mafalda Dias、Aidan Gomez、Joseph K. Min、Kelly Brock、Yarin Gal 和 Debora S. Marks。利用进化数据的深度生成模型进行疾病变异预测。*Nature*，599(7883):91–95，2021年十月。ISSN 1476-4687。doi: 10.1038/ s41586-021-04043-8。URL http://dx.doi.org/10.1038/s41586-021-04043-8。

Ken Fujii、Yutaka Fujii、Takeshi Noda、Yukiko Muramoto、Tokiko Watanabe、Ayato Takada、Hideo Goto、Taisuke Horimoto 和 Yoshihiro Kawaoka。甲型流感病毒 NS节段的编码区和节段特异性非编码区对于其有效整合到病毒粒子中的重要性。《病毒学杂志》，79(6):3766–3774，2005年3月。ISSN 1098-5514。doi: 10.1128/jvi.79.6.3766-3774.2005。URL http://dx.doi.org/10.1128/JVI.79. 6.3766-3774.2005。

Sigrid Gouma、Elizabeth M. Anderson 和 Scott E. Hensley。打造有效流感疫苗面临的挑战。《病毒学年度评论》，7(1):495–512，2020年9月。ISSN 2327-0578。doi: 10.1146/annurev-virology-010320-044746。URL http://dx.doi.org/annurev-virology-010320-044746。

Haogao Gu、Rebecca L Y Fan、Di Wang 和 Leo L M Poon。甲型流感病毒中的二核苷酸进化动态。*Virus Evolution*，5(2)，2019年7月。ISSN 2057-1577。doi: 10.1093/ve/vez038。URL http://dx.doi.org/10.1093/ve/vez038。

James Hadfield、Colin Megill、Sidney M Bell、John Huddleston、Barney Potter、Charlton Callender、Pavel Sagulenko、Trevor Bedford 和 Richard A Neher。Nextstrain：实时追踪病原体进化。*Bioinformatics*，34(23):4121–4123，2018年5月。ISSN 1367-4811。doi: 10.1093/bioinformatics/bty407。URL http://dx.doi.org/10.1093/bioinformatics/bty407。

Alvin X. Han、Simon P. J. de Jong 和 Colin A. Russell。《免疫与季节性流感病毒的协同进化》。《自然综述：微生物学》，21(12):805–817，2023年8月。ISSN 1740-1534。doi: 10.1038/s41579-023-00945-8。URL http://dx.doi.org/10.1038/s41579-023-00945-8。

Brian Hie、Ellen D. Zhong、Bonnie Berger 和 Bryan Bryson。学习病毒进化与逃逸的语言。《科学》，371(6526):284–288，2021年1月。ISSN 1095-9203。doi: 10. 1126/science.abd7331。URL http://dx.doi.org/10.1126/science.abd7331。

E. Hoffmann、G. Neumann、Y. Kawaoka、G. Hobom 和 R. G. Webster。利用八种质粒从甲型流感病毒生成病毒的 DNA 转染系统。美国国家科学院院刊，97(11):6108–6113，2000。doi: 10.1073/pnas.100133697。

Edward C Holmes。《基因组时代的病毒进化》。《*PLoS Biology*》，5(10):e278，2007年10月。ISSN 1545-7885。doi: 10.1371/journal.pbio.0050278。URL http://dx.doi.org/ 10.1371/journal.pbio.0050278。

John Huddleston、John R Barnes、Thomas Rowe、Xiyan Xu、Rebecca Kondor、David E Wentworth、Lynne Whittaker、Burcu Ermetal、Rodney Stuart Daniels、John W McCauley、Seiichiro Fujisaki、Kazuya Nakamura、Noriko Kishida、Shinji Watanabe、Hideki Hasegawa、Ian Barr、Kanta Subbarao、Pierre Barrat-Charlaix、Richard A Neher 和 Trevor Bedford。《整合基因型和表型可改善季节性甲型 H3N2 流感进化的长期预测》。《*eLife*》，9，2020年9月。ISSN 2050-084X。doi: 10.7554/elife.60067。URL http: //dx.doi.org/10.7554/eLife.60067。

Jumpei Ito、Shusuke Kawakubo、Hiroaki Unno、Adam Strange、Spyros Lytras、Kaho Okumura、Alice Lilley、Ruth Harvey、Nicola Lewis 和 Kei Sato。通过人工智能驱动的抗原制图对季节性流感进化进行整合建模。2025年8月。doi：10.1101/2025.08.04.668423。URL http://dx.doi.org/10.1101/2025.08.04.668423。

Yanrong Ji、Zhihan Zhou、Han Liu 和 Ramana V. Davuluri。DNABERT：用于基因组 DNA 语言的预训练双向变换器编码器表示模型。*Bioinformatics*，37(15):2112–2120，2021。doi：10.1093/bioinformatics/btab083。

Melvin Johnson、Mike Schuster、Quoc V. Le 等。Google 的多语种神经机器翻译系统：实现零样本翻译。载于 计算语言学协会会刊，第 5 卷，第 339–351 页，2017 年。doi：10.1162/tacl a 00065。

Alexander Karollus、Johannes Hingerl、Dennis Gankin、Martin Grosshauser、Kristian Klemon 和 Julien Gagneur。物种感知 DNA 语言模型捕捉调控元件及其进化。《基因组生物学》，25(1)，2024年4月。ISSN 1474-760X。doi: 10.1186/s13059-024-03221-x。URL http://dx.doi.org/10.1186/s13059-024-03221-x。

Birgitte F. Koel、D. Burke、T. Bestebroer 等。受体结合位点附近的替换决定了流感病毒进化过程中的主要抗原变化。《科学》，342(6161):976–979，2013年。doi: 10.1126/science.1244730。

Sergey Kryazhimskiy、Georgii A. Bazykin 和 Jonathan Dushoff。甲型流感病毒基因中同义位点和非同义位点核苷酸使用的自然选择。《病毒学杂志》，82(10):4938–4945，2008年5月。ISSN 1098-5514。doi: 10.1128/jvi.02415-07。URL http: //dx.doi.org/10.1128/jvi.02415-07。

Siyuan Li、Zedong Wang、Zicheng Liu、Di Wu、Cheng Tan、Jiangbin Zheng、Yufei Huang 和 Stan Z. Li。Vqdna：释放向量量化在多物种基因组序列建模中的潜力，2024年。URL https://arxiv.org/abs/2405.10812。

Zeming Lin、Halil Akin、Roshan Rao、Brian Hie、Zhongkai Zhu、Wenting Lu、Nikita Smetanin、Robert Verkuil、Ori Kabeli、Yaniv Shmueli、Allan dos Santos Costa、Maryam Fazel-Zarandi、Tom Sercu、Salvatore Candido 和 Alexander Rives。利用语言模型进行进化尺度预测，实现原子级蛋白质结构预测。《科学》，379(6637):1123–1130，2023年3月。ISSN 1095-9203。doi：10.1126/science.ade2574。URL http://dx.doi.org/10.1126/ science.ade2574。

Zining Liu、Lei Zhang、Wenyu Zhang、Yuerong Lai 和 Tao Deng。甲型流感病毒 5′ 端节段特异性非编码区在感染过程中调控竞争性多节段 RNA 转录和选择性基因组包装。《病毒学杂志》，99(9)，2025年9月。ISSN 1098-5514。doi: 10.1128/jvi.00328-25。URL http://dx.doi.org/10.1128/jvi. 00328-25。

Jingzhi Lou、Weiwen Liang、Lirong Cao、Inchi Hu、Shi Zhao、Zigui Chen、Renee Wan Yi Chan、Peter Pak Hang Cheung、Hong Zheng、CaiqiLiu、Qi Li、MarcKa Chun Chong、Yexian Zhang,

Eng-kiong Yeoh、Paul Kay-Sheung Chan、Benny Chung Ying Zee、Chris Ka Pun Mok 和 Mag- gie Haitian Wang。基于位点的突变动力学对流感病毒的预测性进化建模。《自然·通讯》，15(1)，2024年3月。ISSN 2041-1723。doi：10.1038/ s41467-024-46918-0。URL http://dx.doi.org/10.1038/s41467-024-46918-0。

Enhao Ma、Xuan Guo、Mingda Hu、Penghua Wang、Xin Wang、Congwen Wei 和 Gong Cheng。面向 SARS-CoV-2 进化的预测语言模型。《信号转导与靶向治疗》，9（1），2024年12月。ISSN 2059-3635。doi：10.1038/s41392-024-02066-x。URL：http://dx. doi.org/10.1038/s41392-024-02066-x。

Hanover C. Matz 和 Ali H. Ellebedy。每年接种流感病毒疫苗：更新还是缩窄保护屏障?《实验医学杂志》，222（7），2025年4月。ISSN 1540-9538。doi：10.1084/jem.20241283。URL：http://dx.doi.org/10.1084/ jem.20241283。

Aarushi Mehrotra、Sarah Gurev、Noor Youssef 和 Debora Susan Marks。流感 H1N1 大流行预测与季节性进化。载于*ICML 2025*生成式人工智能与生物学（*GenBio*）研讨会，2025年。网址：https://openreview.net/forum?id=HgmjYnO3bz。

Richard A Neher、Colin A Russell 和 Boris I Shraiman。根据谱系树的形状预测进化。*eLife*，第3卷，2014年11月。ISSN 2050-084X。doi：10.7554/elife.03568。网址：http://dx.doi.org/10.7554/eLife.03568。

Richard A. Neher、Trevor Bedford、Rodney S. Daniels、Colin A. Russell 和 Boris I. Shraiman。季节性流感病毒抗原表型的预测、动态变化与可视化。《美国国家科学院院刊》，113(12)：E1701–E1709，2016年。doi：10.1073/ pnas.1525578113。

Alexey D. Neverov、Sergey Kryazhimskiy、Joshua B. Plotkin 和 Georgii A. Bazykin。甲型流感病毒表面蛋白的协同进化。《公共科学图书馆·遗传学》，11(8):e1005404，2015。doi: 10.1371/ journal.pgen.1005404。

Eric Nguyen、Michael Poli、Marjan Faizi、Armin Thomas、Callum Birch-Sykes、Michael Wornow、Aman Patel、Clayton Rabideau、Stefano Massaroli、Yoshua Bengio、Stefano Ermon、Stephen A. Baccus 和 Chris R´e。HyenaDNA：单核苷酸分辨率下的长程基因组序列建模，2023。网址：https://arxiv.org/abs/2306.15794。

Takeshi Noda。甲型流感病毒的选择性基因组包装机制。《冷泉港医学展望》，第 a038497 页，2020年8月。ISSN 2157-1422。doi: 10.1101/cshperspect. a038497。网址：http://dx.doi.org/10.1101/cshperspect.a038497。

Alec Radford、Jeff Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。语言模型是无监督多任务学习器。载于*NeurIPS 2019*会议论文集，2019。URLhttps://api.semanticscholar.org/CorpusID:160025533。

Alexander Rives、Joshua Meier、Tom Sercu 等。生物结构与功能源于将无监督学习扩展至2.5亿条蛋白质序列。《美国国家科学院院刊》，118(15):e2016239118，2021。doi: 10.1073/pnas.2016239118。

Melissa Sanabria、Jonas Hirsch、Pierre M. Joubert 和 Anna R. Poetsch。DNA 语言模型 GROVER 学习人类基因组中的序列上下文。*Nature MachineIntelligence*，6(8):911–923，2024年7月。ISSN 2522-5839。doi: 10.1038/s42256-024-00872-0。URL http://dx.doi.org/ 10.1038/s42256-024-00872-0。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。使用子词单元进行稀有词的神经机器翻译，2015。网址：https://arxiv.org/abs/1508.07909。

Yuelong Shu和John McCauley。GISAID：全球流感数据共享倡议——从愿景到现实。*Eurosurveillance*，22(13)，2017年3月。ISSN 1560-7917。doi: 10.2807/1560-7917. es.2017.22.13.30494。URL http://dx.doi.org/10.2807/1560-7917.ES.2017. 22.13.30494。

Silvie Van den Hoecke、Brian Verhack、Rony Ronsse等。利用下一代DNA测序分析甲型流感病毒的遗传多样性。*BMC Genomics*，16:79，2015。doi: 10.1186/s12864-015-1284-z。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。《Attention is all you need》，2017年。网址：https://arxiv.org/abs/1706.03762。

Ren´e M Vigeveno、Alvin X Han、Robert P de Vries、Edyth Parker、Karen de Haan、Sarah van Leeuwen、Katina D Hulme、Adam S Lauring、Aartjan J W te Velthuis、Geert-Jan Boons、Ron A M Fouchier、Colin A Russell、Menno D de Jong 和 Dirk Eggink。《人季节性流感病毒 A(H3N2) 的长期进化与聚合酶复合物活性增强有关》。*VirusEvolution*，10(1)，2024年1月。ISSN 2057-1577。doi：10.1093/ve/veae030。网址：http://dx.doi.org/10.1093/ve/veae030。

Rui Yang、Minglei Pan、Jiamei Guo、Yong Huang、Qiangfeng Cliff Zhang、Tao Deng 和 Jianwei Wang。绘制甲型流感病毒基因组RNA结构及其相互作用，揭示病毒复制的关键要素。*Cell Reports*，43(3):113833，2024年3月。ISSN 2211-1247。doi: 10.1016/j.celrep.2024.113833。URLhttp://dx.doi.org/10.1016/j.celrep.2024. 113833。

Zhihan Zhou、Yanrong Ji、Weijian Li、Pratik Dutta、RamanaV Davuluri 和 Han Liu。DNABERT-2：面向多物种基因组的高效基础模型与基准测试。载于第十二届国际学习表征会议，2024年。URL https://openreview.net/ forum?id=oMLQB4EZE1。

Marta Łuksza 和 Michael L¨assig。流感预测适应度模型。*Nature*，507(7490): 57–61，2014年2月。ISSN 1476-4687。doi: 10.1038/nature13087。URL http://dx.doi. org/10.1038/nature13087。

A 本研究所用数据

A.1 数据集

表 1 总结了亚型–分支覆盖情况和数据集规模。除两种主要亚型（H3N2 和 H1N1）外，我们还有意纳入了许多低资源亚型——其中 19 种有 15 种的序列数量少于 100 条——以使预训练接触更广泛的抗原/基因组多样性，并促进泛化。随后，这些长尾亚型主要用于对迁移能力进行压力测试，具体是在亚型分类微调任务中进行测试；尽管样本有限，我们仍观察到良好的表现。相比之下，我们的主要预测实验并不依赖罕见亚型：训练和评估均使用 H1N1 和 H3N2，并少量扩展至 H7N9。

此外，表 2 展示了所收集毒株的时间分布，覆盖 2012 年至 2022 年。表 3 还总结了这些预训练序列在美国、中国和澳大利亚等全球主要地区的地理分布，确保用于训练共享模型骨干网络的数据具有广泛性和多样性。因此，预测结论不会受到极端类别不平衡的干扰。

表 1：亚型–分支覆盖及数据集计数

亚型	分支	预训练	微调
H1N1	6B.1；6B.1A；6B.1A.1；6B.1A.2；6B.1A.3；6B.1A.5；6B.1A.5a；6B.1A.5a.1；6B.1A.5a.2；6B.1A.5a.2a；6B.1A.5a.2a.1；6B.1A.5b；6B.1A.6；6B.1A.7；6B.2	20680	5565
H1N2	6B.1；6B.1A.1；6B.1A.5a；6B.1A.5a.2a；6B.1A.5a.2a.1；6B.1A.6	31	74
H2N2		70	78
H3N1	3C.2a1b.2a.2a.3a.1；3C.3a	1	4
H3N2	3C.2；3C.2a；3C.2a1；3C.2a1b.1；3C.2a1b.1a；3C.2a1b.1b；3C.2a1b.2；3C.2a1b.2a；3C.2a1b.2a.1；3C.2a1b.2a.1a；3C.2a1b.2a.1a.1；3C.2a1b.2a.2；3C.2a1b.2a.2a；3C.2a1b.2a.2a.1；3C.2a1b.2a.2a.1a；3C.2a1b.2a.2a.1b；3C.2a1b.2a.2a.2；3C.2a1b.2a.2a.3；3C.2a1b.2a.2a.3a；3C.2a1b.2a.2a.3a.1；3C.2a1b.2a.2a.3b；3C.2a1b.2a.2b；3C.2a1b.2a.2c；3C.2a1b.2a.2d；3C.2a1b.2b；3C.2a2；3C.2a3；3C.2a4；3C.3a；3C.3a1	32758	8979
H3N8	3C.2		3
H5N1	1.1.1；1.1.2；2.1.2；2.1.3.2；2.1.3.2a；2.1.3.2b；2.1.3.3；2.2；2.2.1；2.2.1.2；2.3.2.1a；2.3.2.1b；2.3.2.1d；2.3.2.1e；2.3.2.1g；2.3.4; 2.3.4.1; 2.3.4.2; 2.3.4.3; 2.3.4.4b; 3; 7	104	298
H5N6	2.3.4.4；2.3.4.4b；2.3.4.4e；2.3.4.4g；2.3.4.4h	25	37
H5N8	2.3.4.4b	1	
H6N1		1	
H7N2		1	
H7N3		2	
H7N4		2	
H7N7		2	
H7N9		793	1047
H9N2		37	78
H10N3			3
H10N8		4	3
总计		54512	16169

表 2：预训练毒株按亚型划分的时间分布

亚型	2022	2021	2020	2019	2018	2017	2016	2015	2014	2013	2012 年之前
H10N8									1	3	
H1N1	124	159	2370	4918	2952	1092	2122	786	353	295	5509
H1N2		2	1		4	1	2	1			20
H2N2											70
H3N1										1	
H3N2	1472	3674	963	6723	4086	5470	2515	2018	1117	604	4116
H5N1								2		7	95
H5N6		3	2		4	3	6	4	3		
H5N8			1								
H6N1										1	
H7N2							1				
H7N3											2
H7N4					2						
H7N7										1	1
H7N9				4	1	173	60	68	153	334	
H9N2			3	11	2	2	6	6	2	2	3
总计	1596	3838	3340	11656	7051	6741	4712	2885	1629	1248	9816

表 3：按亚型划分的预训练毒株地理分布

亚型	澳大利亚	巴西	中国	法国	俄罗斯	新加坡	USA	英国		其他
H10N8			4							
H1N1	420	375	1044	638	621	564	9692		878	6448
H1N2		1		1			25			4
H2N2	1		2		6	8	27		2	24
H3N1		1								
H3N2	2006	1136	1283	773	577	403	16532		1895	8153
H5N1			32							72
H5N6			23							2
H5N8					1					
H6N1			1							
H7N2							1			
H7N3										2
H7N4			2							
H7N7										2
H7N9			791							2
H9N2			35							2
总计	26277	14711	3217	2775	2427	1513	1412		1205	975

A.2 预训练数据整理与质量控制

我们保留了包含全部八个片段（PB2、PB1、酸性聚合酶蛋白、血凝素、核蛋白、神经氨酸酶、基质蛋白、非结构蛋白）的分离株。

CDS解析与有效性（按片段）。

- 以 ATG 起始，并以 TAA/TAG/TGA 中的一个终止；
- 长度是 3 的倍数，且不包含内部终止密码子；
- 编码序列长度处于针对该片段经验定义的合理范围内（已移除异常值和明显的截短记录）。

其他卫生检查。

• 移除包含过多歧义字符的记录；
• 确保亚型注释与 HA/NA 一致；
• 使用固定的拼接顺序（PB2→PB1→酸性聚合酶蛋白→血凝素→核蛋白→神经氨酸酶→基质蛋白→非结构蛋白）。

A.3 微调数据整理（仅 HA/NA）

微调采用与附录 A.2 相同的质量控制标准，但仅应用于血凝素和神经氨酸酶。只有当**分离株的**血凝素和神经氨酸酶编码序列通过有效性检查时，该分离株才**符合条件**；其余六个片段无需提供，可以缺失或不完整，即使存在也会被忽略。

输入采样。对于每个区域/亚型，我们从附录 B 的时间窗口（一月至二月、九月至十月、十一月至十二月）中选择三个输入。微调模型仅使用 HA/NA 标记（其他片段如有提供也不会使用）。预测目标是附录 B 中定义的季节性共识。

B 微调样本选择：三输入设置

对于每个目标流感季（ $\rightarrow$ 年十月至 $T{+}1$ 年三月），我们根据采集日期（忽略提交延迟），为每个区域/亚型构建一个三输入样本。每个输入都必须通过附录 A.2 中的质量控制，并至少包含 HA 和 NA。

三个时间窗口（每个窗口一个样本）：

• **九月至十月**， T年。捕捉流感季开始时早期的跨区域引入——这些“种子”日后可能发展为持续传播。非常早的月份（七月至八月）噪声较大，预测能力较弱；而采样过晚则可能错过早期信号。

• **十一月至十二月**， T年。温带地区进入冬季传播期；不同谱系的增长率差异变得更加明显，与十一月下旬和十二月假期相关的跨区域流动也会反映在数据中。

• **一月至二月**， $T{+}1$年。这个时间窗口尽可能贴近世界卫生组织北半球疫苗组分会议，可捕捉季中替代事件以及新近固定的关键突变。考虑到生产时间表，二月之后的信息可操作性较低。

目标输出： 我们预测实验的预测目标与该领域常规给出的毒株选择答案一致：**最近序列**。为实现这一目标，我们构建了一个三输入样本（兼顾提前期与时效性）。我们的模型在这一标准目标上表现**良好**，在进一步针对共识序列目标进行测试时则取得了**更优结果**。这一稳健发现证实了我们预测框架的有效性。

C 训练动态

在匹配的优化与词元预算下，图5显示三种预训练变体之间存在一致的差异。**早期预热阶段：**全基因组模型迅速下降，分段式 改善较慢，而不完整基因组下降极少。**中后期：**全基因组持续下降并收敛至最低渐近值；分段式在更高水平趋于平稳；不完整基因组几乎保持平坦，视觉上偏差更大，方差降低有限。

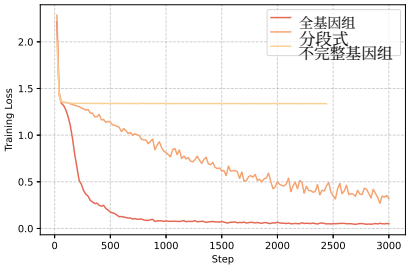

图5：三种变体的预训练损失随步数的变化。

完整的跨片段上下文充当辅助监督信号，提升样本效率；分离片段会消除这些依赖关系；裁剪随机时间窗口会混合不相关的上下文并丢弃边界，从而削弱学习信号。所有变体共享
骨干网络、优化器、调度策略、更新次数和词元预算；曲线在展示时采用相同的平滑处理，且平滑前的排序保持不变。

D 亚型分类

我们在每个预训练骨干网络之上，使用相同的分类头和优化设置微调亚型分类器。在定性分析中，我们提取倒数第二层嵌入，并使用所有变体共用的 t-SNE 配置（相同的困惑度、初始化方式以及困惑度与样本数之比）将其投影到二维空间。t-SNE 仅用于可视化（图6）；我们的主要结论基于正文中报告的定量预测和分类指标。为完整起见，附录表4汇总了在用于绘制 t-SNE 图的同一隐藏层上计算的聚类指标（轮廓系数、调整兰德指数、归一化互信息）：全基因组模型的轮廓系数和调整兰德指数最高，在归一化互信息上则与分片变体并列，表明与不完整基因组和分片基线方法相比，其亚型簇更加紧凑，且与标签的一致性更高。各亚型 F1分数的雷达图进一步凸显了 AntigenLM 的优异性能（图7）。

表 4：用于 t-SNE 图的隐藏层上的亚型嵌入聚类指标。所有指标均为越高越好。

模型	轮廓系数 ↑	调整兰德指数 ↑	归一化互信息 ↑
全基因组	0.877	0.953	0.923
不完整基因组	0.761	0.938	0.890
分段式	0.773	0.920	0.923
仅抗原（核苷酸）	**0.921**	**0.969**	**0.967**

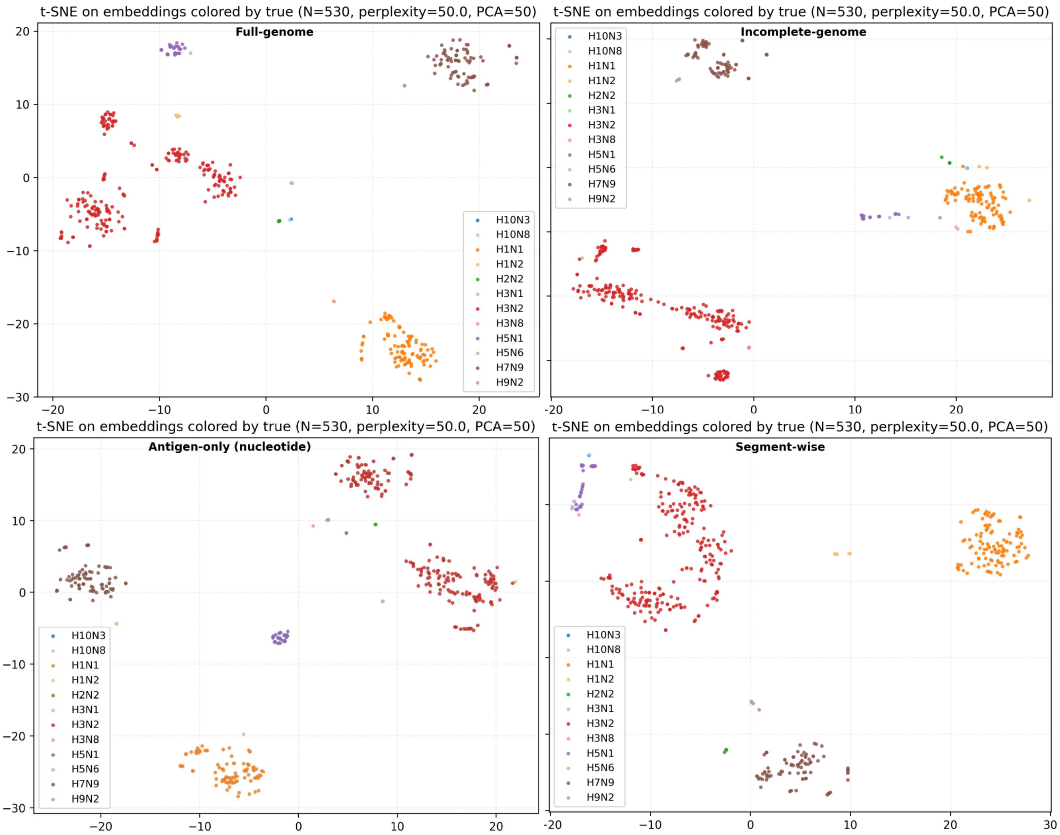

图6：四个微调亚型分类器倒数第二层嵌入的t-SNE，每个分类器均由不同的预训练变体初始化。全基因组和仅抗原（核苷酸）通常形成最紧凑的类别簇；分段式较为松散；不完整基因组介于两者之间。

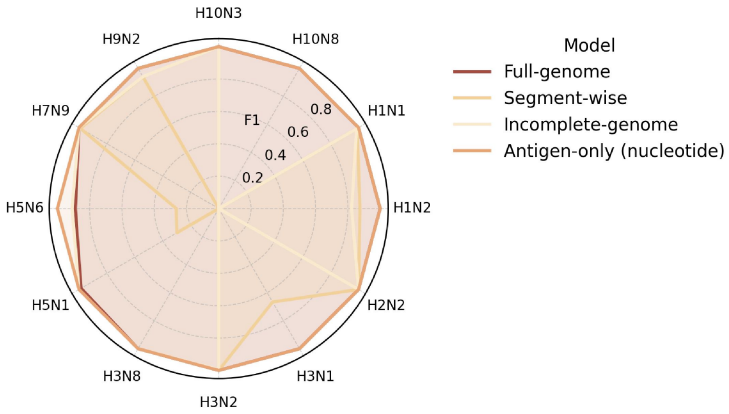

图7：AntigenLM（全基因组）、不完整基因组、分片模型和仅抗原（核苷酸）各亚型的 F1分数雷达图。