[toc]

文献名称：Genome modeling and design across all domains of life with Evo 2

文献URL：https://www.biorxiv.org/content/10.1101/2025.02.18.638918v1

文章总结

模型概述：

Evo 2 是一个大型基因组语言模型，具有最多 400 亿个参数，能够处理最长 100 万个碱基对的序列。它适用于多个复杂层次的预测和生成任务。

Evo 2 训练了两个版本：

- 一个 7B 参数模型，训练数据为 2.4 万亿个标记。
- 一个 40B 参数模型，训练数据为 9.3 万亿个标记。 两个版本的模型都采用 两阶段训练，以捕捉从分子到有机体的生物学尺度。

训练数据：

Evo 2 的训练数据来自生命的所有领域（细菌、古菌、真核生物和噬菌体），并且聚焦于 非冗余的核酸序列。

训练数据总量超过 8.8 万亿个核苷酸。训练数据中排除了感染真核宿主的病毒基因组，验证了排除这些数据后，模型在真核病毒序列上的困惑度较高，表明在这一领域的语言建模表现较差。 数据集以 OpenGenome2 名称公开发布。

训练过程

让我们用一个简化的例子来说明Evo 2的训练过程，并通过一个大致的框架来描述每个步骤。

假设：

- 目标是训练一个基于Evo 2模型的基因组语言模型，用于从DNA序列中预测下一个碱基（类似于文本生成的任务）。

步骤 1：数据准备

首先，你需要一个包含大量DNA序列的数据集。为了简化，我们假设我们有以下的简短DNA序列：

```
AGCTAGCTAGCTAAGCTGAC
```

这些DNA序列是模型的输入数据。为了有效地训练，数据会被切分成更小的token（比如每3个碱基为一个token，或直接按碱基对切分）。然后，每个token会与目标label（下一个碱基）配对。

步骤 2：模型架构

Evo 2使用了**StripedHyena 2**架构，这是一种卷积与注意力相结合的混合架构。它通过不同类型的操作符来处理DNA序列。对于这个例子，假设我们使用简化的变体：

- **卷积操作符**：捕捉DNA序列局部的结构和模式。
- **注意力机制**：帮助模型学习长距离依赖，例如基因组中不同位置之间的关系。

步骤 3：第一阶段预训练（短上下文）

在第一阶段的预训练中，模型会在较短的上下文长度（例如，8,192个token）下进行训练。这一阶段的目标是让模型学会短距离依赖，比如基因的局部功能元素。模型的目标是预测下一个token（碱基）。

例如，给定输入序列“**AGCTAGCTAGCTA**”，模型的目标是预测下一个碱基“**A**”。在这一阶段，模型仅使用较短的序列进行训练，优化在较短范围内的预测能力。

步骤 4：第二阶段中期训练（长上下文）

一旦第一阶段的预训练完成，Evo 2会进入第二阶段的中期训练，这时上下文长度扩展到**1百万个token**。在这个阶段，模型将学习更长距离的依赖关系，例如基因组的不同区域之间的相互作用。

在训练过程中，Evo 2使用了旋转嵌入等方法来扩展上下文，以便处理更长的序列。这是Evo 2的关键创新之一，允许它处理从分子到生物体尺度的长距离依赖。

步骤 5：损失计算与优化

在每个训练步骤中，模型会根据输入序列生成预测（例如，预测下一个碱基），然后计算预测和实际值之间的损失。根据损失反向传播，优化模型的参数。

步骤 6：模型验证与评估

训练完成后，你可以通过在未见过的DNA序列上进行评估来测试模型的效果。例如，验证模型是否能够准确地预测DNA序列中缺失的部分。

例如，如果给定序列“**AGCTAGC**”，模型应该能够预测出接下来的碱基（比如“**T**”）。

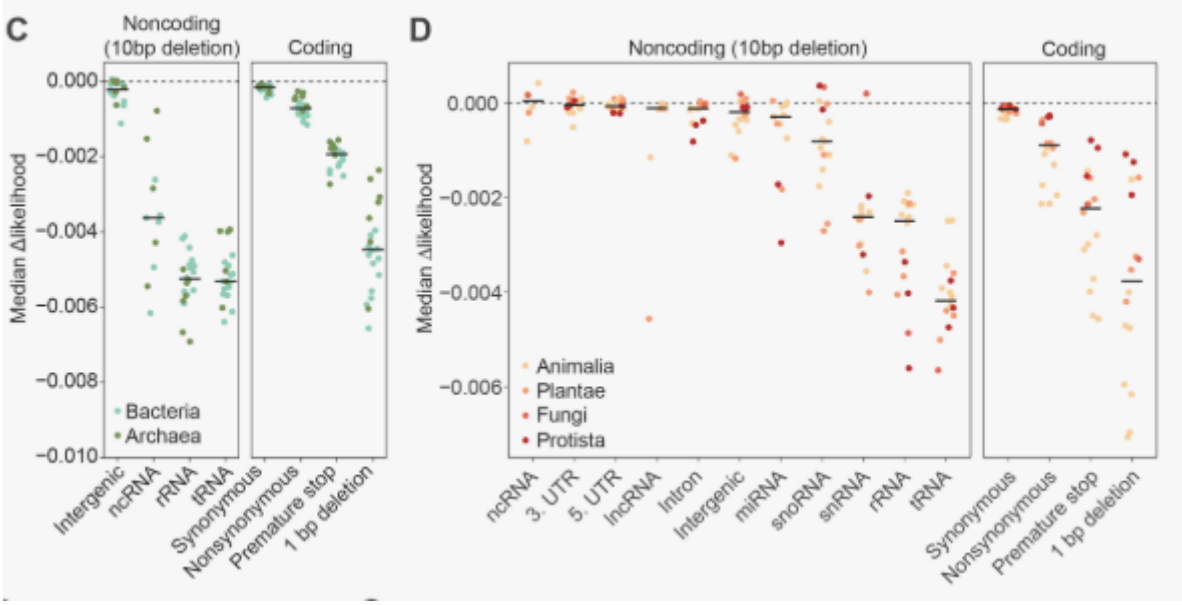
简单总结：

1. **数据准备**：准备大规模DNA序列数据。
2. **模型架构**：使用**StripedHyena 2**架构，结合卷积和注意力机制。
3. **第一阶段预训练**：使用较短的上下文（如8,192 token）训练模型，学习局部依赖。
4. **第二阶段中期训练**：通过旋转嵌入等方法扩展上下文，训练更长的序列，学习长距离依赖。
5. **损失计算与优化**：计算损失，优化模型参数。
6. **评估与应用**：在新数据上评估模型的生成和预测能力。

通过这种逐步扩展上下文长度的训练方法，Evo 2能够高效地学习DNA序列中的复杂模式和长距离依赖关系，从而提高模型的预测能力和生成能力。

模型验证

Evo 2预测突变对蛋白质，RNA和生物体适应性的影响



验证方法：

- **数据来源与基因选择：** 从NCBI获取了20种原核和16种真核物种的参考基因组序列及其注释。每个物种随机选择了1,024个蛋白质编码基因（N. equitans选择了所有536个基因）。对于每个基因，选择了从起始密码子第一碱基的前后20个碱基的基因组坐标，并将每个位置的野生型碱基突变为三种可能的替代碱基，生成单核苷酸变异（SNVs）。
- **计算与对比：** 使用 Evo 2 7B 模型计算突变前后的基因序列似然差异。具体步骤如下：
  - 计算**野生型序列的似然**，即使用 Evo 2 模型预测野生型序列的似然值。
  - 计算**突变型序列的似然**，即将每个位置的野生型碱基替换为突变碱基（如A突变为C），然后预测突变后的序列似然值。
  - 计算**delta likelihood**：即突变型和野生型序列对数似然的差异，公式如下：

```
![alt text](image-1.png)


- 正的 delta likelihood：如果 delta likelihood 是正值，说明突变后的序列比野生型序列的预测似然更高，这意味着突变可能提高了序列的适应性或功能。
- 负的 delta likelihood：如果 delta likelihood 是负值，说明突变后的序列的似然低于野生型序列，通常意味着突变对序列产生了负面影响，例如可能破坏了基因的功能。

 - 通过计算delta likelihood，可以量化突变对基因序列适应性或功能的影响。
```

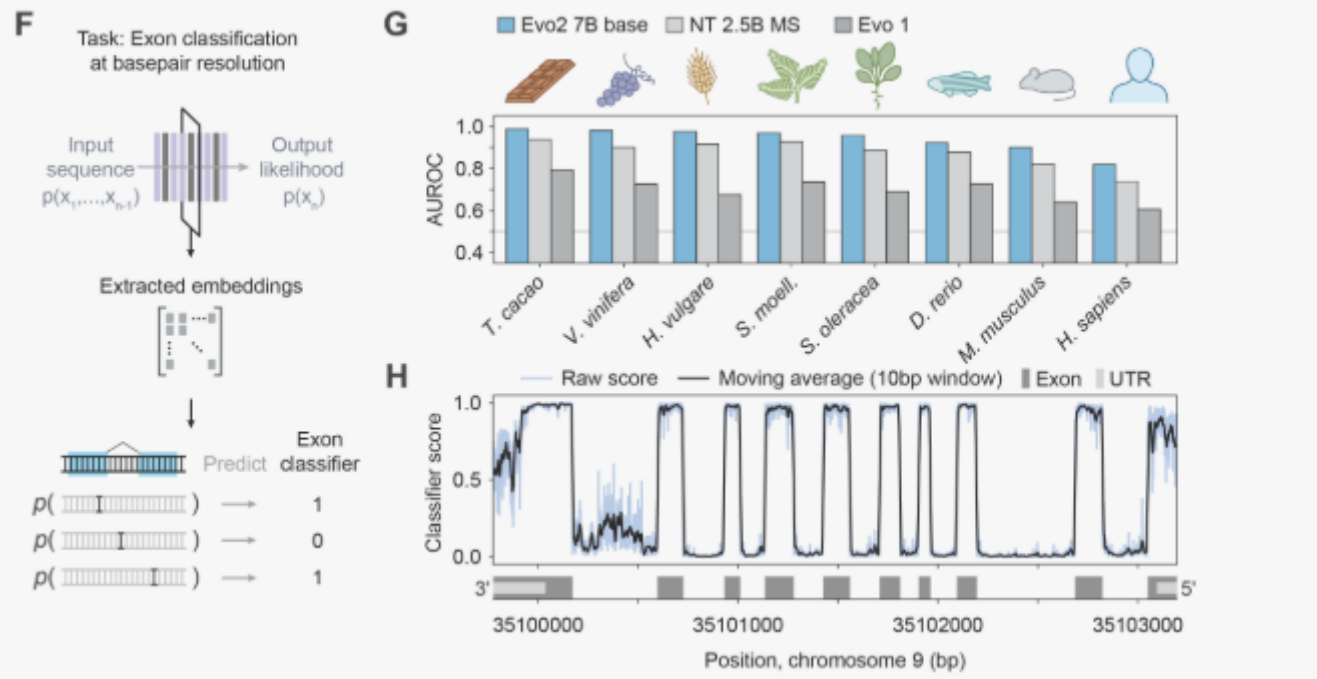
验证结果：

- **零-shot预测：** Evo 2 在没有任何任务特定的微调或监督的情况下，通过零-shot预测突变效应，展示了其强大的通用性。
- **SNV效应：** 在启动密码子附近和起始密码子位置引入突变时，Evo 2 预测的似然发生了显著变化。突变对启动密码子的影响最为明显，接着是基于三碱基的周期性模式，而在摆动位置的变化影响较小。结果显示了Evo 2成功学习了基因组中的核心结构特征。

- **非编码区域：** 在非编码区域中，tRNA和rRNA的缺失对似然的影响大于在基因间区域的缺失，体现了这些RNA在基因组中的重要性。较大的40B模型对miRNA和snoRNA序列的缺失表现出了更高的敏感性，表明更大的模型可能捕捉到了更细致的调控特征。
- **变异类型：** 在编码序列中，非同义变异、提前终止密码子和框移突变引起的似然变化远大于同义突变。Evo 2也能够有效地区分不同遗传密码的提前终止密码子，证明了长上下文窗口在识别这些密码子上的重要性。
- **实验验证：** 与深度突变扫描（DMS）的实验数据进行比较，Evo 2的零-shot预测结果与实验测量的功能效应高度相关，进一步验证了其预测准确性和生物学意义。

总结：
通过零-shot预测，Evo 2在没有任何微调的情况下能够成功预测不同类型的突变对基因功能的影响。其预测与实验数据高度一致，表明Evo 2具有强大的基因组变异预测能力，能够捕捉突变对蛋白质、RNA和整体适应性的影响。

Evo 2 外显子/内含子的分类任务



验证方法：

- **数据集与物种选择：**
  为了评估 **Evo 2** 嵌入向量在外显子与内含子分类任务中的表现，从 **PANTHER19.0** 数据库中选择了94种真核物种。我们将物种随机划分为训练集（80%）、超参数优化集（10%）和测试集（10%）。特别地，**Homo sapiens**（人类）、**Mus musculus**（小鼠）和 **Danio rerio**（斑马鱼）被手动分配到测试集。
  对每个物种，随机采样了来自 **NCBI RefSeq** 注释的长非编码基因和蛋白质编码基因的位置，确保样本的无偏性。

- **模型与训练：**
  使用 **Evo 2**、**Evo 1** 和 **Nucleotide Transformer** 三种模型进行比较。为每个物种从模型的顶层提取嵌入向量，然后通过加权二元交叉熵损失（weighted binary cross-entropy loss）训练分类器。通过 **Tree-structured Parzen Estimators (TPE)** 进行超参数优化，优化过程中包括隐藏层数量、隐藏层维度、学习率、批量大小和权重衰减等。

- **外显子与内含子分类器：**
  通过提取 **Evo 2** 的嵌入向量，开发了一个单核苷酸分辨率的 **外显子/内含子分类器。**该分类器的目标是通过优化后的嵌入向量来预测基因组中位置是否属于外显子或内含子区域。优化后的分类器在 **8个物种**的数据集上进行了评估，并与 **Nucleotide Transformer** 和 **Evo 1** 的分类器进行了比较。

验证结果：

- **分类性能：**
  使用 **Evo 2** 嵌入训练的外显子/内含子分类器在多个物种上的表现优异。分类器的 **AUROC（接收者操作特征曲线下的面积）** 范围从 **0.82 到 0.99**，显示了其在不同物种中的高准确性，表明 **Evo 2** 嵌入能有效捕捉外显子和内含子的典型模式。

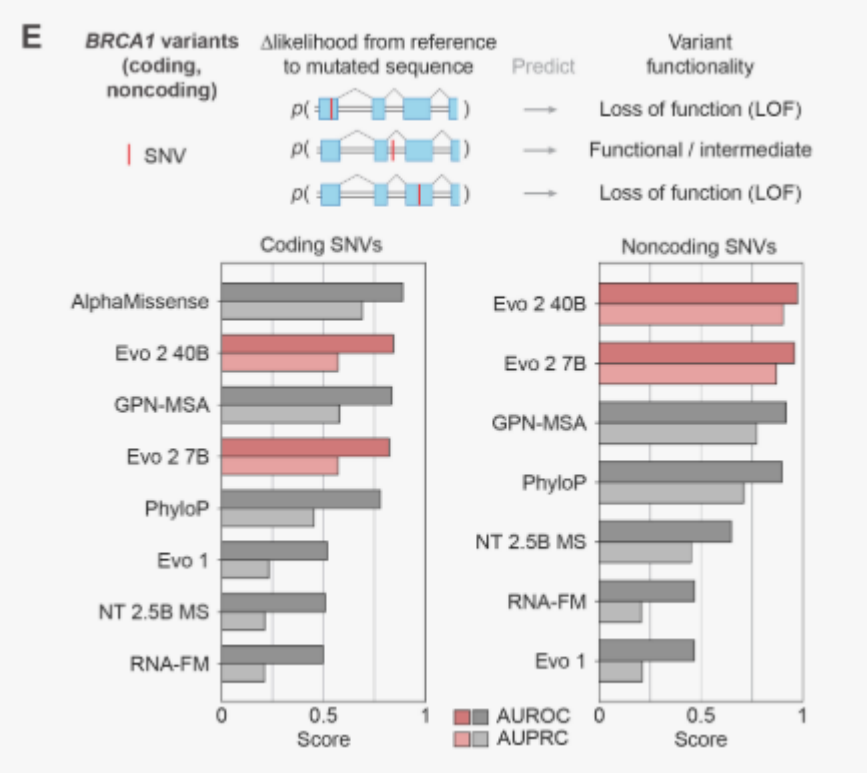
- **人类基因示范：**
  作为示范，使用该分类器扫描了人类 **STOML2** 基因的蛋白质编码区，分类器输出的概率与注释的外显子位置高度一致，能够准确地识别出外显子和内含子的边界。这一结果表明，**Evo 2** 嵌入在实际基因组功能注释中具有很好的应用潜力。

- **跨物种比较：**
  在与 **Nucleotide Transformer** 和 **Evo 1** 的比较中，**Evo 2** 展现出了更优越的性能，说明其在处理外显子与内含子分类任务时具有明显的优势。

总结：

- **Evo 2** 通过嵌入向量在外显子与内含子分类任务中表现出色，特别是在跨多个物种的任务中，AUROC值达到 **0.82-0.99**，证明了其强大的跨物种泛化能力。
- 在具体示范上，**Evo 2** 能够有效识别外显子与内含子的边界，特别是在 **STOML2** 基因的蛋白质编码区域中，分类器与实际注释之间具有高度一致性。
- **Evo 2** 在与其他模型（如 **Nucleotide Transformer** 和 **Evo 1**）的比较中，展示了更高的性能，进一步验证了其在基因组功能注释任务中的潜力。

Evo 2能够准确预测人类临床变异



Evo 2 验证方法：

1. **验证方法：**
    - **数据来源与基因选择：** 对于所有BRCA1和BRCA2的SNV（单核苷酸变异），从原始研究中获取了周围8,192bp窗口的序列。根据原始研究中的注释，将这些SNV分类：
        - 对于BRCA1，将标记为“LOF”的SNV（共823个）视为功能丧失变异（loss-of-function, LOF），将标记为“FUNC”或“INT”的SNV（共3,070个）视为功能/中间变异。
        - 对于BRCA2，将标记为“P strong”、“P moderate”和“P supporting”的SNV（共1,156个）视为LOF变异，将标记为“B strong”、“B moderate”和“B supporting”的SNV（共5,681个）视为功能/中间变异。
2. **zero-shot预测：**
    - 使用 **Evo 2** 进行zero-shot预测，针对BRCA1和BRCA2的编码和非编码变异进行了效果评估。在编码SNV上，Evo 2的预测表现优异，并在BRCA1非编码SNV上设立了新的最先进水平。重要的是，当同时评估编码和非编码变异时，Evo 2超越了所有其他模型，展现出强大的性能。
    - 特别是，Evo 2在没有专门针对人类变异进行训练的情况下，仅利用多物种的变异作为进化约束的代理，仍然能提供准确的zero-shot预测。这一点非常关键，因为许多非编码变异通常会被排除在变异分析报告之外。
3. **有监督分类：**
    - 为了进一步提高预测精度，使用Evo 2的序列嵌入作为特征输入到一个有监督的深度神经网络中。该网络专门针对BRCA1变异进行了训练，以评估是否能够超越zero-shot方法。
    - **神经网络架构：** 神经网络由三层隐藏层组成，每一层后面都有ReLU激活、批量归一化和dropout（p=0.3）。最终的输出层使用sigmoid激活函数进行二分类概率预测。
    - 在训练过程中，使用了早停法、学习率衰减和梯度裁剪。最终的模型在BRCA1编码SNV测试集上取得了 **AUROC = 0.92**，在所有BRCA1 SNV上表现更为出色（AUROC = 0.95），并且在测试集上的AUPRC为 **0.86**。

4. **Evo 2 嵌入的性能：**
    - 进一步分析了Evo 2在变异分类任务中的表现，特别是通过从Evo 2的不同层提取嵌入向量，来评估哪些层包含了最相关的信息。
    - 最终，使用Evo 2的 **第20层嵌入**，在BRCA1 SNV的分类任务中表现最佳，证明了Evo 2嵌入向量在变异分类中的潜力。

****验证结果 ****

- **zero-shot预测：** Evo 2在BRCA1和BRCA2的编码与非编码变异上展示了优异的性能，尤其在BRCA1的非编码区域设立了新的基准。
- **有监督分类：** 利用Evo 2的嵌入进行的有监督训练模型，在BRCA1变异的分类任务中表现出色，超越了现有的所有基准。
- **临床相关性：** 这些结果进一步强调了Evo 2嵌入可以作为基础，用于训练更为专门的模型，特别是具有高度临床相关性的变异分类任务。

总结：
Evo 2通过zero-shot预测和有监督训练，成功展示了其在BRCA1和BRCA2变异预测中的应用价值。它不仅能够准确预测编码和非编码变异的功能效应，还能为更精细的变异分类任务提供强有力的基础，尤其是在临床变异分类和报告生成领域。

Evo 2 特征解释：从分子到基因组尺度

验证方法：

- **模型的机制可解释性：**
  Evo 2 通过学习基因组序列的复杂表示来捕捉不同的生物学特征，而不依赖于显式的生物学标签或注释。为了探究 Evo 2 学到的特征，我们使用了 **稀疏自编码器（SAE）** 来分析其表示（或神经元激活模式），以便将模型拆解为稀疏、高维的表示，其中各个潜在维度通常能够展现出人类可解释的模式。

- **SAE训练与对比特征搜索：**
  我们在 **Evo 2** 的第26层表示上训练了 **Batch-TopK SAE**，该层被初步分析发现是大多数有趣特征的来源。SAE 在10亿个标记上进行了训练，涵盖了完整的真核生物和原核生物基因组。在训练后，我们通过 **对比特征搜索（contrastive feature search）** 找到了与已知生物学概念对齐的特征，具体来说，找到了一些特征，它们在包含特定生物学概念的序列片段中富集。

验证结果：

- **进化信号与移动遗传元件：**
  我们发现 Evo 2 学到的潜在维度能够捕捉基因组中嵌入的进化信号。例如，特征 **f/19746** 与原核生物中的 **前噬菌体（prophage）区域** 密切相关，尤其在大肠杆菌基因组中的注释噬菌体区域（如隐性噬菌体 CPZ-55）上激活。此外，Evo 2 在 **CRISPR阵列的间隔序列** 上也有激活，这些序列在CRISPR适应过程中由外源遗传物质（如噬菌体DNA）整合进入基因组。这表明，Evo 2 将CRISPR间隔序列与噬菌体序列关联，而非直接记忆噬菌体序列。这些特征揭示了Evo 2能够捕捉基因组中的复杂生物学系统，并为基因组注释提供新的视角。

- **基因组组织与功能注释：**
  通过对比特征搜索，Evo 2 学到的特征与基因组中已知的功能注释类型（如开放阅读框（ORFs）、基因

间区域、tRNA 和 rRNA）高度相关。进一步分析表明，Evo 2 还能够捕捉蛋白质二级结构（如 **α-螺旋** 和 **β-折叠**）的结构特征，表明它能够处理DNA之外的高级结构信息，展现了Evo 2的多模态性质。

- **人类基因组与突变效应：**
  我们将对比特征搜索的分析扩展到人类基因组，特别是探讨突变对基因结构的影响。通过对数千个人类编码序列引入突变，发现了一个突变敏感的特征（**f/24278**），该特征在 **框移突变** 和 **提前终止突变** 上表现出更强的激活。这表明，Evo 2 能够识别突变的严重性，捕捉到简单基因结构之外的潜在特征。此外，通过混合的原核生物与真核生物 SAE（第26层），我们还观察到在人类基因的启动子区域存在与已知转录因子结合位点高度相似的DNA基序激活，进一步证明了Evo 2不仅能识别编码序列，还能辨别调控元件。

总结：

- **Evo 2 学到的表示** 捕捉了从基因组到更高层次结构的复杂生物学信息，能够识别基因组中的进化信号、移动遗传元件、编码序列和调控元件。
- 通过 **稀疏自编码器** 和 **对比特征搜索**，我们发现 Evo 2 在不同的生物学和进化层面上学习到了具有生物学意义的特征，这些特征不仅与已知的基因组注释类型（如外显子、内含子）一致，还能提供新的视角帮助挖掘基因组中的未知生物学信息。
- **Evo 2 的内部表示** 不仅能够识别基因结构，还能够洞察突变的严重性，捕捉到突变对基因功能的潜在影响。

Evo 2 在基因组规模上的生成能力验证总结：

验证方法：

- **基因序列生成任务：**
  Evo 2 是一个生成模型，训练目标是预测序列中的下一个碱基对。为了评估 Evo 2 在不同物种和基因组尺度上的生成能力，研究者对多个物种的基因组序列进行了生成实验，包括来自古菌、原核生物和真核生物（如真菌、原生生物、植物和动物）的基因序列。
    - 研究者为每个物种选择了高度保守的代表基因，并提供了包含1,000个上游碱基对和目标基因的前500-1000个碱基对的上下文，评估 Evo 2 在这些上下文中完成基因序列的生成能力。

- **蛋白质生成与比较：**
  通过生成线粒体基因组和原核生物基因组来评估 Evo 2 的生成能力，研究者使用 BLASTp 和 AlphaFold 3 对生成的蛋白质进行分析，检查其与天然基因组序列和蛋白质的相似性。

- **评估生成质量：**
  对生成的基因组进行质量评估，包括序列恢复、同源性分析（如与天然蛋白质的比较）、蛋白质复杂折叠预测等，以验证生成的序列是否符合自然序列的特征。

验证结果：

- **基因序列完成度：**
  Evo 2 在多个物种（包括古菌、原核生物和真核生物）的基因序列生成任务中表现出色，能够准确地恢复目标基因序列。特别是，随着模型规模的增大，Evo 2 40B 和 7B 模型的蛋白质恢复率优于 Evo 1。

- **线粒体基因组生成：**
  使用 Evo 2 40B 生成了250条16 kb的线粒体基因组序列，并且这些生成的基因组包括了正确数量的编码

序列（CDS）、tRNA 和 rRNA 基因。BLASTp 分析表明，生成的基因与天然线粒体基因有不同程度的同源性，且生成的序列保持了正确的基因排列（synteny）和适度的序列多样性。AlphaFold 3 分析表明，生成的蛋白质结构与预期的线粒体蛋白复合体折叠和相互作用一致。

- **原核基因组生成：**
  使用 Evo 2 生成了 M. genitalium（大肠分枝杆菌）基因组，并通过 HHpred 分析确认，生成的基因中约70%包含与天然基因库（Pfam）的显著匹配，明显优于 Evo 1 的表现。此外，生成的蛋白质在结构和序列上与自然蛋白质具有一致性，表明 Evo 2 在生成原核基因组时能够有效地重现天然特征。

- **真核基因组生成：**
  在评估 Evo 2 生成真核基因组的能力时，研究者使用了 S. cerevisiae（酿酒酵母）基因组的10.5 kb序列作为提示，成功生成了330 kb的DNA序列。这些生成的序列具有预测的 tRNA、适当位置的启动子以及包含内含子结构的基因，表明 Evo 2 能够生成符合真核生物特征的基因序列。

- **生成质量改进空间：**
  尽管生成的酵母基因组序列中的 tRNA 和基因密度低于天然酵母基因组，但生成过程是通过简单的无约束自回归生成完成的，表明通过优化推理策略或模型改进可以进一步提高生成基因组的自然度。

总结：

- **Evo 2 在多个生物学领域的生成能力：**
  Evo 2 展示了在从线粒体到原核生物，再到真核生物基因组的生成能力，能够生成包括编码和非编码元素的基因组序列，这些生成的基因序列在结构和序列上与天然序列高度相似。

- **生成的序列质量：**
  Evo 2 成功生成了具有功能性基因、tRNA、rRNA 和合理基因排列的基因组，并能够捕捉天然基因组中的多样性和进化特征。这表明 Evo 2 在生成生物学序列时具备高度的准确性和多样性，具有巨大的应用潜力，特别是在基因组注释和基因设计等方面。

Evo 2 在表观基因组学中的生成能力总结：

验证方法：

- **设计目标与任务：**
  研究的目标是使用 **Evo 2** 生成 DNA 序列，同时控制其在基因组中的表观遗传特征，特别是 **染色质可及性**。染色质可及性决定了哪些 DNA 区域能够被转录机械访问，是基因表达调控的一个重要方面。为了实现这一目标，研究者结合了 **Enformer** 和 **Borzoi** 模型来指导 Evo 2 的生成，优化了生成序列的染色质可及性。

- **生成过程：**
  通过 **Evo 2** 生成 DNA 序列，采用 **Enformer** 和 **Borzoi** 来预测染色质可及性。生成过程中，使用了 **束搜索（beam search）** 来优化设计，即每次生成 128-bp 长度的片段并进行评分，选择最符合目标的片段继续生成。评分标准基于 **Enformer** 和 **Borzoi** 模型的染色质可及性预测结果。

- **评估与度量：**
  研究者使用 **AUROC**（接收者操作特征曲线下面积）来评估生成序列与目标染色质可及性模式的匹配度。通过增加束搜索的宽度（即每次评估更多的片段），显著提高了设计成功率，并且实现了约 0.9 的 AUROC。

验证结果：

- **染色质可及性优化：**
  通过 **Evo 2** 生成的序列成功地匹配了预期的染色质可及性模式，表明 **Evo 2** 能够在生成 DNA 序列时有效地控制染色质的开放性和闭合性。随着束搜索宽度的增加，设计质量逐步提升，生成的序列展现出清晰的染色质高可及性峰值，正好位于设计模式所指定的区域。

- **消息编码与设计成功：**
  研究者还展示了使用 **Evo 2** 来生成 Morse 代码消息（例如，“LO”、“ARC”和“EVO2”），这些设计任务的成功率也很高。通过设计不同长度和位置的染色质可及性峰值，研究者能够将简单的消息编码成 DNA 序列。

- **自然基因组特征：**
  生成的序列保持了与参考基因组（如小鼠基因组）类似的二核苷酸频率分布，显示出自然基因组特征。与使用统一提议生成的序列相比，**Evo 2** 生成的序列不仅在染色质可及性预测上有较好的共识，还保持了更自然的二核苷酸分布。

- **性能改进：**
  通过束搜索和使用生成模型的组合，研究者成功地提升了在复杂设计任务中的表现，证明了 **Evo 2** 能够通过指导性评分函数高效地从功能复杂的序列空间中采样，并生成符合预期的序列。

总结：

- **Evo 2 的生成与表观基因组调控：**
  本研究展示了 **Evo 2** 在表观基因组学中的生成能力，能够通过结合其他预测模型（如 **Enformer** 和 **Borzoi**）来控制染色质可及性，并生成具有特定染色质特征的 DNA 序列。

- **控制生成与设计：**
  研究表明，通过合理调整束搜索宽度和评分函数，**Evo 2** 可以高效地生成符合复杂设计目标的 DNA 序列，这为未来在基因组设计和表观基因组学应用中提供了新的工具。

- **广泛应用潜力：**
  这种生成方法不仅限于染色质可及性设计，其他生物结构或功能预测模型也可以与 **Evo 2** 相结合，用于指导 DNA 序列生成，实现更广泛的生物设计应用。

专业术语

likelihood of sequences

在这里，likelihood of sequences（序列的似然）指的是模型对给定生物序列的概率估计，即模型预测该序列在训练数据中出现的可能性。对于Evo 2这样的序列模型，"似然"描述了模型对某个特定序列（例如DNA、RNA或蛋白质序列）的生成概率，或者说它认为该序列是由自然进化过程生成的可能性有多大。

具体来说，Evo 2是通过训练大规模的进化数据集来学习不同生物序列的概率分布，基于这些数据，模型可以评估一个特定突变（如单核苷酸变异，SNV）如何影响序列的整体“似然”。这意味着模型能够预测突变后的序列相较于原始序列的可能性，反映了突变对该序列功能或结构的潜在影响。

delta likelihood（似然差异）

指的是模型计算出的突变序列（SNVs）与其对应的野生型序列在似然上的差异。具体来说，作者通过以下步骤计算delta likelihood：

突变引入：从每个物种中随机选取了1,024个蛋白质编码基因（其中N. equitans选取了所有536个基因），并在这些基因的起始密码子周围（-20nt到+20nt范围内）引入了单核苷酸变异（SNVs）。每个突变位置都用三个替代碱基之一来替代野生型碱基。

计算似然：对于每个突变位点，作者使用Evo 2模型计算野生型序列和对应的突变序列的似然。这里的似然是指Evo 2模型预测某个序列（包括其8,192nt的基因组上下文窗口）出现的概率。

计算似然差异：delta likelihood是通过比较突变序列与其野生型序列的似然差异来定义的。也就是说，delta likelihood是Evo 2模型在输入了突变序列后，预测该突变序列相对于其野生型序列的似然变化。

平均化：这些delta likelihood在每个位置上对1,024个基因进行平均，并且针对每个物种进行处理。这样可以得到每个突变位置的平均似然差异。

直观理解： Likelihood表示一个序列在Evo 2模型中的生成概率或被模型认为符合进化模式的可能性。 Delta likelihood表示在突变引入后的序列相对于原始序列的似然变化，即突变对序列"自然性"或"功能"的影响。如果delta likelihood值很大，意味着突变对序列的影响显著，可能改变该序列的生物学功能。

zero-shot prediction

是一种机器学习方法，它指的是模型能够在没有专门针对某一任务进行微调或监督的情况下，直接对新任务或新数据进行预测。这种方法的核心思想是通过学习到的通用知识来处理以前未见过的数据或任务，而无需专门的训练。

在传统的机器学习方法中，模型通常需要通过大量的标注数据进行训练，然后才能应用到特定的任务上。然而，zero-shot学习则希望模型能够利用在某一领域或任务上获得的知识，在没有接收到新任务的训练数据时，依然能够有效地做出预测。

在生物学中的zero-shot预测： 在文章中提到的zero-shot mutational effect prediction（zero-shot突变效应预测）是指Evo 2这样的模型能够在没有任何针对特定突变效应的监督数据或微调的情况下，通过学习进化序列数据的概率分布，直接预测突变对生物功能的潜在影响。也就是说，模型在未见过某一特定突变的情况下，依然能够基于已有的序列学习结果，推测该突变的效应。

Zero-shot学习的关键特性： 没有针对性训练数据：模型可以处理没有显式标注或监督的任务。例如，在生物序列预测中，模型没有看到具体突变的标签，但它能基于已知的序列特征进行预测。

依赖通用知识：zero-shot学习通常依赖于模型在大规模数据集上学到的通用知识或模式。例如，Evo 2在不同生物物种的序列数据上训练，从而能够学习到普遍的序列规律，应用到未知的突变预测中。

灵活性和广泛适应性：zero-shot学习使得模型在面对未知任务或数据时具有更大的适应性，能够解决许多不同类型的问题，而不需要为每个新任务单独训练。

举个例子： 假设你训练了一个生物序列模型，学到了DNA、RNA和蛋白质序列的基本结构和规律。这个模型可以用来预测某个未知突变（例如在某个基因的起始密码子处的突变）对生物体的影响，而不需要专门为这个突变进行训练。它只需要基于它之前在其他序列上的学习经验来进行预测，这就是典型的zero-shot方法。

总结来说，zero-shot prediction是一种使模型能够在没有见过特定任务的情况下，通过学习到的通用知识进行有效预测的方法。这种方法在很多领域中，特别是生物学、自然语言处理等领域，都有着非常广泛的应用潜力。

embedding（嵌入）

在这里，**embedding** 是一个机器学习和自然语言处理中的术语，指的是将高维度的、复杂的、通常是稀疏的数据（比如基因序列）转换为更低维、更密集、更容易处理的数值表示（向量）。简而言之，它就是一种把信息压缩成"数字化"表示的方式，便于机器理解和处理。

- 具体到这个问题：

在Evo 2中，**embedding** 指的是模型将基因序列（例如BRCA1或BRCA2的基因序列）转换为一组数值向量。每个基因序列（如DNA序列）可以看作是由许多不同的碱基组成（比如A、T、C、G），这些序列包含了复杂的信息，机器需要通过某种方式"理解"这些信息。

通过**embedding**，Evo 2能够将这些复杂的基因序列转化为一组低维的数值向量（即嵌入向量）。这些向量能够捕捉基因序列的语法结构、模式、变异位置等关键信息。模型通过这种嵌入向量来做出决策，比如判断一个基因变异是否会对健康产生影响。

- 举个简单的例子：

1. **基因序列**：假设我们有一个基因序列"ATCGGCTA"。
2. **embedding的作用**：通过Evo 2，模型把这个基因序列转化为一个数值向量，比如： [[0.12, -0.25, 0.44, 0.78, -0.65, 0.99, -0.31, 0.11, ...]] 这些数值并不直接对应于碱基A、T、C、G本身，而是通过模型学习到的方式来表示这些碱基在整个基因序列中的关系和功能。它们是对基因序列的一种"压缩表达"。
3. **应用：** 通过这种方式，模型能够计算和比较不同基因序列之间的相似性，判断哪些变异对健康可能有影响。这些嵌入向量为机器提供了一种"理解"基因序列的方式，让它能够基于已有的知识做出准确的预测。

- 为什么要使用embedding?
- **减少维度**：基因序列可能非常复杂且冗长，而embedding将其转化为更简洁、更便于处理的表示。
- **捕捉复杂模式**：Embedding使得模型能够捕捉到基因序列中的隐含规律，比如特定的碱基组合可能与某种疾病相关。
- **提高效率**：通过embedding，机器可以更高效地处理基因数据，减少了计算的复杂性。
- 总结：

Embedding 就是将基因序列这样的复杂信息转化为数值向量，使得机器可以理解这些序列的内在规律和结构。这种转化帮助模型更准确地预测基因变异对健康的影响，是Evo 2进行零-shot预测和有监督分类时的关键工具。

稀疏自编码器（Sparse Autoencoders, SAEs）

稀疏自编码器（Sparse Autoencoders, SAEs） 是一种神经网络架构，属于自编码器的一种变体。自编码器（Autoencoder）是一种无监督学习模型，通常由两部分组成：**编码器（Encoder）和解码器（Decoder）**。编码器将输入数据压缩成低维的表示（也叫"隐变量"），而解码器则尝试从这些低维表示中重建输入数据。

在传统的自编码器中，目标是通过最小化输入和输出之间的重建误差来训练模型。然而，**稀疏自编码器（SAE）** 通过加入一种正则化机制，鼓励网络在学习时产生稀疏的表示。也就是说，SAE 会强制模型的隐层中只有少数的神经元在激活（即为非零值），这使得模型能够学习到更加有效和稀疏的特征表示。

- 稀疏自编码器的工作原理：

1. **编码器**：将输入数据压缩成一个较低维度的表示，捕捉数据中的重要特征。
2. **正则化**：通过在训练过程中引入**稀疏性约束**，使得编码器的隐藏层只有少数神经元被激活。这种约束可以通过不同方式实现，例如通过L1正则化（促使激活的神经元的输出值接近零）或KL散度（鼓励神经元的激活概率低）。
3. **解码器**：尝试从低维表示中重建原始输入数据。

这种稀疏性约束有助于模型学到更加紧凑且具可解释性的特征，使得学习到的表示更加具有生物学或物理学的意义。

- 举个简单的例子：

假设我们有一组图像数据，目标是将这些图像压缩成较低维度的表示。

1. **普通自编码器**：学习到的表示可能是一个高维的压缩表示，它能够重建原始图像，但这些表示可能不易解释或冗余。
2. **稀疏自编码器**：学习到的表示会强制仅用少数的神经元来表示图像中的关键信息。例如，对于一张包含很多物体的图像，稀疏自编码器可能会选择一个神经元表示"有动物"或"有汽车"，而其他神经元则保持不激活（值为零），从而生成一个稀疏的表示。

- 示例应用：

1. **图像去噪**：稀疏自编码器可以通过学习图像的稀疏表示来去除图像中的噪声。
2. **特征提取**：在基因组数据或基因序列分析中，SAE可以帮助提取基因组数据中的稀疏特征，从而减少冗余信息并增强模型的泛化能力。
3. **异常检测**：稀疏自编码器在异常检测中也很有用，因为它能够学习到常见数据的稀疏表示，当输入数据与学习到的稀疏表示不匹配时，模型可以识别出异常数据。

- 总结： **稀疏自编码器（SAE）** 通过强制模型学习到稀疏的表示，使得网络能够更高效地捕捉到数据中最重要的特征，并且这些特征通常更具可解释性。在处理复杂数据（如基因组数据、图像数据等）时，SAE能够帮助提取出更有意义、简洁的特征表示。

Enformer 和 Borzoi 解释：

Enformer 和 **Borzoi** 都是用于预测 **染色质可及性** 的模型，它们通过从 DNA 序列中学习，能够预测染色质区域的开放性或闭合性，这对于理解基因的表达调控和表观遗传学至关重要。

1. **Enformer：**
    - **Enformer** 是一种基于深度学习的模型，专门设计用于预测 **染色质可及性**，即基因组中哪些区域易于被转录因子和其他转录机械访问。Enformer 使用了序列到功能的转换方法，通过处理 DNA 序列，学习并预测染色质是否在某些细胞类型中处于开放状态。Enformer 是一个强大的工具，尤其在预测 **染色质可及性** 和其他表观遗传学特征方面，展现了高度的准确性。
    - 它的核心功能是**序列到表观遗传功能的映射**，在基因组学研究中，Enformer 能够帮助研究人员预测不同 DNA 序列在细胞中的实际功能和调控潜力。
2. **Borzoi：**
    - **Borzoi** 是另一个专注于基因组功能预测的模型，尤其是染色质可及性方面。与 **Enformer** 类似，Borzoi 通过深度学习来预测染色质的开放性，帮助识别 DNA 区域是否可能是转录因子和其他调控

因子作用的目标。Borzoi 是一个由多个模型组成的集成模型（ensemble model），通常会结合多个不同的深度学习模型来获得更强的预测性能。
    - Borzoi 也用于基因组学中，通过分析染色质的开放性来帮助理解哪些 DNA 区域与基因表达或基因调控活动相关。

Beam Search 解释：

束搜索（Beam Search） 是一种启发式搜索算法，广泛应用于序列生成任务（如自然语言处理、图像生成、DNA 序列生成等）。与传统的广度优先或深度优先搜索不同，束搜索在每个生成步骤中并不探索所有可能的路径，而是限制探索的路径数量，只选择最佳的候选序列进行扩展。这使得束搜索在生成序列时更加高效。

- 工作原理：

1. **初始化：** 从一个起始状态（例如，给定的 DNA 序列的前缀）开始，束搜索会考虑多个可能的下一步（即可能的序列扩展）。这些候选序列会被评估，并根据预定义的评估标准（如概率、模型得分等）进行排序。

2. **扩展：** 在每个生成步骤中，束搜索会选择一定数量（称为"束宽度"）最有前途的候选序列进行扩展。每扩展一次，就会产生更多的候选序列，而束搜索会限制只保留一定数量的候选序列。

3. **选择最优：** 在每个步骤中，束搜索根据得分选择最佳的候选序列，并继续扩展，直到生成整个序列或达到停止条件。

4. **优势：** 相比于暴力搜索，束搜索减少了计算量，同时保持了较高的生成质量。通过调整"束宽度"，可以平衡计算效率与生成质量。

- 例如： 假设我们正在使用束搜索生成一个 DNA 序列，起始部分是已知的"ATG"。在每个步骤中，束搜索会探索可能的下一步（如添加"A"、"C"、"G" 或 "T"）。然后，根据模型的评分函数（例如，DNA 序列的自然性或染色质可及性），选择得分最高的几个候选序列，继续扩展，直到生成完整的序列。

在 **Evo 2** 的应用中，束搜索被用来优化 DNA 序列的生成，特别是在染色质可及性设计的任务中，通过选择最符合目标染色质模式的序列，并有效提高了设计的成功率和生成序列的质量。