Abstract

信使 RNA（mRNA）疫苗被用于对抗 COVID-19 的传播（参考文献 1–3），但它们仍然存在由 mRNA 不稳定性和降解引起的关键限制，这些是疫苗产品储存、分发和效力的主要障碍[4]。通过增加二级结构可以延长 mRNA 的半衰期，这与优化的密码子一起提高了蛋白质表达[5]。因此，一个基于原理的 mRNA 设计算法必须同时优化结构稳定性和密码子使用。然而，由于同义密码子的存在，mRNA 的设计空间极其庞大——例如，对于 SARS-CoV-2 刺突蛋白，大约有 2.4×10^{632} 种候选 mRNA 序列。这带来了难以克服的计算挑战。

在此，我们提供了一种简单且意料之外的解决方案，使用计算语言学中的经典概念——格解析（lattice parsing），其中寻找最佳 mRNA 序列类似于在相似的句子中识别最有可能的句子[6]。我们的算法 LinearDesign 在仅 11 分钟内为刺突蛋白找到了最优的 mRNA 设计，并能够同时优化稳定性和密码子使用。LinearDesign 显著改善了 mRNA 的半衰期和蛋白质表达，并在小鼠体内相比于 COVID-19 和水痘-带状疱疹病毒的 mRNA 疫苗密码子优化基准，抗体效价提高了高达 128 倍。这一结果揭示了基于原理的 mRNA 设计的巨大潜力，并启发了对以前难以实现但高度稳定和高效设计的探索。我们的工作是疫苗及其他编码治疗性蛋白（如单克隆抗体和抗癌药物）的 mRNA 药物的及时工具[7,8]。

1 Introduction

mRNA 疫苗[9,10] 已被认可为一种可行的工具，能够通过其可扩展的生产、安全性和效力来限制 COVID-19 的传播[1–3]。然而，mRNA 分子化学上不稳定且易于降解，这导致蛋白质表达不足[5]，进而削弱免疫原性和药物可及性。这种不稳定性也成为疫苗储存和分发的主要障碍，需依赖冷链技术，这限制了其在发展中国家的使用[4]。因此，具有增强稳定性的 mRNA 分子是可取的，这将有可能具有更高的效力和良好的临床疗效。

尽管化学稳定性的建模仍然困难，以往研究已确立其与 RNA 二级结构之间的相关性，这可通过热力学折叠稳定性来量化。提高这种结构稳定性，再结合优化的密码子使用，将有助于提高蛋白质表达[5]。因此，一个基于原理的 mRNA 设计算法必须同时优化两个因素——结构稳定性和密码子使用——以增强蛋白质表达。

然而，mRNA 设计问题（本文仅考虑编码区域）极具挑战性，因为其搜索空间呈指数级增长。每个氨基酸由三联密码子编码，即三个相邻的核苷酸——但由于遗传密码的冗余，大多数氨基酸具有多个密码子；对于 20 种常见的氨基酸，共有 4^3（即 64）种密码子。这导致了任何蛋白质序列的候选 mRNA 序列数量极其庞大。例如，SARS-CoV-2 的刺突蛋白有 1,273 个氨基酸，因此可以由大约 2.4×10^{632} 种 mRNA 序列编码（图 1a）。这带来了难以克服的计算挑战，排除了枚举的可能性——若要为刺突蛋白进行枚举计算，则需要 10^{616} 亿年（图 1b）。

相比之下，传统的 mRNA 设计方法通过密码子优化来提高蛋白质表达[11,12]，但对稳定性的提升效果有限，从而忽略了高度稳定 mRNA 的潜在设计空间。优化 GC 含量的效果相似，因为它与脊椎动物的密码子使用相关[13]。因此，大多数高度稳定的设计仍未被探索。

在此，我们介绍了 LinearDesign，这是一种通过适应计算语言学中格解析[6] 经典概念来解决这一挑战的算法（图 1c）。我们展示了在庞大的候选空间中找到最佳 mRNA 类似于在众多相似句子中找到最可能的句子。更具体地，我们使用确定性有限自动机（DFA）来定义 mRNA 设计空间，类似于词格[6]，其紧凑地编码了指数多的 mRNA 候选序列。然后，我们使用格解析在 DFA 中找到最稳定的 mRNA，或在加权 DFA 中找到稳定性和密码子优化之间的最佳平衡。

这种与自然语言解析的意外联系提供了一种有效的算法，其计算复杂度与 mRNA 序列长度成平方关系，并在实践中可扩展。通过这种方式，我们将 mRNA 设计的巨大设计空间转化为优势，而不是障碍。

与密码子优化基准相比，我们的 COVID-19 和水痘-带状疱疹病毒（VZV）mRNA 疫苗显著改善了体外的化学稳定性、细胞中的蛋白质表达以及体内的免疫原性。尤其是，我们的 COVID-19 疫苗在小鼠中将抗体反应提高了 128 倍。这一结果揭示了基于原理的 mRNA 设计的巨大潜力，并启发了对以前难以实现的高度稳定和高效设计的探索。我们的工作为 mRNA 疫苗及其他基于 mRNA 的药物（如编码治疗性蛋白的单克隆抗体和抗癌药物[7,8,14]）的设计提供了一种及时且有前景的工具。

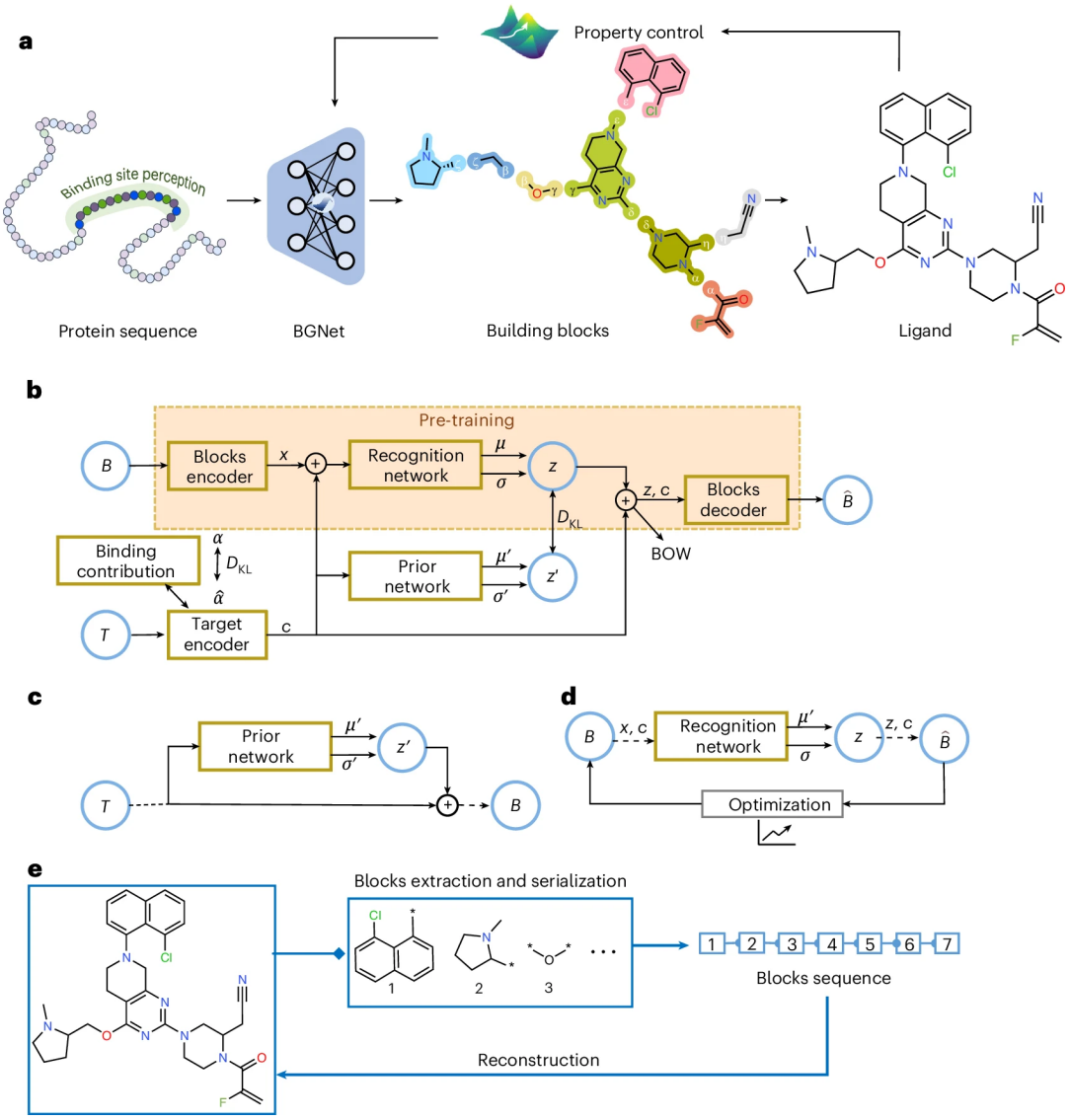


图 1: **Fig. 1 | 针对稳定性和密码子优化的 mRNA 编码设计概述，以 SARS-CoV-2 刺突蛋白为例。** **a,** 由于密码子简并性和组合爆炸，对于编码刺突蛋白的大约有 2.4×10^{632} 种可能的 mRNA 序列。枚举每一个可能的序列需要 10^{616} 亿年。粉色和蓝色路径分别代表野生型和最优稳定（最低自由能）序列的设计结果。nt 表示核苷酸。**b,** 野生型（左）和最优稳定（右）mRNA 的二级结构。野生型 mRNA 主要是单链，并易于在环区（红色）降解，而最优稳定的 mRNA 主要是双链。使用 LinearDesign 优化大约需要 11 分钟。**c,** DFA（确定性有限自动机）和格解析在计算语言学中的应用（左）及其在 mRNA 设计中的适应（右）。mRNA DFA（类似于词格）紧凑地编码了所有 mRNA 候选序列，这些序列通过格解析同时折叠以找到最优 mRNA（图 2）。**d,** mRNA 设计空间的二维可视化，稳定性（用 MFE 表示）为横轴，密码子优化（用 CAI 表示）为纵轴。标准 mRNA 设计方法通过密码子优化提高密码子使用率（粉色箭头），但无法探索高度稳定的区域（虚线左侧）；COVID-19 mRNA 疫苗（BNT-162b2, mRNA-1273 和 CVnCoV）作为示例。LinearDesign 联合优化稳定性和密码子优化（蓝色曲线，其中 λ 为分配给密码子优化的权重）。我们选择了七种 mRNA 设计（图中显示了四种 A–D）以及一个优化基线（H）进行体外和体内实验（图 4）。

2 Formulations and algorithms

先前的研究[5] 确定了 mRNA 设计的两个主要目标：稳定性和密码子优化，这两者协同作用以提高蛋白质表达。为了优化稳定性，给定蛋白质序列，我们的目标是找到在所有可能编码该蛋白的 mRNA 序列中，具有最低自由能（MFE）的 mRNA 序列。具体来说，对于每个候选 mRNA 序列，我们通过标准的 RNA 折叠能量模型[15,16] 来寻找其所有可能的二级结构，并选择其中 MFE 最低的序列。这本质上是一个最小化中的最小化问题（扩展数据图 1a）。这种方法需要数十亿年的时间，因此需要一种无需枚举的高效算法。

我们还希望联合优化 mRNA 的稳定性和密码子优化。密码子优化通常通过密码子适应指数（CAI）[17] 来衡量，该指数定义为 mRNA 中每个密码子的相对适应性的几何平均。由于 CAI 在 0 到 1 之间，而 MFE 通常与序列长度成正比，我们将 CAI 乘以 mRNA 中的密码子数，并使用超参数 CAI 权重（λ）来平衡 MFE 和 CAI（$\lambda = 0$ 时仅优化 MFE）。组合目标函数定义为 $MFE - \lambda|p| \log CAI$，其中 $|p|$ 是蛋白质序列的长度。详情参见“优化选择”及扩展数据图 1b。

接下来，我们通过借用自然语言中的两个思想来解决这两个优化问题：DFA（格）表示和格解析。

2.1 用于 mRNA 设计空间的格表示

受计算语言学中词格表示模糊性的启发（扩展数据图 2a），我们使用类似的方式来表示每个氨基酸的选择——更正式地说，一个 DFA，它是带有核苷酸标签边的有向图（参见图 2a 和图 1c；详情见方法部分“DFA 表示与密码子及 mRNA 候选序列”）。在为蛋白质序列中的每个氨基酸构建 DFA 后，将它们合并为一个单一的 mRNA DFA，其中每一条路径代表编码该蛋白的 mRNA 序列（图 2b 和扩展数据图 1d）。

2.2 格解析

RNA 折叠已被认为等同于自然语言解析，其中随机上下文无关文法（SCFG）可以表示折叠能量模型[18]（扩展数据图 1e, f）。对于 mRNA 设计，难题在于如何同时折叠 DFA 中的所有 mRNA 序列。我们借鉴了格解析的思想[6,19]，将单序列解析推广到处理格中的所有句子，并同时找到最可能的一个（图 1c 和扩展数据图 2）。同样，我们使用格解析来折叠 DFA 中的所有 mRNA 序列，以找到最稳定的序列（图 2b 和扩展数据图 1g, h）。需要注意的是，格解析也是动态规划的一个实例，但在更大的搜索空间上进行操作，而单序列折叠可以被视为单链 DFA 格解析的特例。该过程还可以解释为 SCFG 与 DFA 的交集（扩展数据图 1a），其中 SCFG 用于稳定性评分，而 DFA 则界定候选集。该算法的运行时间随着 mRNA 序列长度按三次方缩放（方法部分“SCFG、格解析和交集”），但在实际应用中按平方缩放（图 3a）。

2.3 带权重 DFA 的格解析

我们现在将 DFA 扩展为带权重的 DFA，以集成密码子优化的边缘权重。由于我们的联合优化公式将 CAI 纳入每个密码子的相对适应性 $w(c)$，因此我们设置每个密码子 DFA 中的权重，使得密码子路径成本为 $-\log w(c)$，这可以解释为从最优路径的“偏差量”。在加权 mRNA DFA 中，起点和终点路径的权重为每个密码子的 $-\log w(c)$ 之和，$w(c)$ 是对应密码子的频率权重（图 2d）。新的格解析使用随机文法（用于稳定性）和加权 DFA（用于密码子使用）解决联合优化问题，同时具有最优性保证，可视为加权 SCFG 与加权 DFA 的交集（扩展数据图 1b 及方法部分“CAI 整合的加权 DFA”）。

a

	Reference	LiGAN	3D-SBDD	Pocket2Mol	TargetDiff	DeepBlock
Vina score (↓)	−7.445 ± 2.276	−6.142 ± 0.847	−6.928 ± 1.558	−7.221 ± 2.132	**−7.284 ± 1.418**	−7.224 ± 0.971
QED (↑)	0.476 ± 0.206	0.460 ± 0.069	0.508 ± 0.120	**0.585 ± 0.098**	0.483 ± 0.108	0.543 ± 0.077
SA (↓)	3.453 ± 1.259	4.798 ± 0.278	4.304 ± 0.857	3.194 ± 1.172	4.546 ± 0.143	**3.186 ± 0.357**
Retro* (%, ↑)	58.00 ± 49.60	7.69 ± 5.13	21.88 ± 19.32	54.29 ± 33.43	22.02 ± 22.99	**60.27 ± 13.90**
Diversity (↑)	–	**0.834 ± 0.031**	0.697 ± 0.115	0.727 ± 0.161	0.742 ± 0.093	0.784 ± 0.071
log*P* (–)	0.894 ± 2.742	0.595 ± 0.766	0.406 ± 1.591	2.029 ± 1.477	1.768 ± 1.378	**1.112 ± 1.244**
Fragment similarity (↑)	–	0.109 ± 0.124	0.176 ± 0.221	0.116 ± 0.144	0.173 ± 0.177	**0.205 ± 0.231**
Scaffold similarity (↑)	–	0.152 ± 0.101	**0.221 ± 0.142**	0.154 ± 0.140	0.147 ± 0.131	0.144 ± 0.217
Time (s, ↓)	–	1,741	3,139	1,962	2,558	9

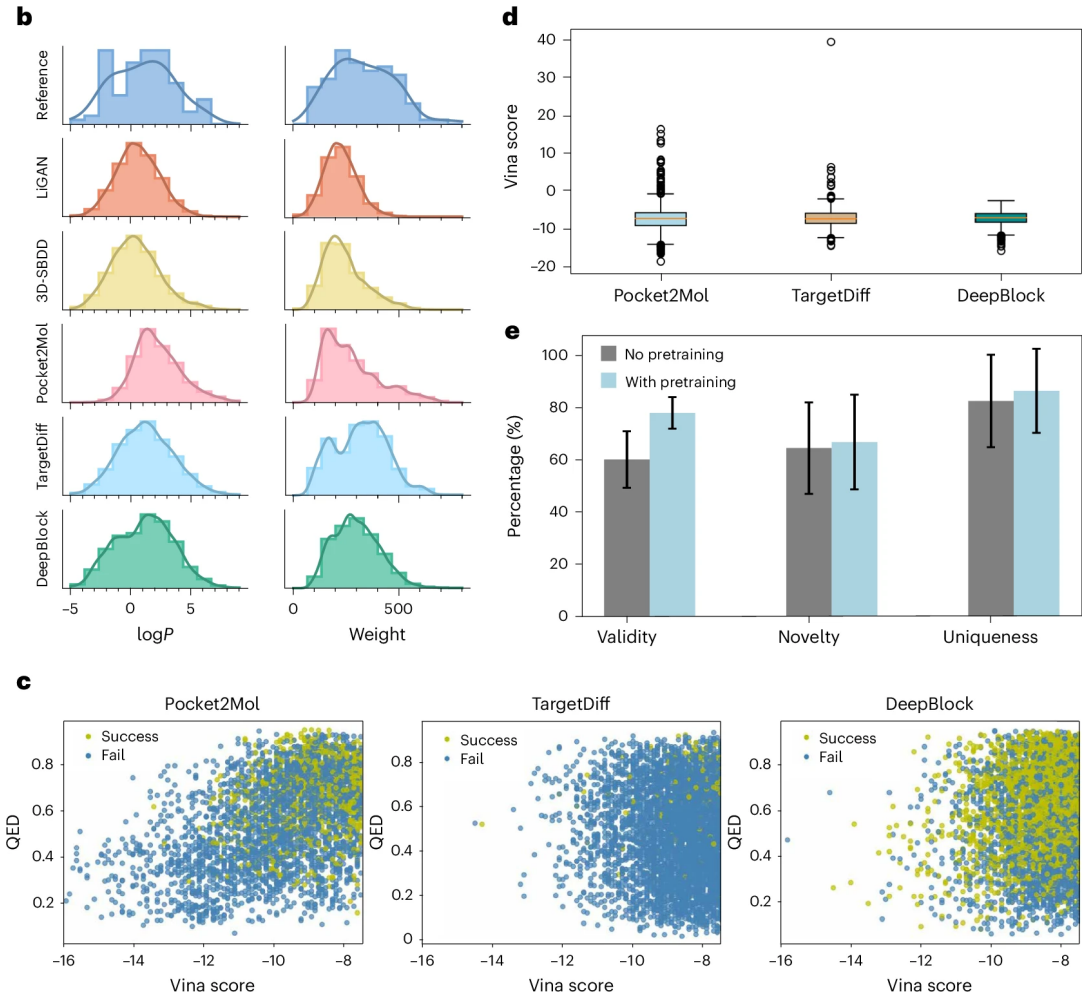


图 2: **LinearDesign 算法的说明。 a,** 密码子 DFA 的示意图。**b,** 一个 mRNA DFA（下方）及其上方的格解析。在 DFA 中，基于简化能量模型的最优 mRNA 序列显示为蓝色路径，同时以点-括号格式展示其最优结构（点表示未配对，括号表示碱基配对）。在格解析中，棕色和黑色弧线也表示碱基配对（两个 GC 对和两个 AU 对），梯形阴影区域表示最优结构的分解。在 DFA 中编码的所有 mRNA 序列中，格解析找到与其最优结构匹配的最优序列，并在该能量模型下达到最低自由能，其中 GC 和 AU 对的能量分别为 -3 和 -2 kcal mol^{-1}（扩展数据图 1e）。注意，此处使用的是简化的能量模型作为示例，而我们的实现使用最近邻能量模型。**c,** 稳定性和密码子优化联合优化的另一示意图。**d,** 通过在带权重的 DFA 中集成密码子优化来优化序列和二级结构。顶部显示苏氨酸和丝氨酸的密码子频率。密码子的相对适应性 $w(c)$ 是密码子 c 的频率与编码相同氨基酸的最常见密码子（白色条）的频率之比，其值显示在条形图右侧。底部，一个带权重的 mRNA DFA 通过使用 $-\log[w(c)]$ 作为边缘权重来编码每个候选序列的 CAI（选择密码子的成本），此带权重 DFA 作为输入，用于在格解析中联合优化稳定性和密码子优化。

2.4 DFA 的表达能力

我们的 DFA 框架足够通用，甚至可以表示替代遗传密码、修饰核苷酸和密码子约束。详情请参见方法部分 “DFA 用于其他遗传密码、密码子约束和修饰核苷酸”、扩展数据图 3 及补充图 5。

2.5 线性时间近似

虽然精确设计算法可能对于较长序列仍然较慢，但由于实验室实验涉及的因素很多，因此次优设计同样值得探索。为此，我们开发了一种近似搜索版本，其通过束搜索在线性时间内运行，每步仅保留前 b 个最有前途的项目（b 为束的大小），灵感来自我们之前的 LinearFold 算法[21]。

2.6 相关工作

此前的两项研究也通过动态规划解决了 “最稳定 mRNA 设计”（我们的目标 1）的问题[22,23]，但使用的是 Zuker 算法的特定扩展，无法同时优化密码子使用（目标 2）。相比之下，我们通过建立 mRNA 设计与计算语言学格解析的联系，提出了一种更简单且更具泛化能力的算法，可以联合优化密码子使用，并引入一个新的目标函数，将 CAI 整合到单个密码子中。我们还验证了这些算法在体内和体外的有效性，为两种 mRNA 疫苗（图 4 和图 5）提供了实验证据。详情参见方法部分 “LinearDesign 算法” 和 “相关工作”。

3 计算结果与分析

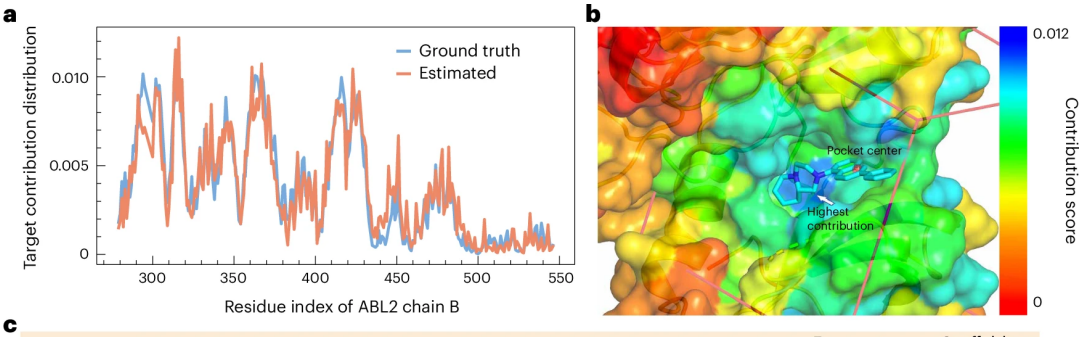
图 3a 展示了 LinearDesign 在 UniProt 蛋白上的运行时间基准测试[24]。LinearDesign 在两种优化目标的组合下进行了测试：仅 MFE（目标 1）和联合 MFE 与 CAI（目标 1 和 2），并通过两种搜索模式进行测试：精确搜索与束搜索（$b = 500$）。经验结果表明，LinearDesign 在实际应用中，随着 mRNA 序列长度 n 呈二次方缩放（$n < 10,000$ nt），这得益于 DFA 表示和格解析（补充图 7 和 8）。接下来，我们的 CAI 集成精确搜索（$\lambda = 4$）具有相同的经验复杂度，并且仅比仅 MFE 版本慢约 15%，这要归功于 DFA 在添加 CAI 时的便利性。最后，我们的束搜索版本（$b = 500$）进一步加快了设计速度，并随着序列长度呈线性缩放，在 SARS-CoV-2 刺突蛋白上仅需 2.7 分钟（而精确搜索则需 10.7 分钟），且近似误差（百分比能量差距，定义为 $(1 - \frac{\mathrm{MFE}_{\mathrm{approx\_design}}}{\mathrm{MFE}_{\mathrm{exact\_design}}}) \times 100\%$）为 1.2%。事实上，随着序列变长，该百分比趋于稳定，表明束搜索质量不会因序列长度而降低（补充图 9）。

对于倾向于 GC 的密码子偏好（如在人类中），传统的密码子优化方法确实改善了稳定性，但主要是正交于密码子优化方向（粉色箭头）（图 3b, c）。相比之下，我们的 LinearDesign 可以直接优化稳定性并找到最稳定的 mRNA。在 SARS-CoV-2 刺突蛋白和 VZV gE 蛋白上，最低 MFE（$\lambda = 0$）比常规密码子优化方法低 1.8 倍。

我们的最优稳定设计主要为双链二级结构（图 3d），预测其降解风险较低[5]。通过将 λ 从 0 变化到 ∞，LinearDesign 计算了 mRNA 设计空间的可行性极限（图 3b, c 中的蓝色曲线；参见扩展数据图 4）。此外，当密码子偏向 AU 富集（如在酵母中）时，密码子优化实际上会降低稳定性（扩展数据图 4b）。

4 COVID-19 mRNA 疫苗的结果

我们在本研究中检查了 SARS-CoV-2 刺突蛋白的 mRNA 序列。使用 LinearDesign 算法设计了七种序列（序列 A–G），作为次优分子（使用束搜索[21,25]）。这些序列广泛分布在低 MFE 设计空间中（MFE $-1,400$ kcal mol^{-1} 的区域，如图 4a 所示），这是传统密码子优化算法无法达到的区域。为了更好地理解 MFE 和 CAI 的生物学效应，我们设计了具有几乎相同 MFE 或 CAI 值的 mRNA 序列（图 3b, c）；序列 B 和 C 具有相似的 MFE，而 D、E 和 F



c

	Vina score (↓)	QED (↑)	SA (↓)	Retro* (↑)	Diversity (↑)	Fragment similarity (↑)	Scaffold similarity (↑)
DeepBlock (predict)	−7.164 ± 0.974	0.539 ± 0.081	3.185 ± 0.394	60.29 ± 14.32	0.788 ± 0.075	0.190 ± 0.217	0.213 ± 0.160
DeepBlock (without)	−7.184 ± 0.942	0.539 ± 0.084	3.201 ± 0.376	60.42 ± 13.34	0.784 ± 0.079	0.194 ± 0.220	0.142 ± 0.213
DeepBlock	−7.224 ± 0.971	0.543 ± 0.077	3.186 ± 0.357	60.27 ± 13.90	0.784 ± 0.071	0.205 ± 0.231	0.144 ± 0.217

图 3: **Fig. 3 | LinearDesign 算法的计算特性。** **a,** 对 UniProt 蛋白质的 mRNA 设计的运行时间分析（补充表 1）。总体而言，我们的精确搜索随着序列长度按平方缩放（补充图 7 和 8），而我们的 MFE + CAI 模式（$\lambda = 4$）仅比仅 MFE 模式慢约 15%。此外，束搜索（$b = 500$）显著加速了长序列的设计，同时缩小了搜索误差（补充图 9）。**b, c,** SARS-CoV-2 刺突蛋白（b）和 VZV gE 蛋白（c）设计的二维（MFE–CAI）可视化，其中蓝色曲线表示可行性极限（最优边界），通过将 λ 从 0 变化到 ∞ 进行计算（参见扩展数据图 4）。GC 偏好用括号表示。人类基因组偏向于 GC 富集的密码子，因此密码子优化（粉色箭头）确实提高了稳定性，但幅度有限，因为密码子优化与稳定性方向大多是正交的。而对于 AU 富集的密码子偏好（如在酵母中），密码子优化反而降低了稳定性（扩展数据图 4b）。**d,** SARS-CoV-2 刺突蛋白和 VZV gE 蛋白设计的 mRNA 二级结构。最优 CAI 设计（顶部，$\lambda = \infty$）主要为单链（约 60% 碱基配对），而最优稳定设计（底部，$\lambda = 0$）主要为双链（约 80% 碱基配对）。中心显示了中间设计（$\lambda = 4$），在稳定性和 CAI 之间达到平衡。

具有相似的 MFE；A、C 和 F 具有相似的 CAI，而 B 和 E 具有相似的 CAI，D、G 和 H 具有相似的 CAI。第八种 mRNA 序列（序列 H）使用 OptimumGene 设计，该基准序列被用于 COVID-19 mRNA 疫苗，该疫苗在两种动物模型中引发了高免疫原性[26]，并已在中国进入 I 期临床试验（由中国疾病控制中心和中国临床试验登记 CTR20210542 共同开发）。所有这些 mRNA 序列均编码刺突蛋白。

4.1 相同氨基酸序列的 COVID-19 mRNA 疫苗

我们检查了全长野生型 SARS-CoV-2 刺突蛋白的 mRNA 序列，使用天然未修饰的核苷酸，并共享相同的 5' 和 3' UTR 序列（详见补充信息部分的序列）。考虑到结构化 5' 引导区可能对翻译效率的潜在负面影响[5]，我们在运行 LinearDesign 时未包括前 5 个氨基酸，而是使用启发式方法选择前 15 个核苷酸。同时有研究表明，长的螺旋结构可能会引发不必要的免疫反应[27]，因此我们在设计中避免了这些情况。这也解释了为什么我们未选择最低 MFE 的候选序列（即接近最优边界的那些蓝色曲线区域），通常包含长茎环结构（如图 4a 所示）。详见方法部分 “其他设计约束”。

4.2 UTR 结构的重要性

除了编码区域设计外，UTR（非翻译区）结构对翻译也至关重要[28]，并且 UTR 工程对蛋白质翻译具有深远的影响。虽然 LinearDesign 不直接优化 UTR，但其设计的 mRNA 分子由于结构更为紧凑，相较于仅进行密码子优化的序列，形成更牢固的碱基配对，因此对常用 UTR 的结构干扰较小（扩展数据表 1）。这一点在我们对 VZV mRNA 疫

苗的不同 UTR 对的实验中得到验证（扩展数据表 2），这些 UTR 导致了蛋白质表达和免疫反应的显著提升（图 5）。这表明，LinearDesign 在不同 UTR 选择下仍然具有稳健性，这与最近的研究一致[29]，该研究表明 LinearDesign 生成的序列在多种 UTR 下在体外蛋白表达上均优于基准序列（参见图 4a 和参考文献 29）；详见方法部分“相关工作”。

4.3 溶液结构紧凑性与化学稳定性

我们还研究了 mRNA 分子的结构紧凑性与化学稳定性，这被假设与折叠自由能变化相关。MFE 较低的 mRNA 分子往往包含更多的二级结构，表现出更紧凑的形状，并具有更小的水动力学尺寸，从而在更高的电泳迁移率下表现出更高的稳定性。我们将 mRNA 样本加载到非变性琼脂糖凝胶上，并发现 mRNA 序列 A–H 的化学稳定性和蛋白质表达。图 4b 显示了具有相似分子量的序列 A–H 的迁移率。序列 A（最低 MFE）表现出最高的迁移率，表明其分子结构更紧凑，而序列 H（最高 MFE）的迁移率最低。数据显示了 LinearDesign 在 MFE 计算上的有效性。为了评估 mRNA 的化学稳定性，我们将 mRNA 在 37℃ 下分别在 10 mM（图 4c）和 20 mM（补充图 5g）Mg^{2+} 缓冲液中孵育，并评估 RNA 完整性。与以往研究类似[29]，序列 A–H 展示了不同的降解速度，其与 MFE 值高度相关（图 4c 和补充图 5g）。序列 A（最低 MFE）的降解速度最慢，在 10 和 20 mM Mg^{2+} 缓冲液中的半衰期（$T_{1/2}$）分别为 12.6 和 10 小时。相比之下，序列 H（最高 MFE）的降解速度最快，$T_{1/2}$ 分别为 3.9 和 3.3 小时。这些结果表明，低 MFE 设计在溶液中更具抗降解性，是生物应用的理想选择。

4.4 细胞蛋白质表达

对于疫苗，抗原表达水平是有效免疫反应的关键决定因素。因此，我们评估了设计的 mRNA 序列在 HEK293 细胞中的转染效率。所有使用 LinearDesign 生成的 mRNA（序列 A–G）在蛋白质表达水平上显著高于基准序列 H（图 4d 和补充图 9）。序列 A 和 B 的 CAI 值接近 H，但其 MFE 更低，表现出更高的蛋白质表达水平。序列 A（最低 MFE）和序列 E（较高 CAI）展示了最高的蛋白质表达水平和细胞外分泌。这些结果与 Maurer 等人的研究一致[31]，表明更低的 MFE 和更高的 CAI 与更高的表达相关。

4.5 体内免疫原性

我们进一步测试了这些设计是否能够在体内增强免疫原性。我们将 mRNA 序列 A–H 使用脂质纳米颗粒递送[30]，并在小鼠中评估其体液和细胞免疫反应。对于每种 mRNA 序列，C57BL/6 小鼠分别肌内注射两剂疫苗（间隔两周）。评估了抗刺突 IgG、中和抗体及刺突特异性干扰素-（IFNγ）分泌 T 细胞的水平。所有由 LinearDesign 生成的 mRNA 分子均能够诱导强大的抗体反应。相比之下，序列 H mRNA 的抗体诱导能力非常有限（图 4e, f）。在抗原特异性 T 细胞反应中也观察到了类似的结果，其中仅 LinearDesign 生成的 mRNA 能够诱导强烈的 T 辅助 1 型偏向 T 细胞反应（图 4g）。序列 A–D 更接近最优边界（图 4a 中蓝色阴影区域），其抗刺突 IgG 抗体效价提高了 57 至 128 倍，中和抗体滴度提高了 9 至 20 倍，相较于基准序列 H。

由于 BNT-162b2（由 BioNTech 和 Pfizer 开发）是目前最广泛使用的 COVID-19 mRNA 疫苗，我们将其与 LinearDesign 生成的 mRNA 序列进行了比较。在这次对比中，我们的 BNT 序列几乎与 BNT-162b2 的序列相同，但有三处差异：（a）BNT-162b2 中用于稳定的两处脯氨酸突变[31] 被转换回野生型序列，（b）BNT 使用相同的 5’ 和 3’ UTR 作为序列 A–H，（c）BNT 中的核苷酸为天然未修饰。包括四种 mRNA 序列 A、C、H 和 BNT 在内的体内研究表明，序列 A 和 C 在溶液中的降解率显著低于 BNT，并且在 HEK293 细胞中表现出显著更高的蛋白质表达（扩展数据图 6）。值得注意的是，BNT 和 H 表现出类似的 MFE、CAI（图 4a）和半衰期。此外，A 和 C 能够引发比 H 和 BNT 更高水平的抗刺突 IgG 和中和抗体（扩展数据图 7）。总的来说，这些数据使我们推测 LinearDesign 优化的 mRNA 分子在体内更稳定，从而导致蛋白质表达改善和免疫原性增强。

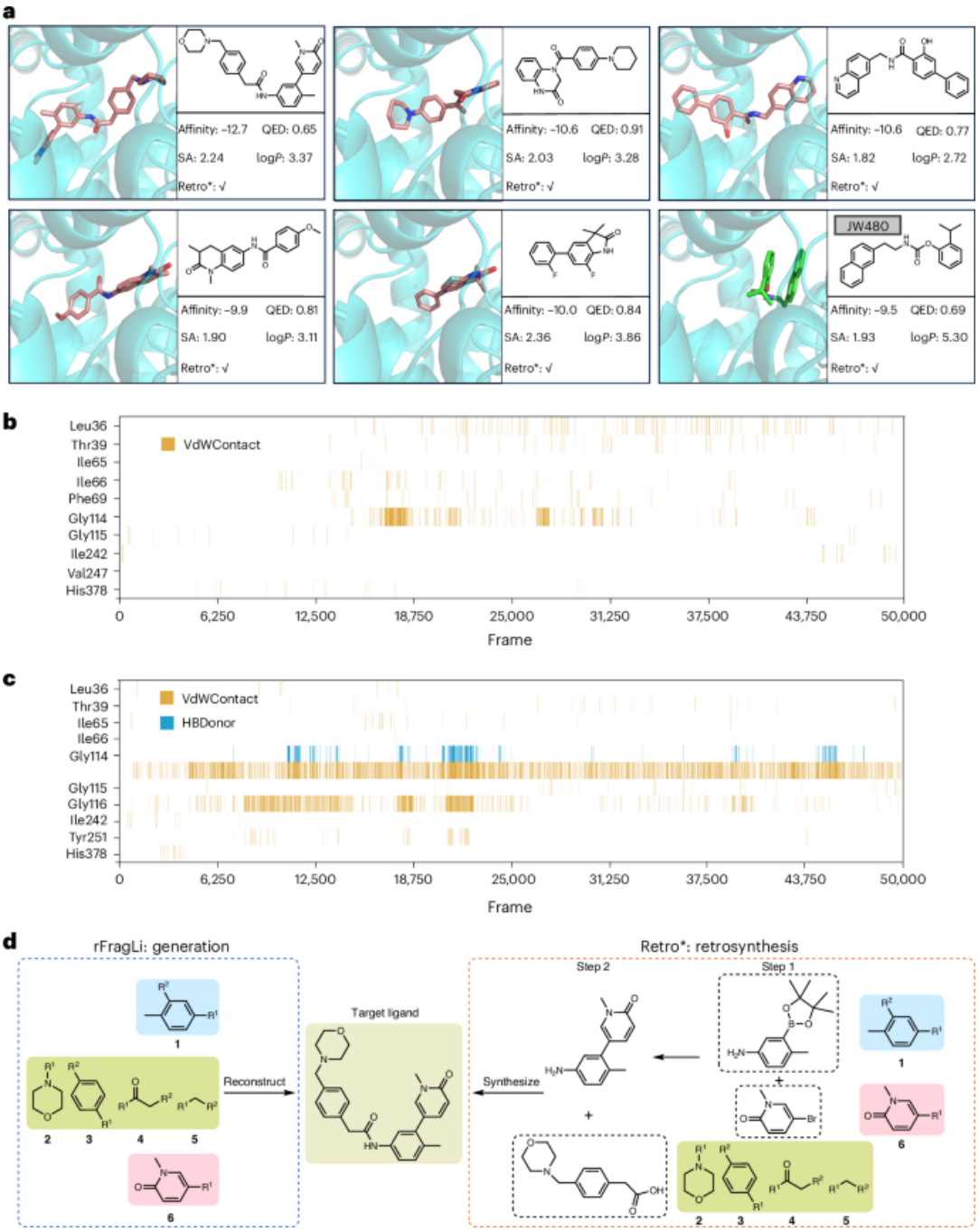
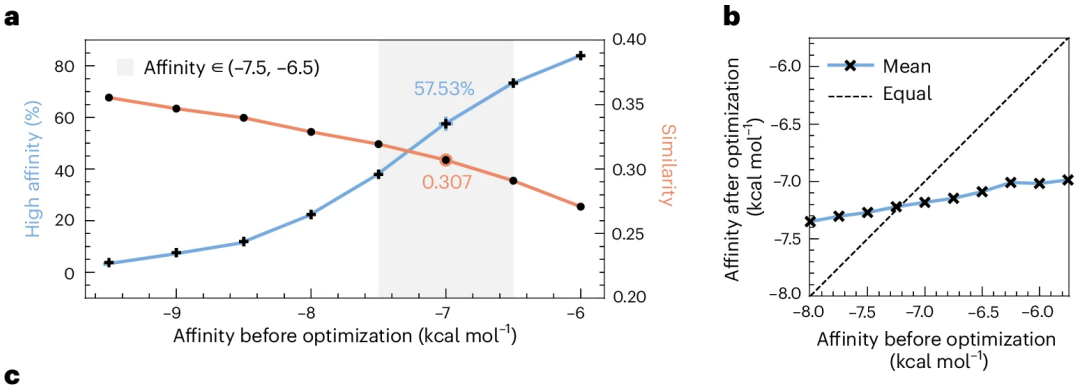


图 4: **对编码 SARS-CoV-2 刺突蛋白的 LinearDesign 生成的 mRNA 序列的实验评估。 a,** 从刺突 mRNA 设计 A–G 及相应的免疫反应（与密码子优化基准 H 相比）总结了化学稳定性和蛋白质表达。疫苗 mRNA-1273 和 BNT-162b2 使用修饰核苷酸，因此其 MFE 是使用标准能量模型计算的。b, 非变性琼脂糖凝胶电泳对 mRNA 的聚集情况显示其最低自由能的全局稳定性。有关凝胶电泳数据，参见补充图 13。**c,** 在 37℃ 下 10 mM Mg^{2+} 缓冲液中 mRNA 的化学稳定性。数据来自三次独立实验。Seq. 表示序列。**d,** 蛋白质表达水平，通过流式细胞仪测定 HEK293 细胞在转染后 48 小时内的蛋白质表达。平均荧光强度（MFI）值来源于三次独立实验。Kruskal–Wallis ANOVA 与 Dunn’ s 多重比较用于组间比较。g, C57BL/6 小鼠（$n = 6$）每隔两周肌内注射两剂制备的 mRNA 疫苗后检测抗刺突 IgG 抗体终点效价。h, 评估中和抗体对全长 SARS-CoV-2 S 蛋白的效价。i, 酶联免疫斑点（ELISPOT）分析 IFNγ 分泌 T 细胞的频率。数据表示为平均值 ± 标准差（s.d.）；P 值来源于 t 检验。* $P < 0.05$，** $P < 0.01$，*** $P < 0.001$。NS 表示无显著性差异。详见补充图 5–7 和补充表 2 以获取详细的计算和实验数据。



c

	Before optimization	After optimization	Better properties (%)
Affinity (kcal mol^{-1}, ↓)	−6.515 ± 0.518	−7.086 ± 0.827	76.04
QED (↑)	0.620 ± 0.227	0.589 ± 0.209	45.91
SA (↓)	2.602 ± 0.664	2.769 ± 0.773	45.27
log*P* (–)	3.051 ± 1.896	2.505 ± 2.049	–
Lipinski (↑)	4.586 ± 0.907	4.662 ± 0.722	85.81

图 5: **对编码 VZV gE 蛋白的 LinearDesign 生成的 mRNA 序列的实验评估。** **a,** 来自 VZV gE mRNA 设计的化学稳定性和相应免疫反应的汇总（诱导抗 gE IgG）。高亮显示的浅蓝色阴影区域表示在最低自由能下的“甜点”区域。b, 非变性琼脂糖凝胶表征 mRNA 的聚集情况，展示了与最低自由能的全局稳定性相关性。有关数据源，请参见补充图 11b。c, 在 37℃ 下 10 mM Mg^{2+} 缓冲液中的 mRNA 化学稳定性。数据来自三次独立实验。d, 转染后 48 小时 HEK293 细胞中的蛋白质表达水平，由流式细胞仪测定。MFI 值取自三次独立实验，使用 Kruskal–Wallis ANOVA 和 Dunn’s 多重比较与 gE-Ther 组进行比较。e, C57BL/6 小鼠（$n = 5$）肌内注射两剂制备的 mRNA 疫苗（间隔一周），检测终点抗 gE IgG 效价。双尾 Mann–Whitney 检验用于统计分析。数据表示为平均值 ± 标准差（s.d.）。* $P < 0.05$，** $P < 0.01$。详见补充数据图 8 及补充表 3 获取详细的计算和实验数据。

5 VZV mRNA 疫苗的结果

为了进一步评估 LinearDesign 的泛化能力，我们将该算法应用于 VZV 疫苗的 mRNA 设计。接种 VZV 疫苗被认为是有效降低带状疱疹风险的策略[32]。使用与刺突 mRNA 设计相同的策略（如图 4a 所示），生成了五种编码全长 VZV gE 蛋白（gE-A 至 gE-E）的 mRNA 序列。这些序列广泛分布在以前未探索的高稳定性区域（图 5a）。这些序列与使用广泛使用的密码子优化工具 GeneOptimizer[33] 设计的基准 gE-Ther 序列进行了比较。基准 mRNA，包括野生型 gE mRNA（gE-WT），共享相同的氨基酸序列和 5' 与 3' UTR（序列见补充信息）。在非变性凝胶上，与刺突 mRNA 数据一致，gE-A mRNA（最低 MFE）表现出最高的迁移率，而 gE-E（最高 MFE）的降解速率明显较低（图 5b）。这表明联合优化 CAI 和 MFE 的重要性。最高表达的分子是那些 CAI 和 MFE 均在可行区域内的分子（图 5a 中的浅蓝色阴影区域）。我们进一步评估了 VZV mRNA 疫苗在 C57BL/6 小鼠中的免疫反应。LinearDesign 生成的 mRNA（gE-B、gE-C 和 gE-E）在抗 gE IgG 抗体效价上显著高于 gE-Ther 和 gE-WT（图 5e）。

6 讨论

有效的 mRNA 设计策略对于 mRNA 疫苗的发展至关重要，而这些疫苗在对抗 COVID-19 大流行中显示出了巨大潜力。然而，由于庞大的搜索空间，这一任务仍然极具挑战性。我们提出了一种将 mRNA 设计问题转化为计算语言学中的经典问题的简单解决方案。该方法实现了一种高效的算法，可以在 11 分钟内设计出编码 SARS-CoV-2 刺突蛋白的最优 mRNA，并能联合优化稳定性和密码子使用。这种方法基于最近语言学和生物信息学交叉研究的成果[35,36]。

在此研究中，我们全面表征了 LinearDesign 生成的 mRNA 序列，并展示了其在病毒抗原中的优越性，与传统的密码子优化基准相比，通过三项指标衡量其对疫苗性能的关键贡献：化学稳定性、蛋白翻译和体内免疫反应。特别地，我们的 SARS-CoV-2 刺突蛋白 mRNA 设计在体内显著增加了高达 128 倍的抗体效价和 9 至 20 倍的中和抗体滴度。对于 VZV gE mRNA 设计——由于其更紧凑的结构而在溶液中表现出更高的稳定性——我们观察到更高的蛋白质表达和免疫反应。我们的结果表明，LinearDesign 是 mRNA 疫苗开发的一种有效工具，并为未来的疫苗设计提供了新的思路。实际上编码区域设计和 UTR 工程[3] 是互补的，并且在未来的工作中可以结合起来。值得注意的是，我们设计的 mRNA 并未使用化学修饰，而化学修饰被广泛认为是近期 mRNA 疫苗成功的关键因素[1,2,10,37,38]。然而，我们的 mRNA 仍表现出高水平的稳定性、翻译效率和免疫原性，并且在制造成本上具有额外的优势。LinearDesign 方法很可能补充化学修饰的策略，一旦相应的能量模型可用，也可以轻松适配修饰核苷酸。我们的工作仅考虑了稳定性和密码子使用，但由于其格表示的泛化能力，也可以用于优化与 mRNA 设计相关的其他参数。通过开放以前难以访问的高稳定性和高效序列区域，这种方法为 mRNA 疫苗开发提供了一种及时且有前景的工具，可能在未来的大流行中发挥关键作用。这也是一种在 mRNA 药物设计领域中的基础方法，可用于包括单克隆抗体和抗癌药物在内的所有治疗性蛋白质的设计。