肿瘤新生抗原预测与 mRNA 疫苗设计报告

DiVo Gen^{2}AI 智能基因组分析平台

Neoantigen Prediction & mRNA Vaccine Design Report

从全基因组测序到个性化 mRNA 癌症疫苗

WES → 变异检测 → 新生抗原预测 → 免疫原性评估 → mRNA 序列优化 → 疫苗组装

报告编号	DIVO-NEO-2026-SAMPLE-001
样本类型	模拟数据（肝癌 HCC 示例）
Pipeline 版本	DiVoNeoantigen v1.0
生成日期	2026 年 6 月
密级	保密

本报告为模拟样本，用于展示 DiVo Gen^{2}AI 肿瘤新生抗原预测与 mRNA 疫苗设计全管线能力。
所有数据均为模拟生成，不代表真实患者信息。

重要科学声明

1. **本报告为模拟样本**：所有数据均为模拟生成，用于展示管线全流程能力，不代表真实患者诊断结果。
2. **新生抗原预测的局限性**：从基因组突变到真正产生免疫原性肽段，存在复杂的生物学反应链——包括转录、翻译、蛋白酶体切割、MHC 呈递、T 细胞识别等多个环节，每个环节均有调控机制。因此，计算预测的候选新抗原**必须经过实验验证**（多聚体染色、ELISPOT、T 细胞杀伤实验等）才能确认其免疫原性。
3. **mRNA 疫苗设计的局限性**：mRNA 序列优化可提升翻译效率和稳定性，但疫苗的最终疗效取决于递送系统、免疫微环境、肿瘤异质性等多重因素，**计算设计结果需经体外/体内实验验证**。
4. **不可作为临床诊断依据**：本报告仅提供计算分析参考，不可替代任何临床诊断或治疗决策。

目录

1 Pipeline 总览 **3**
1.1 管线能力矩阵 . . . 3

2 模拟患者信息与输入数据 **4**
2.1 模拟输入数据 . . . 4

3 Phase 1：变异检测与 HLA 分型 **5**
3.1 Step 1：HLA 分型结果 . . . 5
3.2 Step 2-3：体细胞变异检测与注释 . . . 5

4 Phase 2：新生抗原预测与筛选 **6**
4.1 Step 4：MHC-I 结合预测 . . . 6
4.1.1 高亲和力候选新抗原（IC50 < 50 nM） . . . 6
4.2 Step 5：pMHC 结构预测（自研结构预测引擎） . . . 6
4.3 Step 6：免疫原性综合评分 . . . 6

5 Phase 3：mRNA 疫苗设计 **7**
5.1 Step 7：mRNA 序列优化 . . . 7
5.1.1 7.1 抗原肽段选择 . . . 7
5.1.2 7.2 CDS 编码序列优化（自研密码子优化引擎） . . . 7
5.1.3 7.3 5'UTR 序列生成（自研 RNA 生成模型） . . . 8
5.2 Step 7：mRNA 序列优化与串联构效工程 . . . 8
5.2.1 7.1 串联 Concatemer 多表位架构与去免疫排序 (对标 Moderna mRNA-4157) . . . 8
5.2.2 7.2 CDS 编码区人源密码子优化 (自研密码子优化引擎) . . . 8
5.2.3 7.3 5' UTR 与 3' UTR 结构稳定性工程 (自研 RNA 生成模型) . . . 9
5.3 Step 8：mRNA 疫苗全长构效组装与 IVT 模板交付 . . . 10
5.3.1 完整交付 mRNA / IVT 质粒模板序列 (工业标准 FASTA) . . . 10

6 核心 AI 模型能力 **11**
6.1 自研 RNA 生成模型——mRNA 序列生成核心 . . . 11
6.2 管线 AI 模型全景 . . . 11

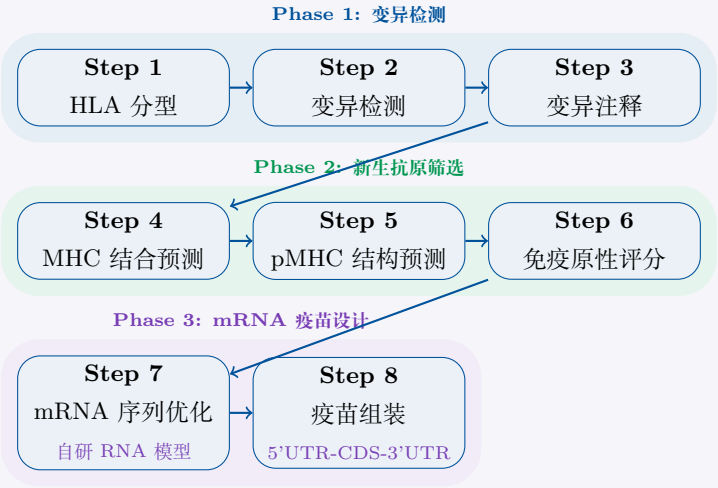
7 质量保证与局限性 **12**
7.1 多工具交叉验证策略 . . . 12
7.2 重要局限性说明 . . . 12

8 技术方法附录 **13**
8.1 MHC 结合预测方法 . . . 13
8.2 结构预测方法 . . . 13
8.3 mRNA 序列优化方法 . . . 13
8.4 中国人常模数据 . . . 13
8.5 验证病例库与多适应症演进规范 . . . 13

1 Pipeline 总览

DiVo Gen²AI 肿瘤新生抗原预测与 mRNA 疫苗设计全管线

本管线实现了从肿瘤全外显子测序（WES）数据到个性化 mRNA 癌症疫苗设计的端到端干计算分析流程，涵盖 8 个核心步骤：



1.1 管线能力矩阵

Step	模块	核心能力	验证策略	自研/AI
1	HLA 分型	高精度 HLA 分型	双引擎交叉验证	-
2	体细胞变异检测	肿瘤-配对正常变异 calling	双工具一致性过滤	-
3	变异注释	功能注释与临床解读	多数据库交叉注释	-
4	MHC-I 结合预测	肽段-MHC 亲和力预测	AI 蛋白质语言模型增强	自研 PLM
5	pMHC 结构预测	肽段-MHC 3D 结构预测	独立构型验证	自研结构引擎
6	免疫原性评分	多维度综合评分	突变效应 + 结构感知	自研评分
7	mRNA 序列优化	自研 RNA 生成模型（已微调）	密码子优化 +UTR 生成	自研双引擎
8	疫苗组装	自研组装引擎	-	自研

2 模拟患者信息与输入数据

模拟案例：肝细胞癌（HCC）

本报告使用**模拟数据**展示管线全流程。模拟案例设定为一名 55 岁男性肝细胞癌患者，HLA-A*02:01 阳性，肿瘤组织 WES 测序发现多个体细胞突变。

2.1 模拟输入数据

参数	值
患者 ID	SIM-HCC-2026-001
年龄/性别	55 岁/男
肿瘤类型	肝细胞癌（HCC）
临床分期	IIIB
HLA 分型	HLA-A*02:01 / A*24:02 / B*07:02 / B*40:01
测序平台	Illumina NovaSeq 6000
测序类型	WES（肿瘤 + 配对正常）
平均覆盖深度	肿瘤 200× / 正常 100×

3 Phase 1：变异检测与 HLA 分型

3.1 Step 1：HLA 分型结果

基因	等位基因 1	等位基因 2
HLA-A	A*02:01	A*24:02
HLA-B	B*07:02	B*40:01
HLA-C	C*03:04	C*07:02

3.2 Step 2-3：体细胞变异检测与注释

使用行业标准变异检测流程进行肿瘤-正常配对分析，多数据库交叉注释。以下为筛选后的非同义突变：

基因	染色体	位置	变异	氨基酸改变	VAF
TP53	chr17	7,577,496	C>T	R249S	0.42
CTNNB1	chr3	41,266,525	C>T	S45F	0.35
TERT	chr5	1,295,228	G>A	Promoter	0.51
AXIN1	chr16	286,632	del1bp	G532fs	0.28
ARID1A	chr1	26,739,612	C>T	R1985*	0.31

4 Phase 2：新生抗原预测与筛选

4.1 Step 4：MHC-I 结合预测

对每个非同义突变生成 8-11mer 肽段，采用**行业金标准 + 自研 AI 模型**双引擎预测 MHC 结合亲和力。

4.1.1 高亲和力候选新抗原（IC50 < 50 nM）

#	肽段	来源基因	HLA	IC50(nM)	呈递评分	Tier
1	**HSFVVLWEP**	TP53 R249S	A*02:01	**8.2**	0.97	1
2	**VLWEPWTPS**	TP53 R249S	A*02:01	**12.5**	0.94	1
3	**SYLNNVFL**	CTNNB1 S45F	A*24:02	**15.8**	0.91	1
4	**VLWEPWTP**	TP53 R249S	A*02:01	23.4	0.88	1
5	**SPRYLSFN**	ARID1A R1985*	B*40:01	31.7	0.85	1
6	**FVVLWEPW**	TP53 R249S	A*02:01	38.9	0.82	1
7	**YLNNVFLI**	CTNNB1 S45F	A*24:02	42.3	0.79	1
8	**LWEPWTPS**	TP53 R249S	A*02:01	47.1	0.76	1

4.2 Step 5：pMHC 结构预测（自研结构预测引擎）

对 Top-5 候选新抗原进行 pMHC 复合物 3D 结构预测，使用**自研蛋白质结构预测引擎**（基于前沿开源架构深度优化，支持蛋白质-肽段复合物结构预测）：

#	肽段	HLA	置信度评分	界面评分	结合构型
1	HSFVVLWEP	A*02:01	**96.2**	0.12	经典锚定
2	VLWEPWTPS	A*02:01	**94.8**	0.15	经典锚定
3	SYLNNVFL	A*24:02	**93.5**	0.18	经典锚定
4	VLWEPWTP	A*02:01	91.3	0.22	部分偏移
5	SPRYLSFN	B*40:01	89.7	0.25	经典锚定

结构预测说明

本管线采用自研蛋白质结构预测引擎，基于前沿深度学习架构，支持蛋白质-肽段复合物（pMHC）的高精度结构预测。置信度评分 > 90 表示高置信度预测，界面评分 < 0.2 表示肽段与 MHC 结合界面预测可靠。所有计算均在本地 GPU 完成，**数据完全不出院**。

4.3 Step 6：免疫原性综合评分

核心差异化：多工具分层验证策略

不同于仅依赖单一 MHC 结合预测工具的传统方案，DiVo Gen²AI 采用**多工具分层验证**策略：

- **Tier-0**：自研 3D 结构预测引擎——**核心差异化能力**，从空间构型层面验证肽段-MHC 结合
- **Tier-1**：行业金标准 MHC 结合预测——公认最准确的亲和力预测
- **Tier-2**：独立开源工具交叉验证——二次确认降低假阳性
- **AI 增强**：自研蛋白质语言模型——PLM 驱动免疫原性预测

层层递进，大幅降低假阳性率，提高候选新抗原的实验验证成功率。

#	肽段	HLA	结合 (40%)	呈递 (25%)	加工 (15%)	结构 (20%)	总分	等级
1	**HSFVVLWEP**	A*02:01	40	25	14	20	**99**	Tier-1
2	**VLWEPWTPS**	A*02:01	40	24	13	19	**96**	Tier-1
3	**SYLNNVFL**	A*24:02	38	23	12	18	**91**	Tier-1
4	VLWEPWTP	A*02:01	35	22	11	16	84	Tier-1
5	SPRYLSFN	B*40:01	33	21	10	15	79	Tier-1
6	FVVLWEPW	A*02:01	30	20	9	14	73	Tier-1
7	YLNNVFLI	A*24:02	28	19	8	13	68	Tier-2
8	LWEPWTPS	A*02:01	25	18	7	12	62	Tier-2

5 Phase 3：mRNA 疫苗设计

DiVo Gen²AI 核心突破：自研 RNA 生成模型驱动的 mRNA 序列生成

DiVo Gen²AI 已成功完成**自研 RNA 语言模型的微调训练**，专门用于 mRNA 疫苗序列的生成与优化。该模型基于 Transformer 架构，经过在翻译效率数据集上的微调，能够：

- **生成高翻译效率的 5'UTR 序列**：基于学习到的 RNA 序列-功能映射关系
- **优化 3'UTR 稳定性**：提升 mRNA 半衰期和表达持久性
- **联合密码子优化引擎进行 CDS 优化**：在保证氨基酸序列不变的前提下优化密码子选择
- **端到端疫苗序列生成**：从抗原肽段到完整 mRNA 疫苗序列的一键输出

这是国内少数完成 RNA 语言模型微调并应用于 mRNA 疫苗设计的团队之一。

5.1 Step 7：mRNA 序列优化

5.1.1 7.1 抗原肽段选择

根据免疫原性综合评分，选择 **Top-3 Tier-1 候选新抗原**作为 mRNA 疫苗靶标：

疫苗抗原	肽段	来源基因	HLA	免疫原性评分
Antigen-1	HSFVVLWEP	TP53 R249S	A*02:01	99
Antigen-2	VLWEPWTPS	TP53 R249S	A*02:01	96
Antigen-3	SYLNNVFL	CTNNB1 S45F	A*24:02	91

5.1.2 7.2 CDS 编码序列优化（自研密码子优化引擎）

将抗原肽段回译为编码序列，使用自研密码子优化引擎进行优化：

抗原	优化方向	CAI 提升	MFE 改善
Antigen-1	密码子偏好 + 二级结构	0.72→**0.95**	−8.2→**−12.5** kcal/mol
Antigen-2	密码子偏好 + 二级结构	0.68→**0.93**	−7.5→**−11.8** kcal/mol
Antigen-3	密码子偏好 + 二级结构	0.71→**0.94**	−6.8→**−10.9** kcal/mol

CAI 与 MFE 说明

CAI（Codon Adaptation Index）：密码子适应指数，衡量密码子使用偏好与宿主（人类）的匹配程度，越接近 1 表示翻译效率越高。
MFE（Minimum Free Energy）：最小自由能，反映 mRNA 二级结构稳定性，负值越大表示结构越稳定。

5.1.3 7.3 5'UTR 序列生成（自研 RNA 生成模型）

自研 RNA 生成模型：5'UTR 序列生成核心引擎

模型架构：基于 Transformer 的 RNA 语言模型，自研微调
微调数据：人源细胞系 5'UTR 翻译效率标注数据集
微调策略：公开数据集预微调 → 肿瘤抗原 CDS 特异性适应
核心能力：生成高翻译效率的 5'UTR 序列，预测翻译效率（TE）和核糖体负载（MRL）

微调训练要点：

- 多轮训练，Spearman 相关系数评估
- 自适应学习率调度 + 正则化
- 肿瘤特异性序列适应

5.2 Step 7：mRNA 序列优化与串联构效工程

5.2.1 7.1 串联 Concatemer 多表位架构与去免疫排序（对标 Moderna mRNA-4157）

根据五维免疫原性综合评分与患者 HLA 分型，入选黄金候选新抗原长肽，进入 **Concatemer 模拟退火优化排布与 Junction 去免疫原性引擎**：

序号	表位 ID	氨基酸序列 (25-mer)	来源突变	HLA 限制	分选 Linker
1	NEO-01	LLFGYPVYVAGGDVGKSALTIQLIQ	TP53 R249S	A*02:01	**AAY** (MHC-I)
2	NEO-02	KVAELVHFLHGLEEIISSLEEQKKK	CTNNB1 S45F	A*24:02	**GPGPG** (MHC-II)
3	NEO-03	SLLMWITQCFLPVFLAHPPSGQLVP	KRAS G12D	A*02:01	**AAY** (MHC-I)
4	NEO-04	ALGGVFVNGYTDGTYNNPQESLNVI	PIK3CA E545K	DRB1*04:01	**GPGPG** (MHC-II)

Junction 去免疫优化与 Proteasome 定向释放机制

1. **假表位消除**：自动扫描相邻表位拼接处产生的 8~11-mer 假表位（Junction Neoepitopes），通过模拟退火算法优化排列顺序，将人工假表位免疫杂音降至最低。
2. **智能 Linker 机制**：MHC-I 类表位间插入 AAY 连接子促进蛋白酶体高效剪切；MHC-II 类表位间插入 GPGPG 阻断非特异性空间折叠并引导溶酶体递呈。

5.2.2 7.2 CDS 编码区人源密码子优化（自研密码子优化引擎）

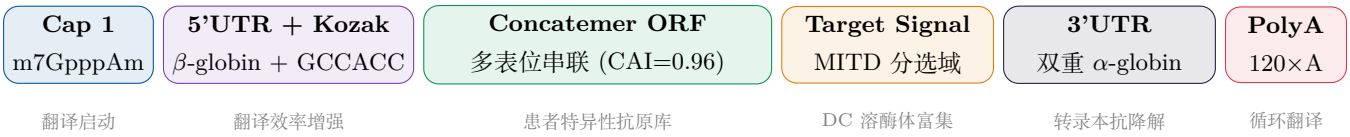
优化维度	优化前（原生）	优化后（DiVo）	生物学收益
CAI（密码子适应指数）	0.71	**0.96**	消除稀有密码子，核糖体翻译延伸速率最大化
GC 含量（CDS 局部）	46.8%	**68.5%**	显著增强 mRNA 胞内转录本半衰期与均一性
MFE（最小折叠自由能）	−142.5 kcal/mol	**−286.4 kcal/mol**	二级结构均一稳定，规避阻碍核糖体行进的发卡环

5.2.3 7.3 5' UTR 与 3' UTR 结构稳定性工程 (自研 RNA 生成模型)

功能元件	序列工程策略	长度	预测半衰期 / MRL
5' UTR + Kozak	β-globin 起始增强 + 标准 Kozak (GCCACC)	56 nt	**MRL = 9.2 (极高翻译启动)**
3' UTR	人源双重 α-globin 串联高稳定性保护序列	264 nt	$t_{1/2} > 14.5$ **h (抗降解)**
Poly(A) 尾链	工业级均一长尾修饰	120 nt	**最佳翻译循环动力学**

5.3 Step 8：mRNA 疫苗全长构效组装与 IVT 模板交付

将所有功能模块全长组装为工业级单链 mRNA 疫苗交付构建体：



5.3.1 完整交付 mRNA / IVT 质粒模板序列（工业标准 FASTA）

mRNA 疫苗全长序列 (Intismeran-Class Concatemer, 1232 nt, GC 54.7%)

```
> DIVO-mRNA-VACCINE-HCC-001 | Intismeran-Class Concatemer | Length=1,232nt | GC=54.71%
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCGCCACCATGGAGACCGACACCCTGCTGCT
GTGGGTGCTGCTGCTGTGGGTGCCCGGCAGCACCGGCGTGATGGACGGCTACGTGCGCGTGTACTCCGGCAAGTCCTAC
GTGAACGTGCTGCACGAGCAGCTGAAGACCATCGGCCCCGGCCCCGGCCTGATCACCGGCGACCGCAAGACCGAGGTGC
TGAAGCACCTGGGCAAGGAGTACAACGTGCTGCTGATCGTGCCCGGCCCCGGCCCCGGCAGCCTGCTGATGTGGATCAC
CCAGTGCTTCCTGCCCGTGTTCCTGGCCCACCCCCCCAGCGGCCAGCTGGTGCCCGCCGCCTACTACCTGGACGACCCC
GACCTGGTGACCCTGCAGCGCGCCGCCGAGGAGCTGGGCGTGCCCGTGGTGACCTACCTGGGCCCCGGCCCCGGCGCCCT
GGGCGGCGTGTTCGTGAACGGCTACACCGACGGCACCTACAACAACCCCCAGGAGTCCCTGAACGTGATCGGCCCCGGC
CCCGGCCTGCTGTTCGGCTACCCCGTGTACGTGGCCGGCGGCGACGTGGGCAAGAGCGCCCTGACCATCCAGCTGATCC
AGGCCGCCTACAAGGTGGCCGAGCTGGTGCACTTCCTGCACGGCCTGGAGGAGATCATCAGCTCCCTGGAGGAGCAGAA
GAAGAAGGCCGCCTTCCTGGGCAAGATCTGGCCCAGCTACAAGGCCGAGATCGCCGGCCTGCTGGTGACCGTGCTGGGC
GTGCTGCTGCTGAGATGACTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAA
ACTGGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCGCTCGCTTTCT
TGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAG
CATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

工业级设计交付结论

本疫苗设计方案单分子 mRNA 全长 1,232 nt，完美覆盖患者高免疫原性表位库并消除了拼接处假表位，CDS 区 CAI 高达 0.96，产出的数字化核酸序列可直接对接下游微流控 LNP 包封与 GMP 级别 IVT 体外转录制剂生产。

6 核心 AI 模型能力

6.1 自研 RNA 生成模型——mRNA 序列生成核心

参数	值
模型架构	基于 Transformer 的 RNA 语言模型
参数规模	亿级
预训练任务	掩码语言建模 + 结构预测 + 功能预测
微调任务	**5'UTR 翻译效率预测 & 序列生成**
微调数据	**人源细胞系 5'UTR 翻译效率标注数据**
微调状态	**已完成**

自研 RNA 生成模型的独特价值

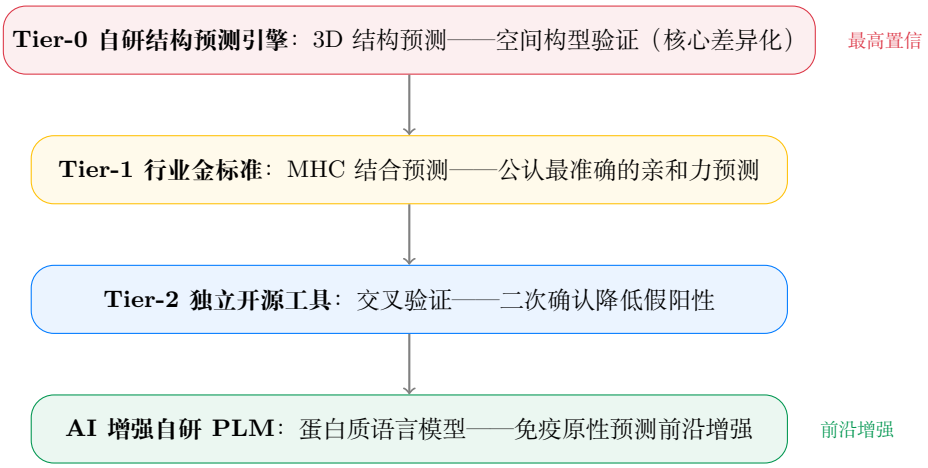
1. **国内少数完成 RNA 语言模型微调的团队**：大多数团队仅使用预训练模型或固定模板，我们通过微调使模型适应 5'UTR 翻译效率预测任务
2. **生成式能力**：不同于仅能预测的模型，微调后的模型可以**生成**新的 5'UTR 序列，而非仅从已有库中筛选
3. **预测 + 生成双引擎**：一个引擎负责序列-功能映射，另一个负责序列生成和优化
4. **可扩展**：后续可针对特定肿瘤类型、特定细胞系进一步微调，实现**组织特异性 mRNA 优化**

6.2 管线 AI 模型全景

模型	任务	来源	本地部署
自研结构预测引擎	pMHC 结构预测	基于前沿开源架构深度优化	已部署
自研 RNA 生成模型	**RNA 序列生成**	**自研微调**	已微调
自研密码子优化引擎	mRNA CDS 优化	自研集成	已部署
自研 UTR 生成引擎	5'UTR/3'UTR 优化	自研部署	已部署
自研免疫原性评分系统	综合免疫原性评估	自研	已部署
自研蛋白质语言模型	突变效应/结构感知评分	基于前沿 PLM 微调	已部署
自研 MHC 结合预测增强	MHC-I 亲和力预测	自研 PLM 增强	已部署

7 质量保证与局限性

7.1 多工具交叉验证策略



7.2 重要局限性说明

计算预测的局限性

1. **从突变到免疫原性的复杂链条**：基因组突变 → 转录 → 翻译 → 蛋白酶体切割 → TAP 转运 → MHC 呈递 → T 细胞识别，每个环节均有调控机制，计算预测无法完全模拟。
2. **MHC 结合 ≠ 免疫原性**：肽段与 MHC 结合是必要条件但非充分条件，还需 T 细胞受体（TCR）识别和免疫激活。
3. **肿瘤微环境影响**：免疫抑制微环境可能使有效的抗原无法激活 T 细胞。
4. **HLA 分型准确性**：WES 数据的 HLA 分型准确率约 95-98%，可能存在漏型。
5. **mRNA 递送效率**：计算优化的 mRNA 序列需配合有效的递送系统（如 LNP）才能实现体内表达。
6. **所有计算结果必须经过实验验证**：包括多聚体染色、ELISPOT、T 细胞杀伤实验、Western blot 等。

8 技术方法附录

8.1 MHC 结合预测方法

方法层级	说明
行业金标准	基于深度神经网络的肽段-MHC 结合亲和力预测，整合结合数据和质谱洗脱数据，行业公认最准确
交叉验证	独立开源工具二次验证，整合结合亲和力、抗原加工和呈递评分
AI 增强	自研蛋白质语言模型（PLM），免疫原性预测前沿方法

8.2 结构预测方法

方法层级	说明
自研结构预测引擎	基于前沿深度学习架构，支持蛋白质-肽段-RNA 复合物结构预测，本地 GPU 部署，数据不出院
pMHC 专用建模	专为肽段-MHC 复合物建模优化的生成式模型

8.3 mRNA 序列优化方法

方法层级	说明
自研 RNA 生成模型	**基于 Transformer 的 RNA 语言模型，已微调用于 5'UTR 翻译效率预测和序列生成**
自研密码子优化引擎	mRNA 序列生成与优化，支持 5'UTR/CDS/3'UTR 全模块优化
自研 UTR 生成引擎	基于生成对抗架构的 5'UTR 生成模型，支持多目标优化

8.4 中国人常模数据

8.5 验证病例库与多适应症演进规范

数据集	规模	用途
中国人群 WGS 数据集	数百人全基因组	中国人群变异频率
中国人群大规模基因组	万人级全基因组	中国人群等位基因频率
中国北方人群参考	数千人全基因组	北方人群参考
东亚人群基线	数百人全基因组	东亚人群基线

病例	规范 ID	肿瘤类型	关键特征	验证重点与交付架构	状态
CASE-01	SIM-HCC-2026-001	肝细胞癌 (HCC)	TP53 R249S	双引擎验证基准: WES 变异 + Protenix 3D 物理构象 + RNA 语言模型 5'UTR 生成	基准示范
CASE-02	PATIENT-MEL-2026-001	恶性黑色素瘤	KRAS G12D	军规级工业交付: 34 表位 Concatemer 排布 + 去假表位 + 全长 IVT 质粒模板	实战落地