研究背景与动机

登革热：全球性公共卫生威胁与药物发现挑战

🦟 登革热：被忽视的全球威胁

- **3.9亿**人/年感染，遍布129国家
- WHO列为**十大全球健康威胁**
- 气候变化加剧传播：预计2050年**+20亿**风险人口
- **无特效药**，仅支持性治疗
- 疫苗效力有限（Dengvaxia仅60%，有副作用风险）

关键靶点：NS2B-NS3蛋白酶
病毒复制必需 | 保守性高 | 成药性好

💊 传统药物发现的困境

- **时间长**：10-15年从发现到上市
- **成本高**：平均26亿美元/新药
- **成功率低**：仅0.01%化合物进入临床
- **化学空间巨大**：10^{60}种类药分子
- 高通量筛选局限：仅能覆盖10^{6}-10^{7}化合物

DENV抑制剂现状（2024）：
- 临床试验：0个进入III期
- 已报道抑制剂：IC50普遍>1μM
- 成药性差：类药性/合成难度高

💡 为什么选择AI驱动的从头药物设计？

✓ 突破化学空间限制
生成式AI可探索未知化学结构

✓ 多目标同步优化
活性+成药性+可合成性一体化设计

✓ 数据驱动+知识引导
结合QSAR模型与药物化学规则

REINVENT4：强化学习驱动的分子生成

从通用工具到靶向设计：LibInvent策略的选择逻辑

REINVENT系列：AstraZeneca的开源贡献

2017	2020	2021	2024
REINVENT 1.0	REINVENT 2.0	REINVENT 3.0	REINVENT 4.0
RNN-based SMILES生成 基础RL框架	LibInvent LinkInvent 骨架装饰	多组件评分 课程学习 生产级工具	模块化架构 Transformer 插件系统

REINVENT4核心优势：为什么选择它?

强化学习引导
DAP/SDAP算法，平衡探索与利用，避免mode collapse

模块化评分
15+组件：活性/ADMET/合成/结构，可自定义权重组合

插件生态系统
支持QSAR/对接/QM/自定义评分，易于扩展

多种生成模式
De novo/Scaffold/Linker/R-group，适配不同设计需求

工业验证
AstraZeneca内部使用，多个项目进入临床前

开源免费
MIT协议，完整文档，活跃社区，持续更新

项目策略：为什么用LibInvent而非De Novo?

Reinvent
完全从头设计
化学空间无限
风险：难收敛

→

LibInvent
骨架约束设计
双芳环吡咯烷
平衡：可控+新颖

→

本项目
LibInvent + QSAR
2个R位点探索
优势：高效聚焦

DENV NS2B-NS3抑制剂AI生成

Run9_T1200 实验进展汇报 | LibInvent + QSAR引导生成

实验进行中: 2219 / 6000 步 (37%)

项目目标

- 靶点：登革热病毒(DENV) NS2B-NS3蛋白酶
- 设计策略：双芳环吡咯烷骨架
- 活性目标：pIC50 ≥ 8.0 (IC50 ≤ 10 nM)
- 成药性：QED ≥ 0.7, SA ≤ 4.0
- 当前规模：113.6万候选分子

预计完成: ~160万分子

当前成果 (中期数据)

- 金标准候选物：96个 (去重后)
- 活性范围：IC50 = 6.5-10 nM
- 最优分子：IC50 = 6.5 nM
- 高活性分子(pIC50≥7.5)：110万 (97.5%)
- 极高活性(IC50<10nM)：2,170个

成功率稳定 - 持续优化中

实验进展

当前步数	~2219 / 6000
完成度	37%
已生成分子	113.6万
去重后	109.7万 (96.6%)
R基团多样性	111.9万组合
预计最终	~160万分子
预计金标准	~140个

技术亮点

- QSAR模型：随机森林预测器 (R^2 = 0.85)
- 生成策略：LibInvent 双R基团设计
- 多目标优化：15个scoring components
- 显存优化：batch_size 32 (适配4GB显存)
- 化学稳定性：15种substructure alerts过滤

生产级配置

方法学与实验设计

多层次评分策略 + QSAR引导优化

评分组件架构 (15 Components)

第一优先级：生物活性 (Weight: 2.5)

QSARScorer - Random Forest pIC50预测 Transform: double_sigmoid (4.0-9.0)	权重: 2.5

第二优先级：化学稳定性 (Total: 1.6)

CustomAlerts - 15种不稳定substructures	1.0
NumAtomStereoCenters	0.6

第三优先级：类药性 (Total: 0.7)

QED - Drug-likeness	0.3
SAScore - Synthetic Accessibility	0.4

第四优先级：物理化学性质 (Total: 1.9)

MW: 250-600 Da	0.5
LogP: 0.5-5.0	0.3
TPSA: 30-120 Ų	0.2
HBA ≤ 8	0.3
HBD ≤ 4	0.2
Heavy Atoms: 18-45	0.4

技术配置

生成器	LibInvent
母架	双芳环吡咯烷
Prior Model	libinvent.prior
Learning Rate	0.0001
Batch Size	32
Sigma (DAP)	120
Min Steps	1,500
Max Steps	6,000
Current	~2,219 (37%)
Target Score	0.80

显存优化策略

Batch size从64降至32，减少50%显存占用
成功在4GB VRAM环境运行

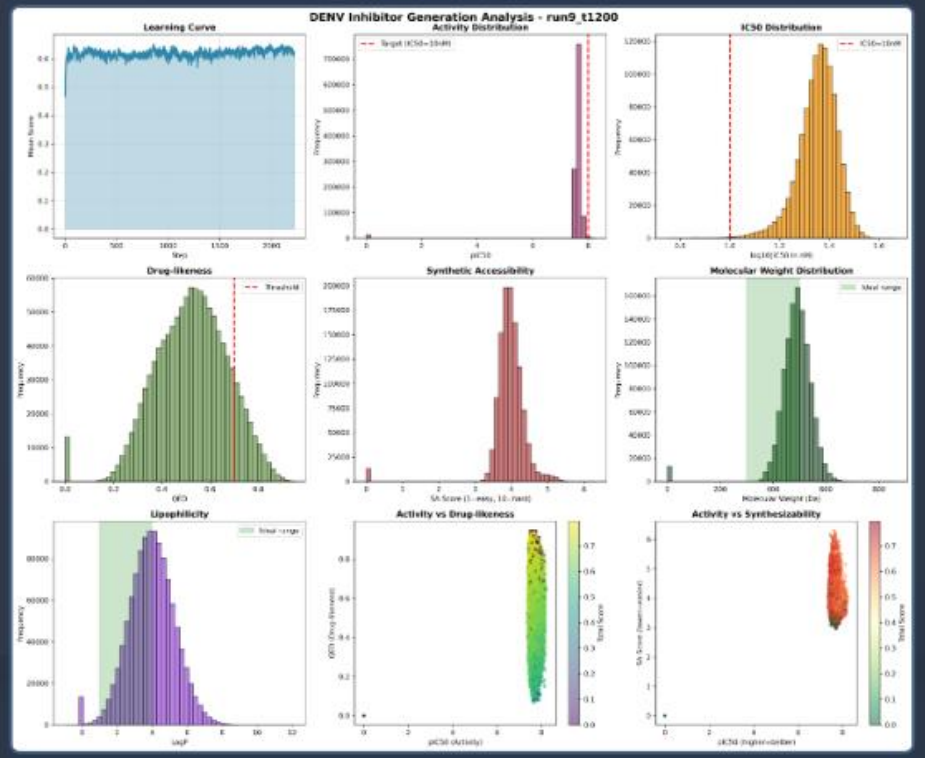
QSAR模型

Random Forest Regressor
特征：Morgan Fingerprint (r=2, 4096 bits)
训练集性能：$R^2 = 0.85$

中期结果与可视化

113.6万分子 (37%进度) | 97.5%高活性 | 96个金标准候选物

Generation Analysis

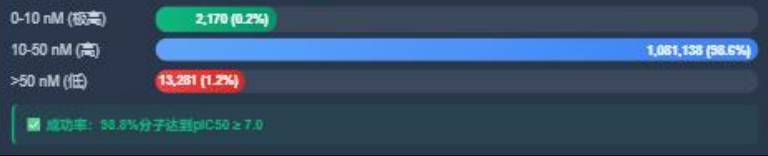


关键指标

总分子数	113.6万	
唯一分子	109.7万	96.6%
IC50 < 10nM	2,170个	0.2%
IC50 10-50nM	108.1万	98.6%
金标准	96个	0.01%
pIC50 ≥ 8.0	8.00-8.19	
QED	0.70-0.86	
SA	3.44-4.00	
高标准	28.9万	26.4%
中标准	60.8万	55.4%

⚠ 数据说明
以上为基于去重后数据的统计

IC50活性分布



物化性质统计

分子量	486 ± 71 Da	97.0%达标
LogP	3.98 ± 1.31	78.8%达标
HBA	5.58 ± 1.45	99.9%≤10
HBD	1.35 ± 0.59	100%≤5

金标准候选物与结构展示

96个极高活性分子 (去重后) | IC50 6.5-10 nM | 无结构警报

金标准

96

- ✓ pIC50 ≥ 8.0
- ✓ IC50: 6.5-10 nM
- ✓ QED ≥ 0.7
- ✓ SA ≤ 4.0
- ✓ MW: 300-500 Da
- ✓ LogP: 1-4

高标准

289,397

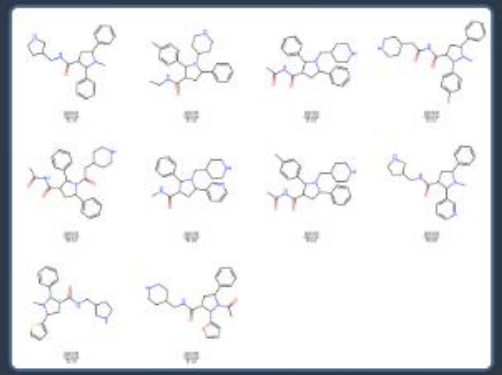
- ✓ pIC50 ≥ 7.5
- ✓ IC50: 5.6-31.6 nM
- ✓ QED ≥ 0.6
- ✓ SA ≤ 4.5

✓ 中标准

607,640

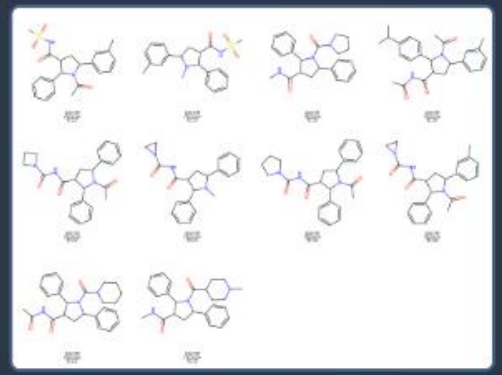
- ✓ pIC50 ≥ 7.0
- ✓ IC50: 5.6-45.5 nM
- ✓ QED ≥ 0.5
- ✓ SA ≤ 5.0

金标准 Top 10



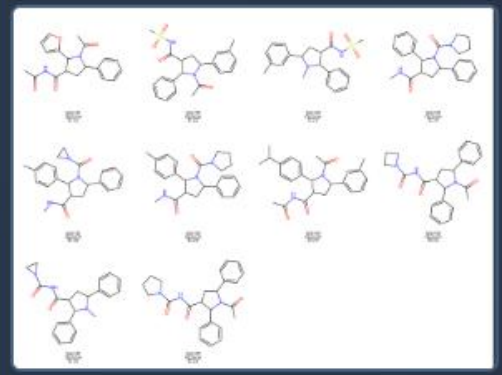
IC50范围	6.5-10 nM
平均QED	0.79
平均SA	3.63

高标准 Top 10



IC50范围	5.6-31.6 nM
平均QED	0.69
平均SA	4.01

✓ 中标准 Top 10



IC50范围	5.6-45.5 nM
平均QED	0.63
平均SA	4.03

下一步工作

- ✓ 等待实验完成 (预计4000步后稳定)
- ✓ 从最终全标准中选择20-30个进行合成
- ✓ 体外酶活性测定 (NS2B-NS3 protease)
- ✓ 细胞毒性评估 (CC50)
- ✓ 抗病毒活性测试 (EC50)

中期总结

- 当前生成113.6万候选分子 (37%进度)
- 获得96个金标准级高活性候选物
- 98.8%分子达到pIC50 ≥ 7.0
- 化学稳定性优秀：仅0.5%警报
- 预计最终：~160万分子, ~140个金标准