

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 $>1\mu\text{M}$
- 成药性差：类药性/合成难度高



为什么选择AI驱动的头药物设计？

✓ 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

REINVENT 1.0

RNN-based
SMILES生成
基础RL框架

2020

REINVENT 2.0

LibInvent
LinkInvent
骨架装饰

2021

REINVENT 3.0

多组件评分
课程学习
生产级工具

2024

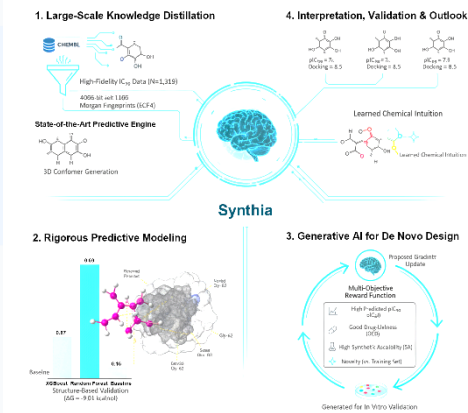
REINVENT 4.0

模块化架构
Transformer
插件系统

Synthia工作流程：知识蒸馏到生成式AI

整合策略：

- ✓ 大规模知识蒸馏：ChEMBL数据库提取高质量IC50数据(N=1,319)
- ✓ 严格预测建模：Random Forest ($R^2=0.85$) + 基于结构验证($\Delta G=-9.01$ kcal/mol)
- ✓ 生成式AI设计：LibInvent多目标奖励函数(活性+类药性+合成可及性+新颖性)
- ✓ 体外验证通路：生成候选物→体外酶活性测定→抗病毒活性评估



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
双芳环吡啶烷
平衡：可控+新颖

本项目(Synthia)

LibInvent + QSAR
2个R位点探索
优势：高效聚焦

决策依据：

- ✓ 文献已报道该骨架有中等活性 (pIC₅₀ ~6.0)，但成药性差
- ✓ 晶体结构显示骨架结合稳定，R基团伸向溶剂可及区
- ✓ LibInvent可同时优化2个R位点，探索10¹⁰+组合空间
- ✓ 避免从头设计的高风险 (60%分子可能违反Lipinski规则)
- ✓ QSAR模型基于该骨架训练 ($R^2=0.85$)，预测准确性高



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: $pIC_{50} \geq 8.0$ ($IC_{50} \leq 10$ nM)
- 成药性: $QED \geq 0.7$, $SA \leq 4.0$
- 当前规模: **113.6万** 候选分子

预计完成: ~160万分子

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

当前成果 (中期数据)

- 金标准候选物: **96个** (去重后)
- 活性范围: $IC_{50} = 6.5-10$ nM
- 最优分子: $IC_{50} = 6.5$ nM
- 高活性分子($pIC_{50} \geq 7.5$): **110万** (97.5%)
- 极高活性($IC_{50} < 10$ nM): **2,170个**

成功率稳定 - 持续优化中

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

活性权重: 34.5%

稳定性+成药性: 31.7%

物化性质: 26.2%

生物活性 (2.5)

QSARScorer - Random Forest pIC_{50} 预测

化学稳定性 (1.6)

CustomAlerts - 15种不稳定substructures

NumAtomStereoCenters

类药性 (0.7)

QED

SAScore

物理化学性质 (1.9)

MW: 250-600 Da

LogP: 0.5-5.0

TPSA: 30-120 Å²

HBA ≤ 8

HBD ≤ 4

Heavy Atoms: 18-45

结构复杂度 (0.55)

Rotatable Bonds ≤ 9

Ring Count: 2-5

Aromatic Rings: 1-3

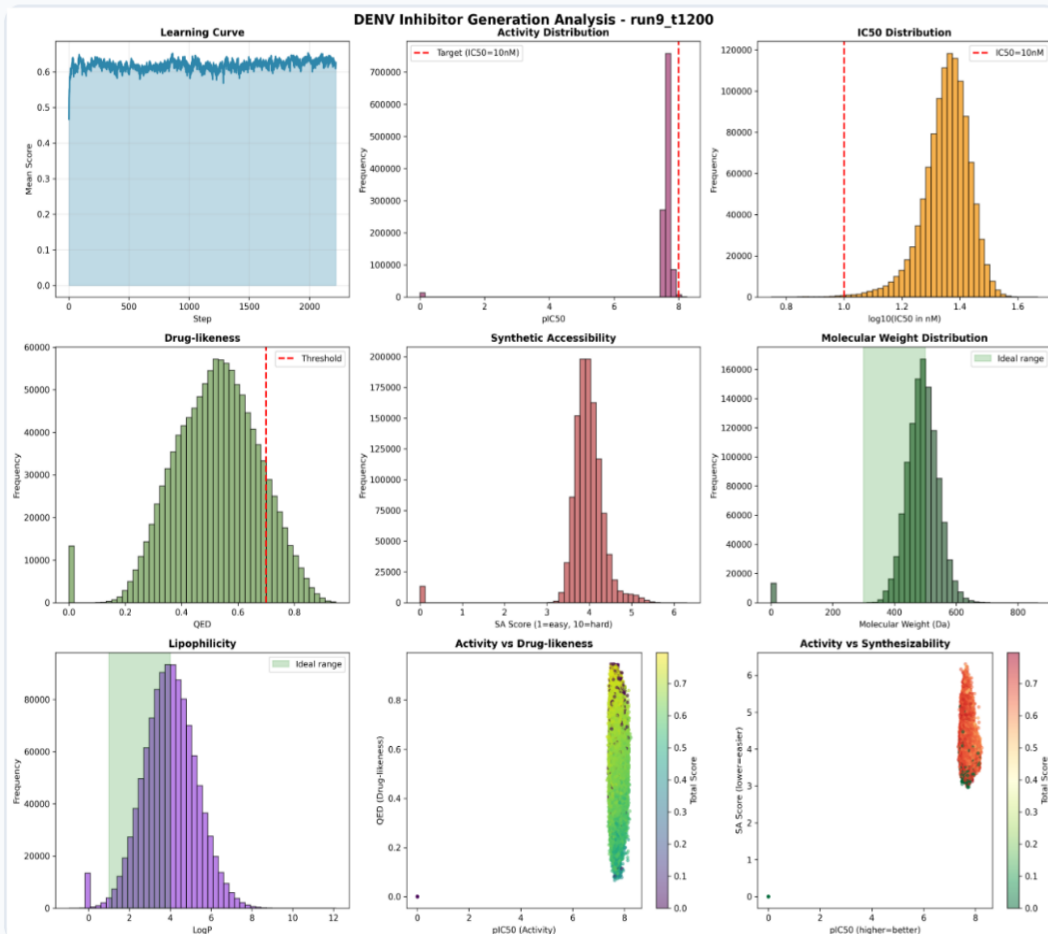


中期结果与可视化

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



技术配置

生成器	LibInvent
骨架	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

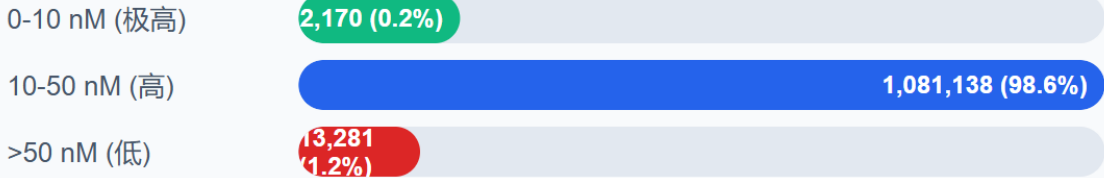
QSAR模型

Random Forest Regressor

特征: Morgan Fingerprint (r=2, 4096 bits)

训练集性能: $R^2 = 0.85$

🎯 IC50活性分布



✅ 成功率: 98.8%分子达到pIC50 ≥ 7.0

⚙️ 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
TPSA	88.2 ± 23.7 Å²	97.8%达标
QED	0.52 ± 0.15	
SA Score	3.95 ± 0.53	98.8%在1-6

📅 实验进展预测

当前 (37%)	
总分子	113.6万
金标准	96个
预计 (100%)	
总分子	~160万
金标准	~140个

⚠️ 化学稳定性

触发警报	5,740	0.5%
[*;r8]	2,266次	
[NX3][NX3]	1,718次	
[SH]	975次	

✅ 99.5%分子通过稳定性检查



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

★ 高标准

289,397

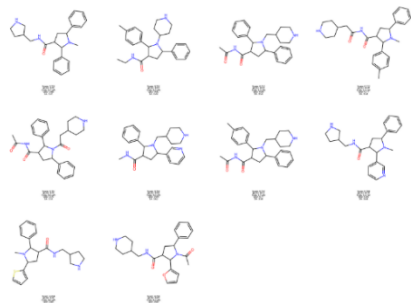
- ✓ pIC50 ≥ 7.5
- ✓ IC50: 5.6-31.6 nM
- ✓ QED ≥ 0.6 , SA ≤ 4.5

✓ 中标标准

607,640

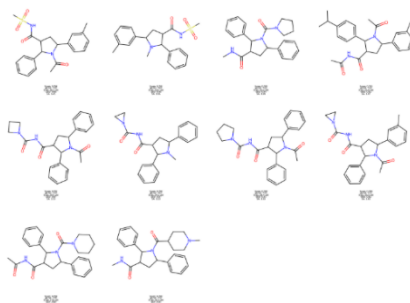
- ✓ pIC50 ≥ 7.0
- ✓ IC50: 5.6-45.5 nM
- ✓ QED ≥ 0.5 , SA ≤ 5.0

◆ 金标准 Top 10



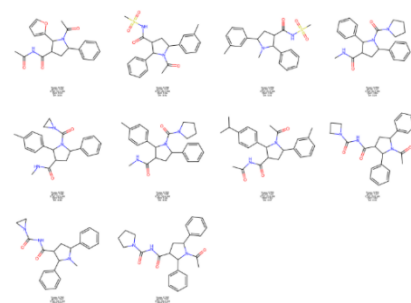
IC50范围	6.5-10 nM
平均QED	0.79
平均SA	3.63

★ 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)
- ✓ 从最终金标准中选择20-30个进行合成
- ✓ 体外酶活性测定 (NS2B-NS3 protease)

中期总结

- 当前生成113.6万候选分子 (37%进度)
- 获得96个金标准极高活性候选物
- 98.8%分子达到pIC50 ≥ 7.0

感谢广医药学院 — —各位领导给予的 帮助与支持

