

从过拟合到高通量：构建智能药物虚拟筛选系统

一项旨在解决药物发现中小样本学习挑战的深度学习实践

药物研发的困境：“双十定律”与高通量筛选的瓶颈

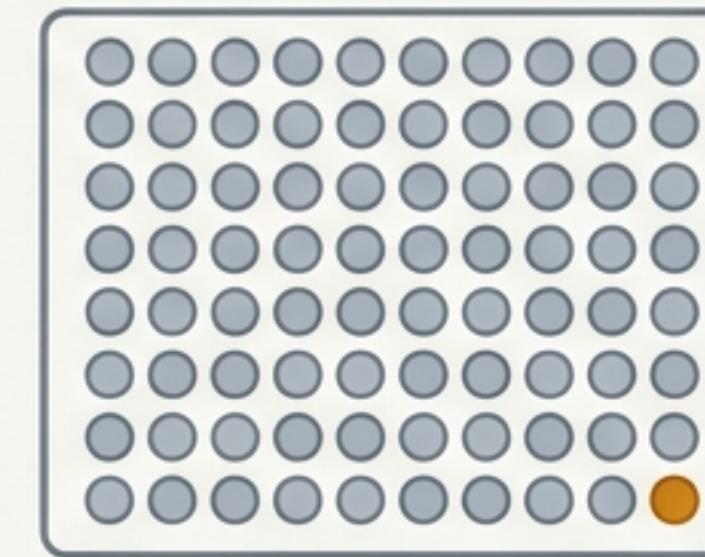
“双十定律”



传统药物研发平均耗时10年，耗资超过10亿美元，成功率不足10%。

传统 vs. 现代方法

传统高通量筛选 (HTS)



需对数百万化合物进行实验测试，成本高昂，命中率通常低于0.1%。

虚拟筛选 (VS)



利用计算机模拟和机器学习算法进行预筛选，大幅缩小实验范围，提高研发效率。

深度学习：加速药物发现的核心引擎

为什么选择深度学习？

与传统机器学习相比，深度神经网络能够自动从原始数据中学习层次化的特征表示，无需人工设计复杂的分子描述符。

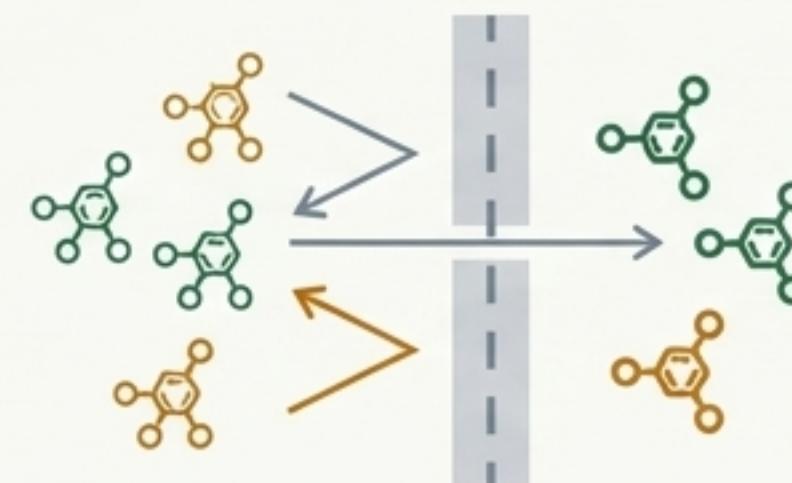
项目聚焦的核心预测任务

BBBP (分类任务)

血脑屏障穿透性预测 (Blood-Brain Barrier Penetration)

任务类型：二分类 (Classification)

意义：对中枢神经系统药物研发至关重要。

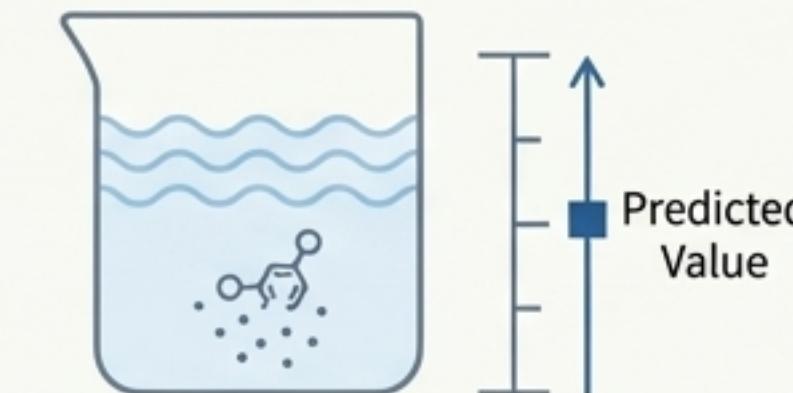


ESOL (回归任务)

水溶解度预测 (Estimated SOLubility)

任务类型：回归 (Regression)

意义：药物成药性评估的关键指标，影响人体吸收效率。



蓝图：一个端到端的模块化系统架构

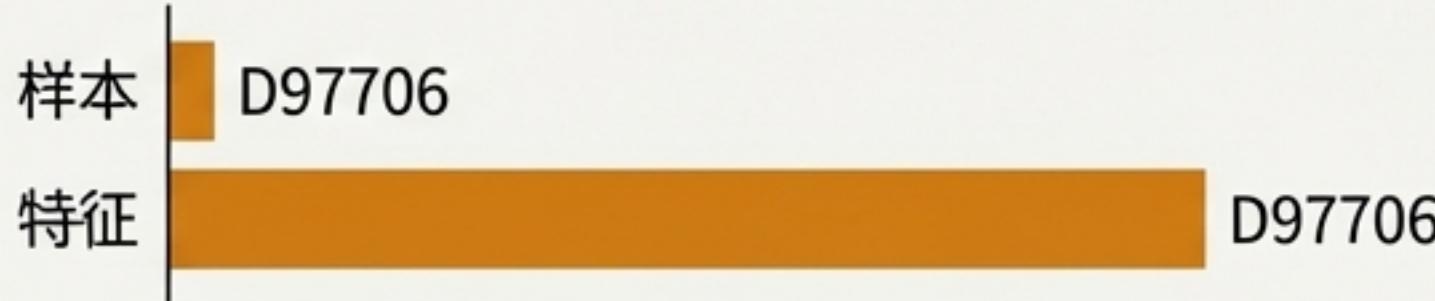


潜藏的巨龙：小样本数据下的严重过拟合

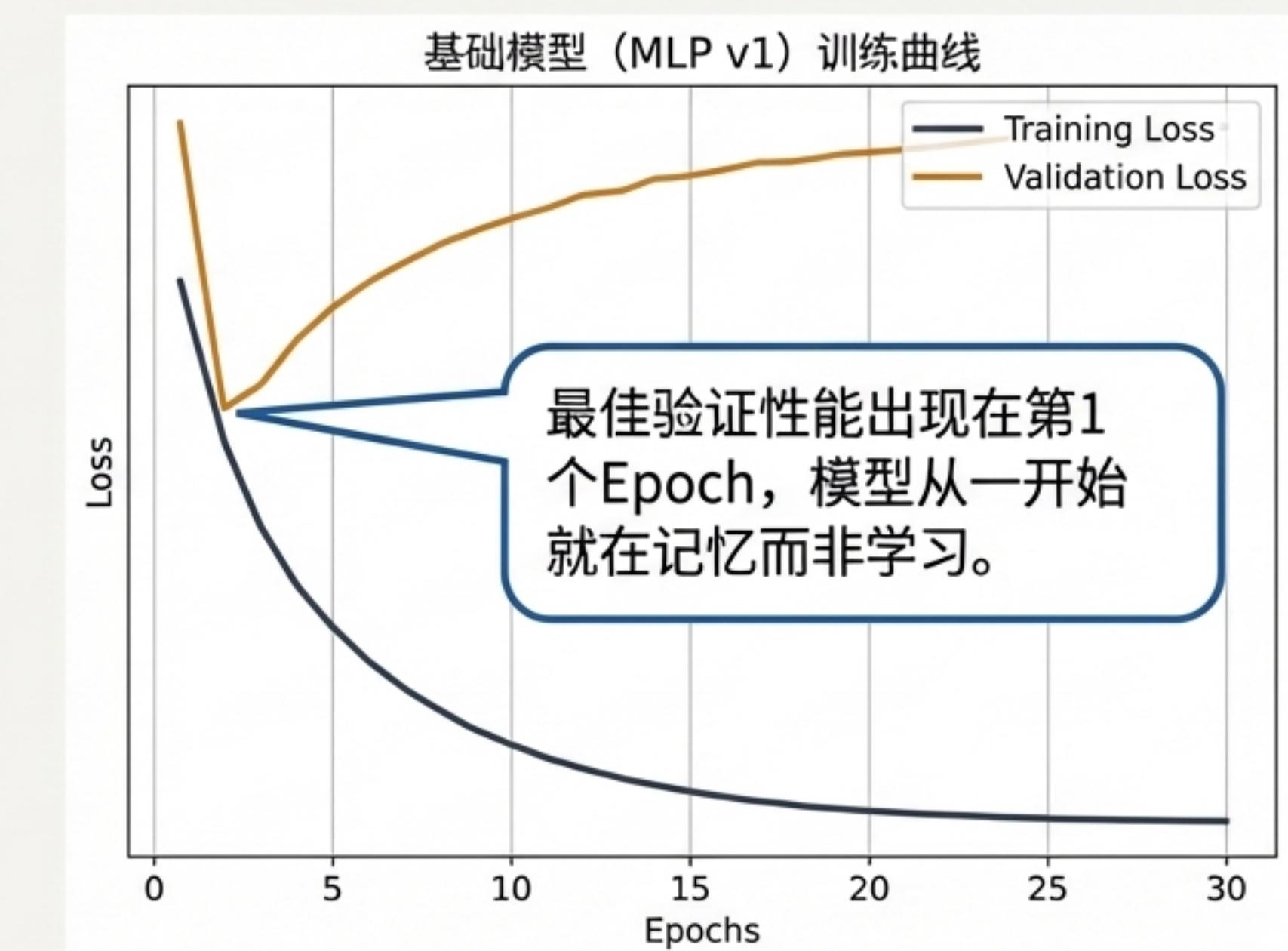
问题的根源

药物数据集的典型困境：样本量小 vs. 特征维度高。

- BBBP数据集：约2000个样本
- ESOL数据集：约1100个样本
- 分子指纹特征：1024维，稀疏二进制向量

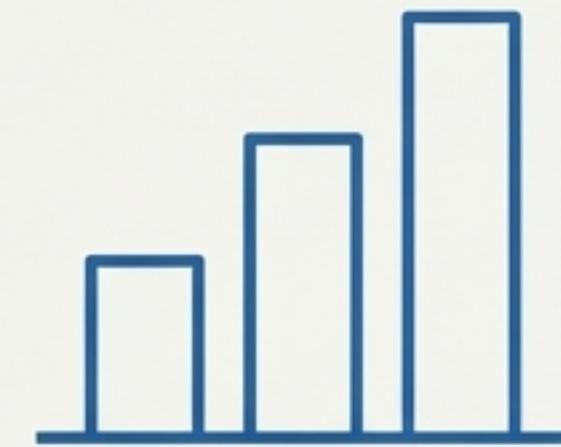
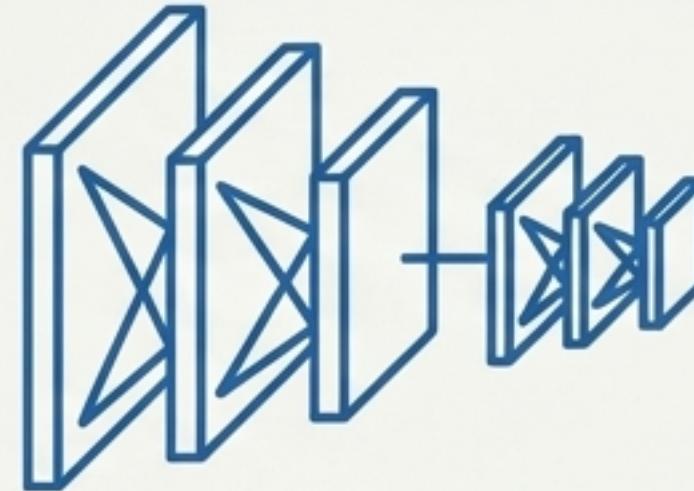


无可辩驳的证据



我们的对策：针对过拟合的三叉戟式攻击

引入增强模型 DrugPredictorMLPv2，其设计核心是多重正则化策略。



策略一：轻量化网络

将隐藏层从 [512, 256, 128] 压缩至 [256, 128, 64]，参数量减少 56%。更小的模型容量限制了其死记硬背的能力。

策略二：输入层正则化

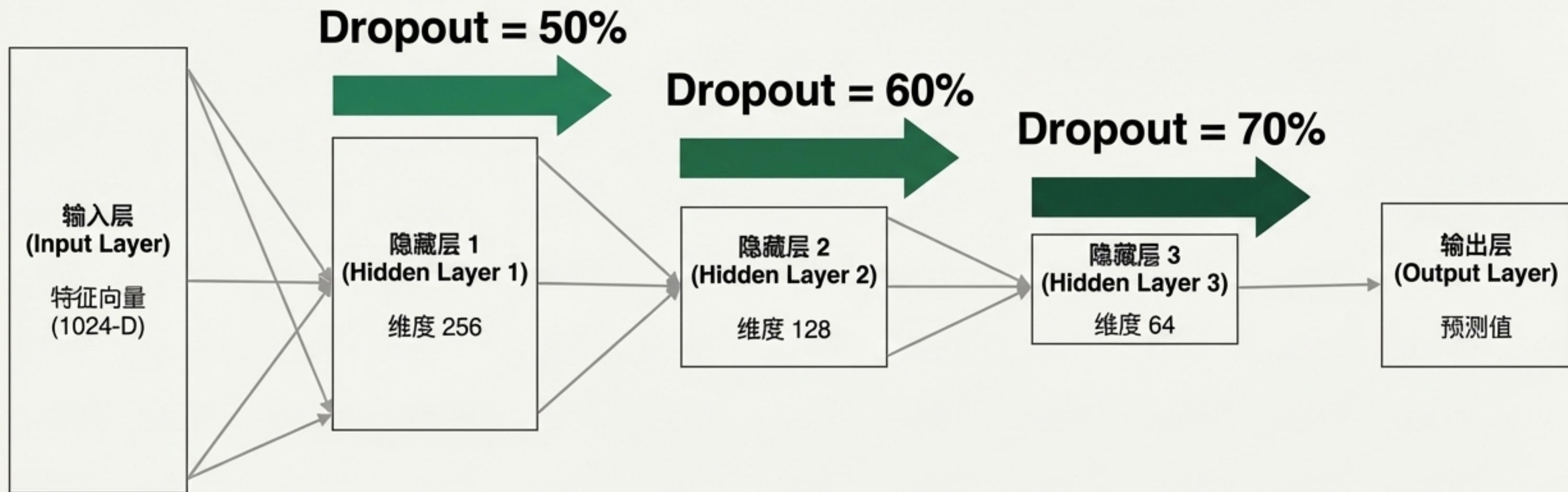
在1024维输入特征上应用25%的 Dropout，随机屏蔽部分分子指纹位，迫使模型学习更鲁棒的特征。

策略三：渐进式Dropout

创新性地设计随网络深度递增的 Dropout策略，对更深、更抽象的特征层施加更强的正则化。

核心创新：渐进式Dropout机制详解

“网络越深，学习到的特征越抽象、越复杂，因此也越容易过拟合到训练数据中的特定模式。我们的对策是：更深的网络层需要更强的正则化。”

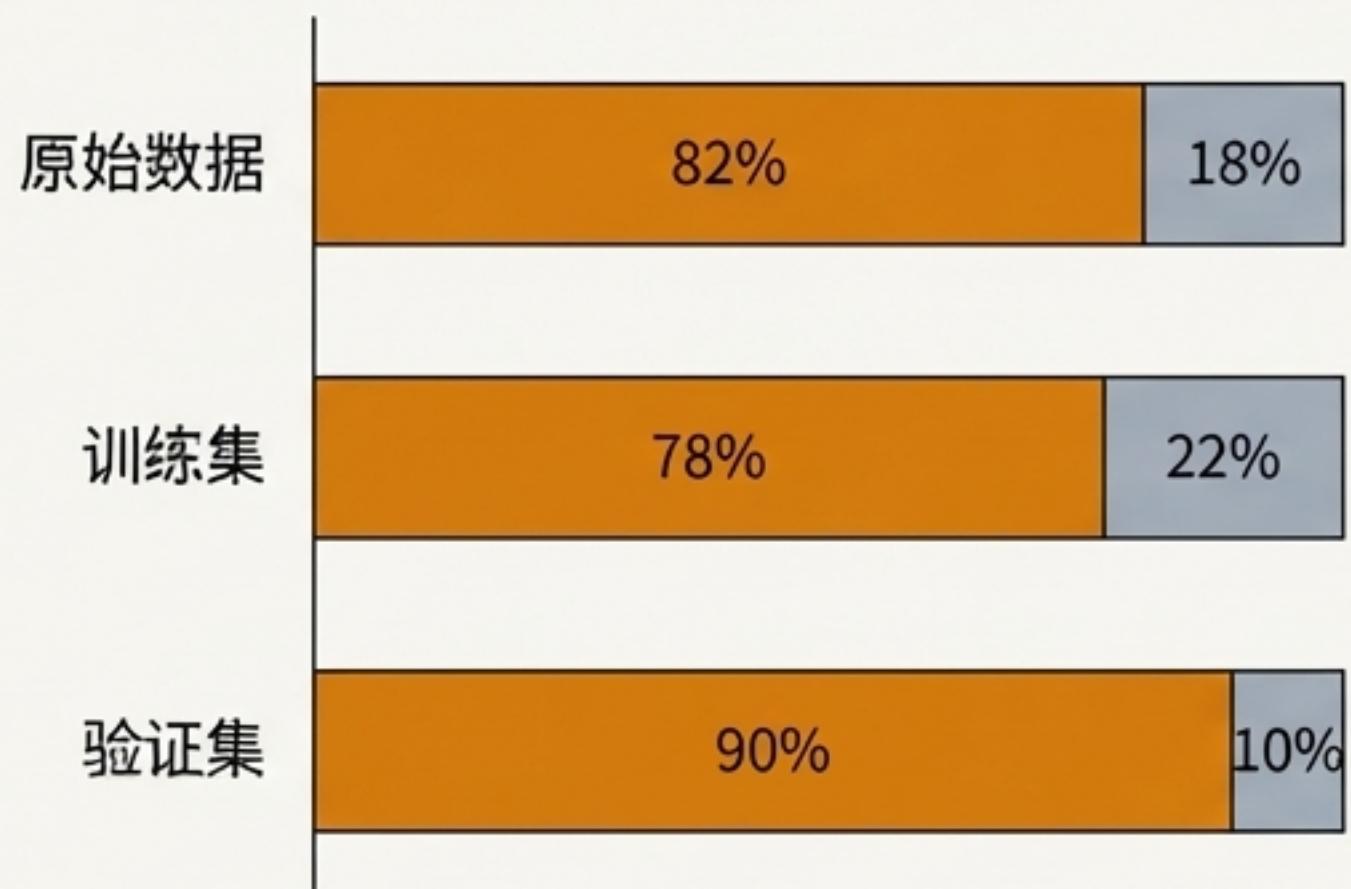


这种设计确保了正则化强度与特征的复杂性相匹配，从而有效抑制过拟合。

沉默的守护者：用分层采样保证数据一致性

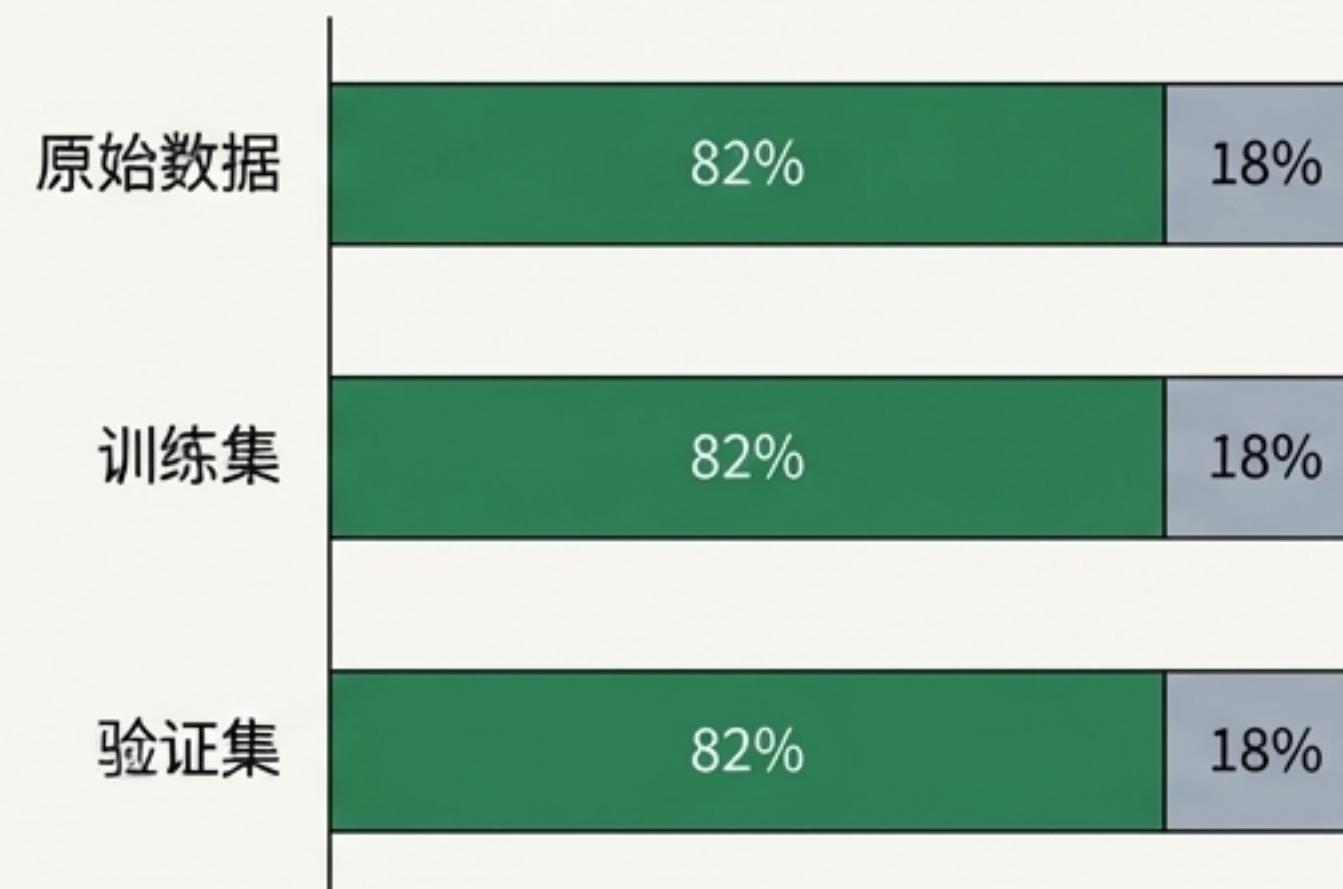
BBBP数据集存在典型的类别不平衡问题：约82%的分子为正例（能穿透），18%为负例。随机划分会导致训练集、验证集和测试集的类别分布不一致，使评估结果不可靠。

随机划分 (Random Split)



导致分布偏移，模型评估失真。

分层采样 (Stratified Split)



保证了评估的可靠性与公平性。

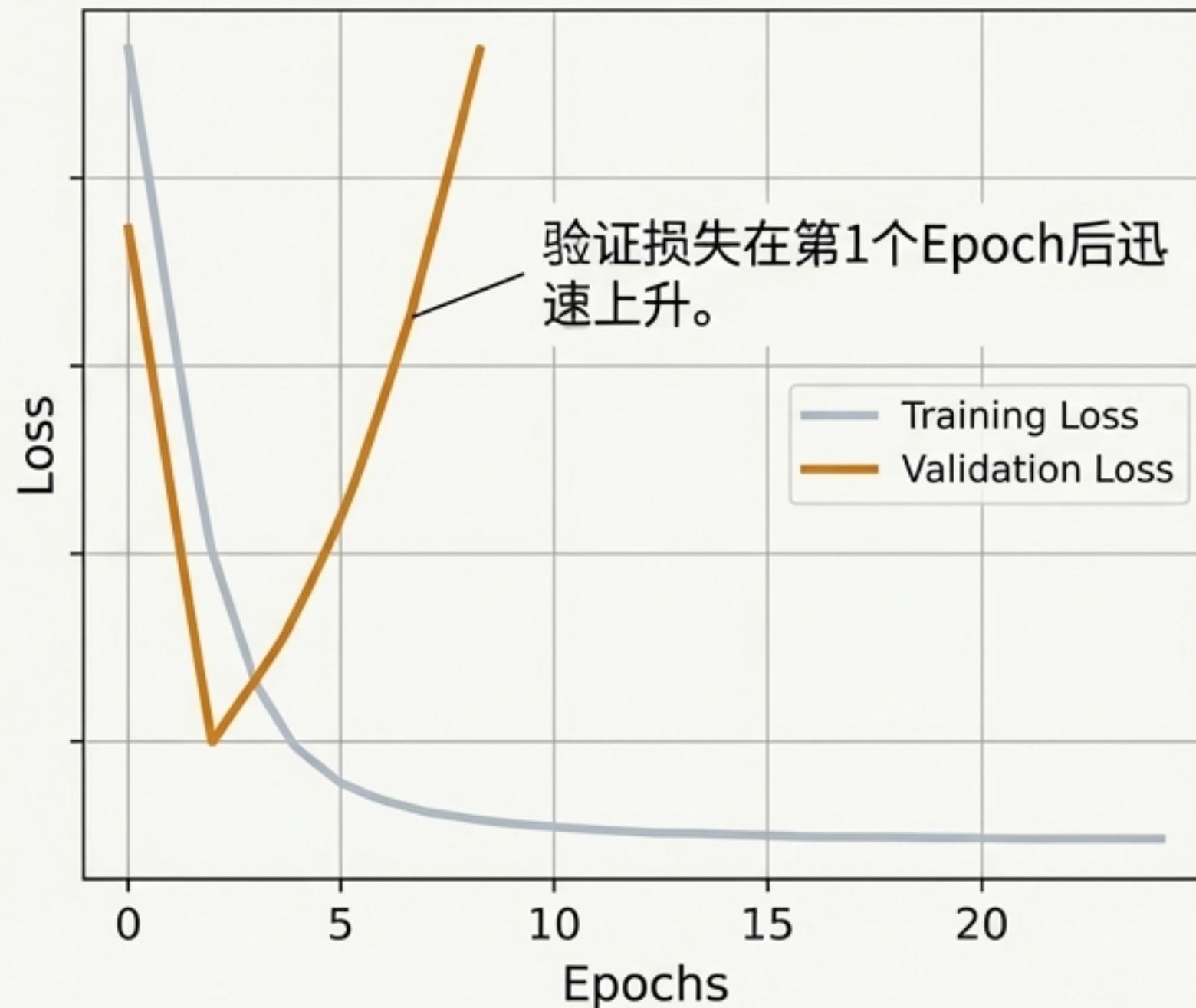
性能的飞跃：量化验证我们的解决方案

任务 (Task)	指标 (Metric)	基础模型 (MLP v1)	优化模型 (MLP v2)	提升 (Improvement)
BBBP	AUC-ROC	0.70	0.91	+30%
BBBP	F1-Score	0.69	0.92	+33%
ESOL	R ²	0.47	0.68	+45%
ESOL	RMSE	0.75	0.55	-27% (越低越好)

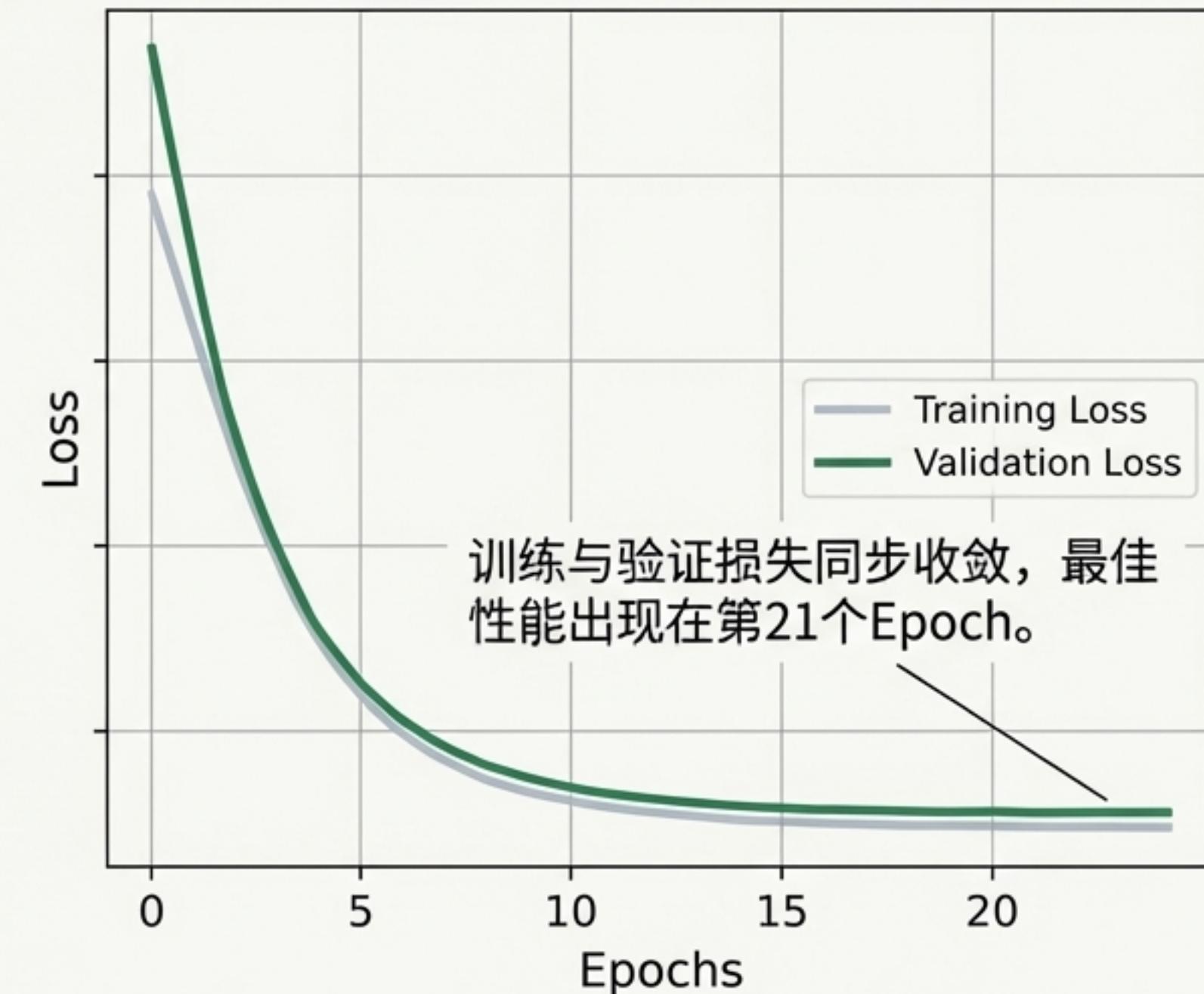
结合了分层采样与多重正则化策略的MLP v2模型，在分类和回归任务上均实现了性能的质变。

可视化证明：被驯服的过拟合曲线

优化前 (Before) : MLP v1



优化后 (After) : MLP v2

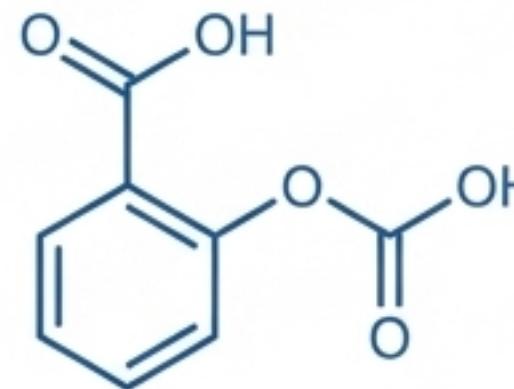


从模型到应用:一个功能完备的Web筛选工具

输入SMILES字符串
(e.g., 阿司匹林)

CC(=O)OC1=CC=CC=C1C(=O)O

自动生成分子2D结构图



Results

BBB穿透概率: 72.3%

	分子量	232.08
LogP	-7.03	
HBD	2.34	
HBA	0.89	

单分子预测界面
(Single Molecule Prediction UI)

支持上传包含SMILES的CSV文件

Upload CSV

SMILES	Score	Lipinski Violations
C1=CC=CC=C1...	0.94	0
C1CCNCC1...	0.88	1
C1=NC2=C(N1)...	0.76	0
C1=CC=CC=C1...	0.77	1
C1=NC2=C(N1)...	0.76	0
C1=CC=CC=C1...	0.94	1

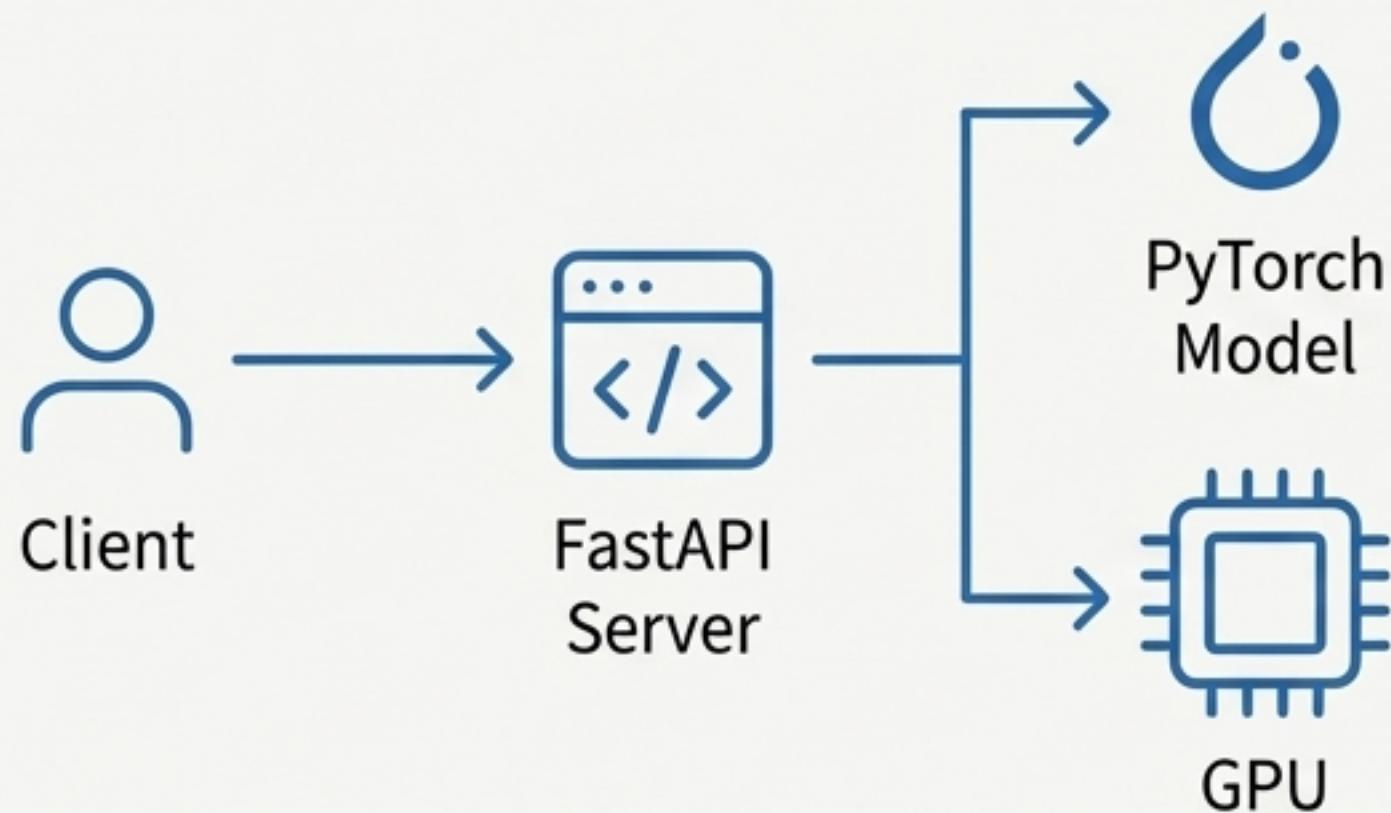
结果按预测分数排序，并进行Lipinski五规则检查

Download Results

批量筛选与结果展示
(Batch Screening & Results Display)

引擎盖下：为可扩展性而生的API服务

架构图



基于FastAPI构建的RESTful API服务，为自动化工作流和系统集成提供标准接口。

性能亮点

`POST /predict`：单分子预测

`POST /screen`：批量筛选

GPU加速带来20倍性能提升

在NVIDIA RTX 3050 Ti上，对10,000个分子的批量筛选任务，耗时从CPU的~60秒缩短至~3秒。

CPU: ~60s

GPU: ~3s

一个完整的解决方案：从架构到应用



模块化系统架构

设计了职责分明、松耦合的六层架构，确保代码的清晰度、可维护性与未来扩展性。



高性能预测模型

通过渐进式Dropout等创新正则化策略，成功攻克小样本过拟合难题，将BBBP任务AUC提升至0.91。



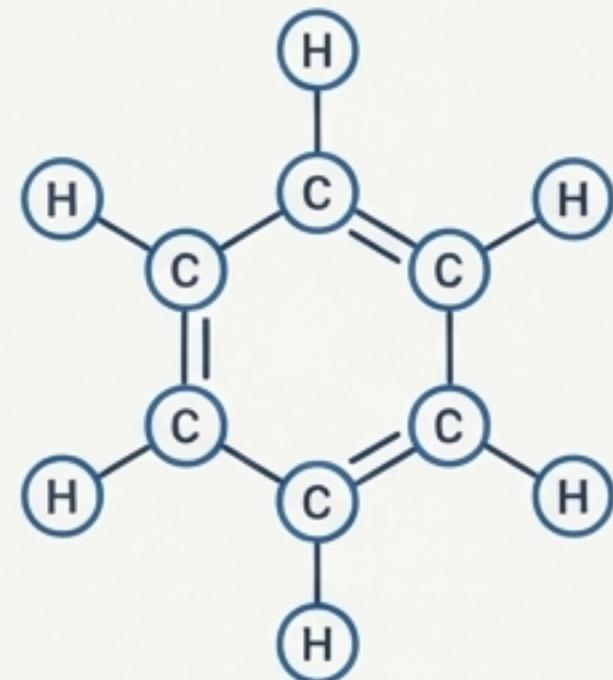
可部署的应用系统

提供了直观的Streamlit前端界面与稳健的FastAPI后端服务，将模型能力转化为实用工具。

未来图景：超越分子指纹的探索

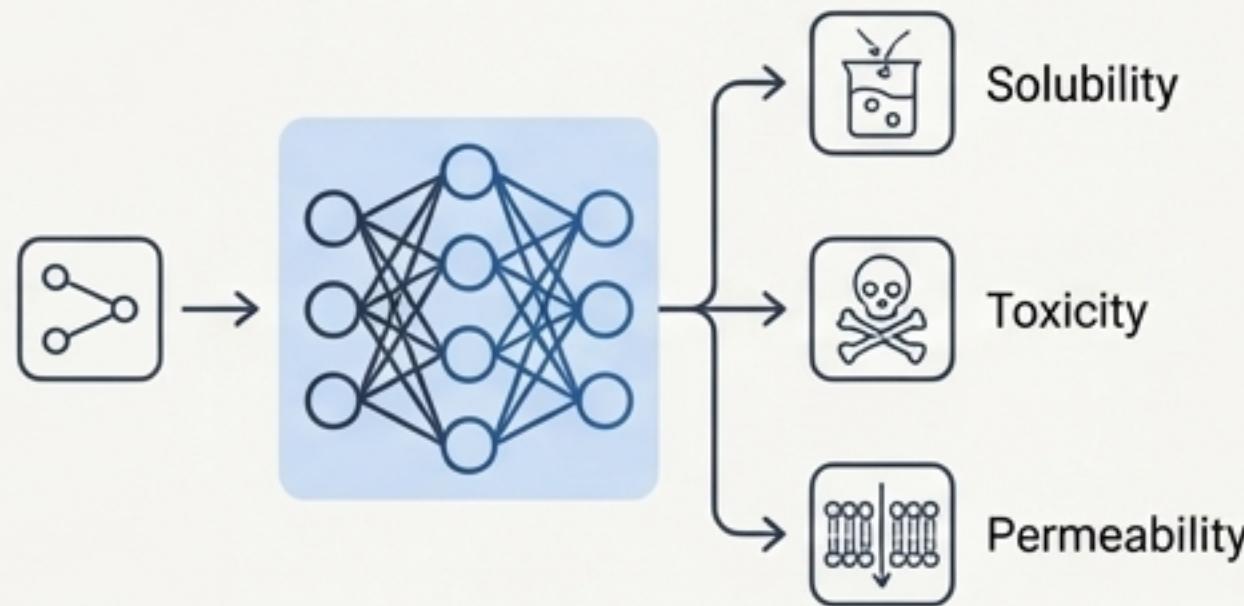
尽管MLP模型表现优异，但其依赖于信息有损的分子指纹。未来的工作将探索更先进的表示学习方法。

方向一：图神经网络（GNN）



直接对分子的图结构进行建模，无需预先计算指纹，能够学习原子间的相互作用，保留完整的结构信息以提高预测精度。

方向二：多任务学习



构建一个能同时预测多个ADMET属性（吸收、分布、代谢、毒性等）的统一模型。通过任务间的知识共享，缓解数据稀缺问题并提升整体性能。

通过攻克过拟合这一基础性挑战，我们构建了一个实用且强大的工具，为加速药物发现的漫长征程贡献了坚实一步。



github.com/project/drug-screener