基于ESM2和孪生神经网络的病毒与受体相互作用预测研究

传染病作为由病毒、细菌等病原体引发的全球性健康威胁，在人类历史上造成了深远影响。病毒性传染病因为其高变异性和传播复杂性，至今仍然对公共卫生安全构成严峻挑战。病毒入侵宿主细胞的关键步骤在于其与宿主受体的特异性结合，这一过程直接影响病毒的感染能力与致病机制。传统病毒受体研究方法如亲和层析、噬菌体展示技术等存在通量低、成本高等局限，难以应对日益增长的病毒多样性。随着蛋白质组学与微生物组学数据的爆发式增长，基于机器学习的预测方法为解析病毒-受体相互作用提供了新途径。然而，现有模型在特征表征深度、数据利用效率及跨物种泛化能力方面仍有不足，亟需发展更精准高效的计算模型以支撑抗病毒药物研发与新兴传染病预警。

研究内容：开发高精度的计算模型来预测病毒与宿主受体蛋白之间的相互作用

数据收集

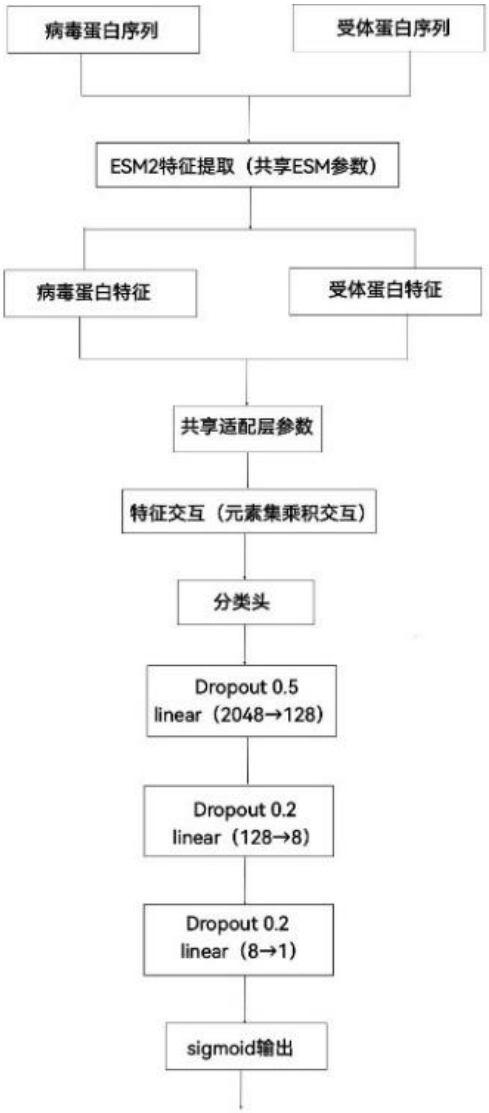
本实验中的数据集来源为 Sho Tsukiyama 等人在构建 LSTM-PHV 的模型中收集到的数据，这些数据可在 http://kurata35.bio.kyutech.ac.jp/LSTM-PHV 免费获得[错误!未找到引用源。]，LSTM-PHV 模型将在下文进行介绍。

Sho Tsukiyama 等人提供了多份数据，包括一份正样本，包含 22383 条数据；负样本集由 223,830 条非互作样本构成，构成 1:10 的正负样本比例；独立测试集涵盖 49,243 条验证样本；五折交叉验证集划分为五个独立子集，各子集训练队列含 157,576 条样本，对应验证队列为 39,394 条样本。

在实际参与训练时，本研究选择从正样本中抽取 7500 条，负样本中抽取 7500 条，维持 1:1 的类别均衡，构建 15000 条样本的训练集并按 1:9 比例分割为验证集与训练子集；测试集则沿用原始测试集数据，选取了 5000 条数据作为现有的测试集。

Ppi.py：实现了一个基于 ProtBert 蛋白质语言模型和 PyTorch Lightning 框架的蛋白质-蛋白质相互作用预测系统，通过自定义数据加载器处理蛋白质序列数据，构建包含预训练编码器和分类头的深度学习模型，支持训练、验证和预测全流程，并提供了可配置的参数和训练策略，为生物信息学中的蛋白质相互作用研究提供了高效的计算工具。

ppi_virhostnet_finetuning.py：将 Ppi.py 复现的 ProtBert 蛋白质语言模型改为一个基于 ESM-2 蛋白质语言模型，来对比两个模型的效果



esm+CNN.py：去除 Ppi.py 与 ppi_virhostnet_finetuning.py 复杂的框架，仅使用 esm 输出的 embedding，并使用简单的 CNN 框架来进行预测，查看模型效果

esm+cross_attention.py：去除 Ppi.py 与 ppi_virhostnet_finetuning.py 复杂的框架，仅使用 esm 输出的 embedding，并使用简单的 cross_attention 架构来进行预测，查看效果

结果：

	AUROC	Acc	F1	Pre	Recall
STEP	0.8412	0.7330	0.7329	0.8511	0.7329
ESM2+CNN	0.7657	0.7070	0.6440	0.8204	0.5300
ESM2+交叉注意力机制	0.7765	0.6830	0.5931	0.8280	0.4620
本模型	0.8842	0.7799	0.8000	0.8823	0.7799