nature machine intelligence

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s42256-020-0152-y

按照作者提供的格式，未经编辑。

利用类视觉问答系统预测药物与蛋白质的相互作用

按照作者提供的格式，未经编辑

补充材料

数据集详情

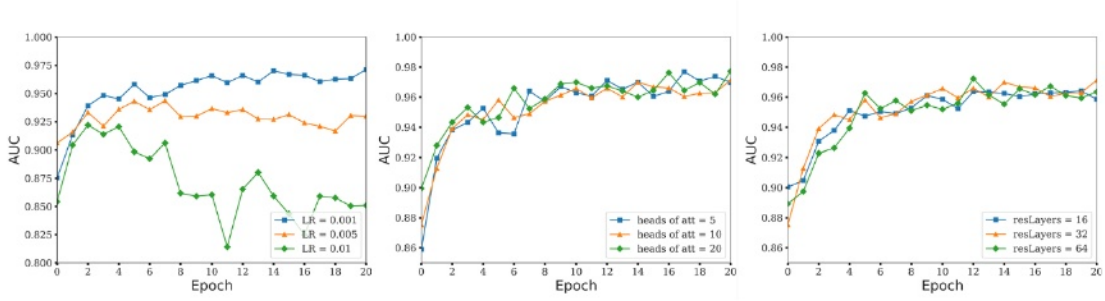
考虑到所用 GPU（GTX1080Ti 12GB）的内存有限，我们将蛋白质序列的最大长度设定为 1000 个字符（氨基酸）。该最大长度分别覆盖了原始 DUD-E、BindingDB 和 Human 数据集的 100%、91% 和 93%。原始数据集可从相关文献中轻松下载。我们从蛋白质数据库（PDB）获取结构数据。对于没有晶体结构的蛋白质，我们使用 blast 从 PDB 中检索序列同源性最高的同源蛋白质。序列同源性低于 40% 的蛋白质被排除在外。

神经网络训练及性能详情

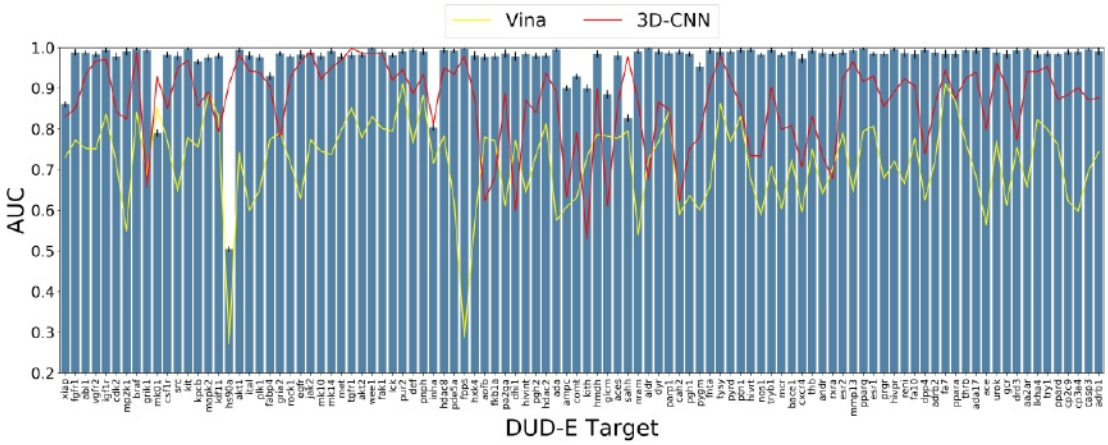
我们在补充表 1 中定义的超参数空间内进行了搜索。在人类验证集上评估的最佳超参数集以粗体突出显示。在人类验证集上使用不同超参数的学习曲线见补充图 1。我们注意到，在其他两个数据集上，除了学习率之外，其余超参数均未进行调整，因为我们发现模型性能对合理设置的超参数不敏感。

补充表 1. 超参数空间，最佳模型的参数以粗体显示。

关键参数	可能的值
卷积神经网络的残差块	16、**32**、64
双向长短期记忆网络的隐藏单元	**64**、128、256
注意力的隐藏单元	50、**100**、150
注意力的焦点	5、10、**20**
学习率	**0.001**，0.003，0.01
输出脱落	**0.2**，0.5，0.8



补充图 1. 在人类验证集上使用不同超参数的学习曲线。在所有学习曲线中，除非另有说明，我们使用以下超参数：双向长短期记忆网络（BiLSTM）的隐藏单元数 = 64，输出层的丢弃率 = 0.2，注意力机制的隐藏单元数 = 100，L2 正则化系数 = 0.001。

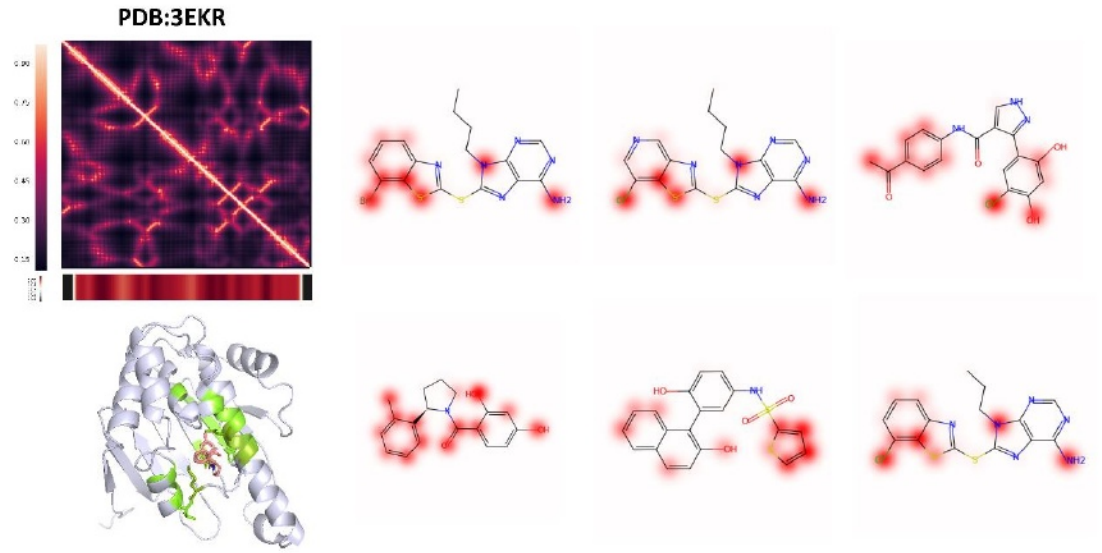


补充图 2. DrugVQA 模型在 DUD-E 基准测试中的交叉验证性能与 Vina 评分函数和 3D-CNN 模型的对比。

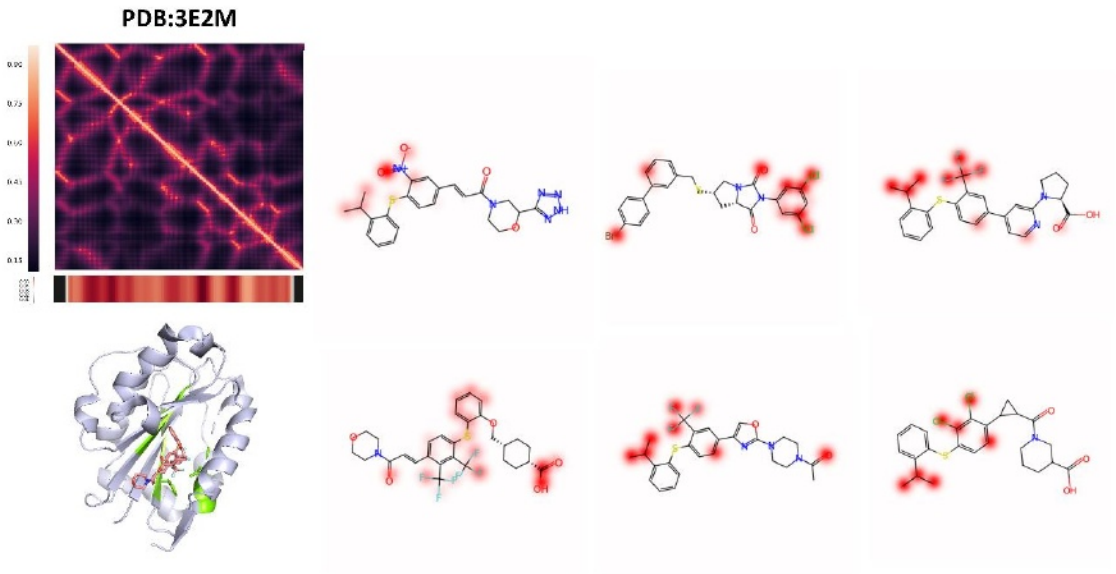
可视化细节

由于注释矩阵 A^p 和 A^m 的存在，该模型的解释十分直观。对于每个矩阵，我们对所有的注释向量进行求和，然后将得到的权重向量归一化为总和为 1。权重最高的部分分别对应药物分子和蛋白质中的原子和氨基酸，它们分别用绿色和红色进行了标注。

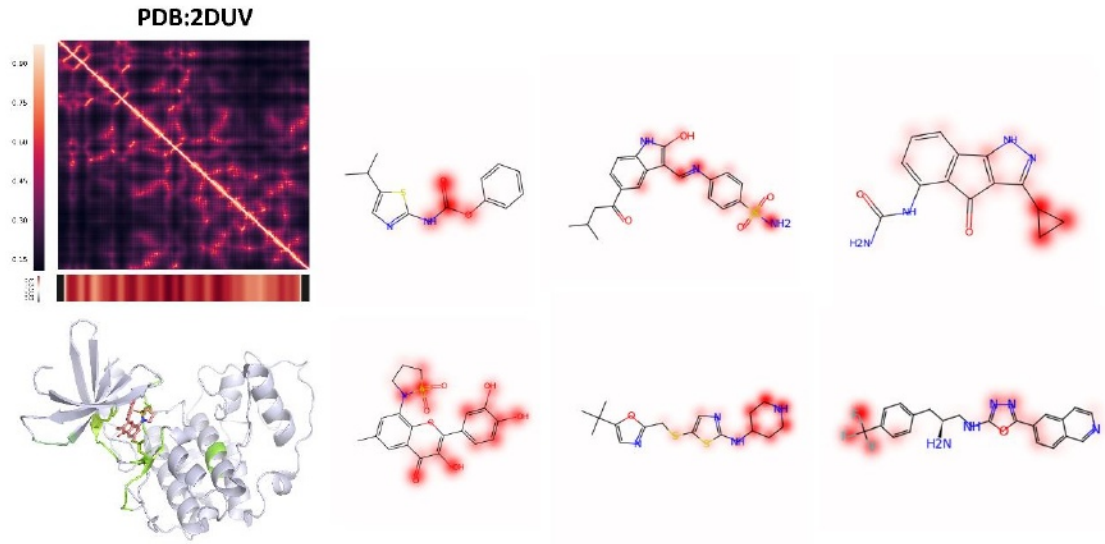
在此，我们展示了针对 CDK2、Hsp90 和 LFA-1 的更多示例可视化结果，以及它们的前几位预测活性分子（补充图 3 - 5）。我们将从蛋白质序列注意力图中检索到的前十五位贡献氨基酸用绿色标注，并将分子注意力权重映射到活性化合物的原子上，用红色表示。



补充图 3. CDK2（PDB：3EKR）的重要性可视化及其相应活性物质。



补充图 4. LFA-1（PDB：3E2M）的重要性可视化及其相应的活性位点。



补充图 5. 热休克蛋白 90（PDB：2DUV）及其相应活性位点的重要性可视化。