1.引言

在生物大分子中，蛋白质是一类特别引人关注的类别，因为它们参与了从结构和机械角色到信号传递和调节功能的大量细胞过程。为了发挥其生物学功能，蛋白质通常需要与其他分子直接相互作用，包括蛋白质/肽、核酸、膜、底物以及具有高度特异性和亲和力的小分子配体。对于新药的合理设计和开发，结构数据和蛋白质-配体结合机制的需求要求对分子识别的本质有深入的理解。[bernetti2017]

药物靶点相互作用网络的识别是药物发现流程中的一个基本步骤。[引用目录Knowles2003A]药物靶点相互作用网络的研究是药物开发的一个重要课题。仅通过实验来确定化合物-蛋白质相互作用或潜在的药物靶点相互作用既耗时又昂贵。作为补充，计算预测方法可以及时为我们提供非常有用的信息。[高维随机投影[12]]蛋白质相互作用网络等互作组中存在许多未被发现的实体间联系。在这些网络中验证每一个联系的实验成本极高，因此，指导搜索可能联系的计算方法具有极高的价值。[蛋蛋]

2.相关工作

3.数据整理及模型训练

3.1数据整理

BindingDB[ZeroBind[12]] 是一个 DTI 相互作用公共数据库，其中存储了药物（类药物分子）与靶蛋白之间的结合亲和力数据。该数据库目前包含超过 2,600,000 个经实验测定的蛋白质-药物复合物的结合亲和力，涉及 8,000 多个蛋白质靶点和 1,100,000 多种小分子。

为了创建 ZeroBind 的训练和测试数据集，我们采用了多个过滤和预处理步骤来创建高质量的基准数据集。首先，数据点的 “目标类型 ”属性为 “单个蛋白质”，“标准类型 ”属性为动力学常数 Ki、Kd、IC50 和 EC50。此外，所有目标蛋白质都应是人类或类似人类的蛋白质，因此在 “目标源生物 ”属性中使用 “智人 ”进行筛选。在排除了没有 SwissProt 名称的蛋白质和 RDKit[ZeroBind [44]] 无法处理的分子后，我们收集到了 1,500,000 对蛋白质-药物配对。我们使用 AI-bind[ZeroBind [9]] 中的阈值，将动力学常数 Ki、Kd、IC50 和 EC50<1000 nM 视为阳性样本，将 >106 nM 视为阴性样本。[ZeroBind]

BindingDB是一个庞大的、可公开访问的蛋白质-配体结合数据知识库。我们从BindingDB官网获取了BindingDB\_All\_202411\_tsv.zip，其中包含了截止至2024年11月的BindingDB中的所有结合数据。将其中的异常值，模糊值，及无法

3.2 异常数据排除

在本文中，我们首先在训练集上使用自编码器进行异常检测，以获取包含反应条件的蛋白质-配体相互作用数据集

显然，药物结合的一个重要特征是高亲和力、大负自由能（ΔG）和慢解离动力学（koff）。然而，要使药物成为安全的药物，其结合必须提供选择性反应，从而达到可耐受的治疗指数。在平衡态下的竞争反应被认为是热力学控制的(平衡常数和浓度)，而那些不在平衡态下的反应被认为是动力学控制的(竞争的动力学速率)。[10.1002@9783527673025.ch16]

几十年来，解离常数 Kd 及其替代指标IC50/EC50（即导致生物活性半数抑制的药物浓度）一直被认为是体内药效的有效替代指标。[bernetti2017] Kd是配体亲和力和选择性比较以及预测体内有效药物浓度的重要基础参数。药物-受体相互作用的热力学研究需要对反应条件，进行仔细的标准化，这样才能对来自不同实验室的不同数据集进行有意义的比较。反应条件，如温度及PH值，可能会极大地影响配体的亲和力。[British J Pharmacology - 2010 - Hulme - Ligand binding assays at equilibrium validation and interpretation]因此，结合反应条件进行药物-受体相互作用模式的学习，更具有实际意义。传统上，平衡结合参数，即半数最大效应所需浓度（EC50 或 IC50）和抑制常数（Ki），通常用于评估药物与目标的相互作用。[British J Pharmacology - 2023 - Liu - The translational value of ligand‐receptor binding kinetics in drug discovery (1)]

我们假定AA所报道的b数据集bindingd是相对无噪声的。为了整合数据集bindingdb中的相互作用模式及BindingDb中的反应条件及实验测定值(如Kd,EC50,IC50及Ki等)，我们根据配体及靶标序列的完全一致性对这些数据集进行匹配。具体而言，对于每一组来自各数据集的记录，若配体及靶标序列均完全一致，则将该记录合并。经过该操作，最终获得的数据集中既包含了反应条件及测定值信息，又包含了相互作用模式判定，且该数据集是相对无噪声的。

实验测定值可以通过阈值转换为二分类标签，即相互作用模式(如0或1,1表示药物-配体之间存在积极的相互作用，0表示其之间存在消极的相互作用)。[1 受体结合测定值的阈值转换 3 机器学习中的阈值学习] 结合亲和值通常用解离常数（Kd）、抑制常数（Ki）和半最大抑制浓度（IC50）等指标来表示。[Neural Networks2024-AttentionMGT-DTA A multi-modal drug-target affinity prediction using graph transformer and attention mechanism] 但是，由于实验条件不一致等原因，尽管该分类标签大多是依据Kd, IC50, Ki, EC50的值进行设定，但是关于阈值的设置并未统一[2 阈值的选择标准]。例如，[oup-accepted-manuscript-2016]使用其固定的阈值进行二分类。而[jain1996]使用统计模型优化阈值。因此，为了识别新数据集中的实验测定值与相互作用模式之间的阈值关系，减少数据融合所带来的冲突，我们使用决策树及逻辑回归对数据中潜在的阈值关系进行学习，并采用异常检测剔除异常数据。

3.3 模型搭建

3.3.1编码

药物化合物通过功能团进行编码，而蛋白质则通过包括生化和物理化学性质在内的生物学特征进行编码。

4.实验

对比算法包括  
(1)drugVQA(2020Nature machine intelligence: Predicting drug-protein interaction using quasivisual question answering system)

(2)DrugBAN(2023 Nature machine intelligence:DrugBAN)

(3)mlanDTI(2024AAAI: Multilevel Attention Network with Semi-supervised Domain Adaptation for Drug-Target Prediction)

(4)PretrainDPI [ 2021Bioinformatics: Bayesian neural network with pretrained protein embedding enhances prediction accuracy of drug-protein interaction ?]

(5)

5.未来设想

药物设计中结合效应通常通过平衡结合来评估，并且在大多数情况下反映了药物与药物靶标结合时在晶体结构中观察到的结合模式[。[10.1002@9783527673025.ch16]](mailto:。[10.1002@9783527673025.ch16])

从热力学角度讲，优化结合亲和力意味着提高结合自由能，结合自由能有两个组成部分：结合焓和结合熵。结合焓和结合熵可能对结合自由能产生积极或消极的影响，进而影响结合亲和力。特异性相互作用（如氢键、盐桥、范德华接触）会带来焓的增益，而极性基团的脱溶则导致焓的损失。熵的增益通常与配体与靶蛋白结合时的脱溶有关，而配体侧和受体侧的构象变化则会产生熵的损失。

基于这些考虑，优化工作可能会导致不同的热力学结果。当具有最佳几何结构的特定极性相互作用形成时，极性基团的引入可以提高结合焓（图16 - 4a）。如果极性基团的位置不正确，那么基本上会实现极性去溶的焓损失（图16 - 4b）。当极性空腔被填充时，添加非极性基团可能是焓有利的（图16 - 4c）。由于无极性相互作用通常不太依赖方向，对束缚腔的次优拟合仍然会给予一定的熵增益；然而，空间类别——尽管存在配体去溶带来的熵增益——导致了剧烈的焓损失（图16 - 4d）。因此，热力学数据最好与结构信息联系起来使用，这有助于理解化合物逐步结构优化的热力学结果。

热力学会影响化合物的物理化学性质，从而影响其吸收、分布、代谢和毒性(ADMET)特征，此外还会影响配体结合的亲和力。