



ZeroBind: 一种针对蛋白质特异性的零样本预测器，采用子图匹配来预测药物-靶点相互作用

收到时间: 2023 年 4 月 15 日

收稿日期: 2023 年 11 月 13 日

Published online: 29 November 2023

 检查更新Yuxuan Wang¹, Ding Xia¹, Junqi Yan², Ye Yuan², Hong-Bin Shen¹ & Xiaoyong Pan¹  

现有的药物-蛋白质相互作用 (DTI) 预测方法通常无法很好地泛化到新的 (未见的) 蛋白质和药物。在本研究中, 我们提出了一个蛋白质特异性元学习框架ZeroBind, 它使用子图匹配来预测蛋白质-药物相互作用, 从它们的结构中预测。在元训练过程中, ZeroBind制定了一个蛋白质特异性模型的训练任务, 这也被视为一个学习任务, 每个任务使用图神经网络 (GNNs) 来学习蛋白质图嵌入和分子图嵌入。受分子与蛋白质中的结合口袋结合而非整个蛋白质这一事实的启发, ZeroBind引入了一个弱监督子图信息瓶颈 (SIB) 模块, 以识别蛋白质图中的最大信息量和压缩性子图作为潜在的结合口袋。此外, ZeroBind将单个蛋白质的模型作为多个任务进行训练, 其重要性通过任务自适应注意力模块自动学习, 以做出最终预测。结果表明, ZeroBind 在 DTI 预测方面的表现优于现有方法, 尤其是对于那些未见过的蛋白质和药物, 并且在对那些已知结合伴侣较少的蛋白质或药物进行微调后表现良好。

确定药物与靶点 (蛋白质) 之间的相互作用在药物发现过程中起着至关重要的作用¹⁻³。然而, 传统的实验方法用于解析药物-蛋白质复合物的晶体结构以确定药物-靶点相互作用既昂贵又耗时³⁻⁵。为了降低成本, 计算方法越来越受到关注。与其将大量候选物带入体外筛选, 使用计算方法在体外筛选之前虚拟筛选掉大多数候选物更高效且成本更低。一般来说, 计算方法分为两大类, 对接模拟和基于数据驱动的学习方法。对接模拟利用药物分子和靶蛋白的三维结构来确定它们的潜在结合位点, 这也非常耗时^{1,6,7}。相比之下, 由于机器学习⁸的快速发展, 利用

从蛋白质和药物中提取的特征来确定它们之间的相互作用, 既能实现高精度, 又能降低成本。

数据驱动的学习方法通常将药物-靶点相互作用 (DTI) 预测定义为二元分类或回归任务^{8,11}, 其中从现有的数据库如BindingDB¹²、ChEMBL¹³、PDBind¹⁴、15和DrugBank¹⁶中提取蛋白质和药物的相互作用对。由于效力值的性质是对数性的, 从微摩尔到纳摩尔水平的动力学常数的降低会导致指数级变化。因此, 回归任务的输出值通常被定义为动力学常数的负对数, 即 K_i 、 K_d 、 IC_{50} 和 EC_{50} 。对于分类任务, 根据 K_i 、 K_d 、 IC_{50} 和 EC_{50} , 17, 18设定阈值来定义结合或非结合。

¹Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China. ²Department of Computer Science and Engineering, and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China. e-mail: 2008xypan@sjtu.edu.cn



基于机器学习的从分子和蛋白质特征中获取的方法主要集中于学习分子和蛋白质的良好表示，然后将这些表示输入分类/回归模型以执行预测任务^{8–11, 19}。

最近，深度学习在通过已知药物-靶点相互作用来学习来预测药物-靶点相互作用方面取得了令人兴奋的结果。然而，这些方法在那些未见过的蛋白质和药物上不能很好地泛化。基于相似性/距离的^{20–23}和基于网络的^{24–26}方法利用蛋白质-蛋白质相似性、药物-药物相似性和已知的药物-靶点相互作用，在转捩测试中表现过度估计，其中测试集中的蛋白质和配体都出现在训练集中。此外，对于新发现的靶蛋白，通常已知结合的药物很少，这使得在归纳测试中难以开展工作，在归纳测试中测试集中的蛋白质和配体都不在训练集中。目前，大多数现有方法都集中在如何有效地学习分子和蛋白质的表示，然后将表示输入分类或回归模型^{8–11, 19}。像DeepPurpose⁸这样的表示学习方法将分子指纹和SMILES27字符串作为分子特征，氨基酸序列作为蛋白质特征。然后，他们应用卷积神经网络（CNN）²⁸、循环神经网络（RNN）²⁹或Transformer³⁰模型来嵌入输入特征。为了考虑空间信息，一些方法将卷积神经网络应用于从蛋白质和分子结构中得出的三维图像³¹。近年来，图神经网络（GNN）³²因其能够处理图作为非欧几里得结构化数据³³而备受关注。其他方法利用图神经网络对蛋白质和分子的图进行嵌入以进行药物-靶点相互作用（DTI）预测¹⁰。由于药物和靶蛋白都是天然的图结构数据，DTI预测也受益于图神经网络的快速发展。GEFA34将预训练的蛋白质嵌入和图内图神经网络与注意力机制相结合，以捕获药物和蛋白质残基之间的相互作用。因此，使用图神经网络进行蛋白质和药物的表示学习是合理的。

然而，经典的成对输入方法通常将药物-蛋白质对作为训练样本，通过连接并将其表示输入到密集层来识别相互作用。然而，这些方法在针对模型训练期间未见过的蛋白质和药物的归纳测试中可能会失败，这是由于它们可能存在潜在的捷径³⁵，记住了训练集中结合注释的程度比率，而不是学习了相互作用的分子特征⁹。为了对未见过的蛋白质和药物进行泛化，AI-Bind利用网络推导的负样本和预训练来解决捷径学习问题⁹。此外，蛋白质可能与药物具有不同的结合模式，这可以通过为单个蛋白质训练蛋白质特异性模型来捕获。最近，在元学习框架下，开发了一种基于配体的零样本预测方法MetaDTA³⁶，用于具有少量已知结合药物的蛋白质的少量样本预测。然而，MetaDTA不支持对训练集中没有已知结合药物的未见过的蛋白质的零样本预测。

为了应对这些挑战，我们在元学习框架下提出了一种用于药物-蛋白质相互作用预测的特定蛋白质零样本预测器ZeroBind。ZeroBind采用元学习框架MAML++³⁷作为训练策略，每个基础模型对特定蛋白质的结合药物进行预测，其主要包含四个模块：

- （1）图卷积网络（GCN）编码器学习分子图和蛋白质图的嵌入；
- （2）子图信息瓶颈（SIB）模块生成蛋白质图作为潜在结合口袋的关键IB子图；
- （3）多层感知机（MLP）模块将蛋白质IB子图嵌入和分子嵌入连接起来进行DTI预测；
- （4）任务自适应自注意力模块用于衡量不同任务的重要性，其中不同的DTI任务对元学习器的贡献不同。

生成的任务权重用于加权平均损失，并进一步纳入元学习过程。

总之，我们的贡献如下：（1）我们将未见过的药物和蛋白质的药物-蛋白质相互作用预测作为一个零样本学习问题。从现有蛋白质、药物及其相互作用中通过元学习策略学习到的关于药物-蛋白质相互作用预测的一般知识，比现有方法对未见过的蛋白质和药物进行泛化的能力更强。（2）我们为每个蛋白质训练一个药物-蛋白质相互作用任务模型，其中任务自适应自注意力被设计用于计算多个药物-蛋白质相互作用任务对蛋白质特定元学习器的贡献。因此，每个蛋白质特定模型捕获对药物的个体结合模式。（3）我们提出了一种模型不可知的IB子图学习，以自动发现蛋白质中作为潜在结合口袋的压缩子图，而不是从整个蛋白质中推导出的冗余图信息。（4）我们在三个独立的零样本测试集和一个少量样本测试集上进行了广泛的实验。结果表明，ZeroBind始终优于现有方法。对现实世界中新冠病毒（SARS-CoV-2）药物靶点结合预测的进一步验证表明，ZeroBind预测的可靠性，并且通过IB子图学习检测到的子图与蛋白质中已知的结合位点高度吻合。

结果

零绑定概述

在本研究中，ZeroBind将DTI预测表述为一个元学习任务，并提出了一个元学习框架来解决DTI预测中未见过的蛋白质和药物的泛化问题。具体而言，元学习任务被定义为特定蛋白质的结合药物预测，其中利用IB子图学习来自动发现蛋白质中作为潜在结合位点的压缩子图，并设计了一种自注意力机制来学习蛋白质每个任务的权重。ZeroBind的流程图如图1所示。

具体来说，ZeroBind使用基于网络的负采样作为数据扩充来缓解标注不平衡（图1a，方法部分）。图1b、c展示了训练集中基于网络负采样前后正样本的比例，表明基于网络的负采样在一定程度上缓解了标注不平衡。然后，它将DTI采样到支持集和查询集中（图1d），其中支持集用于训练元学习器，查询集用于训练特定任务的模型。重复N个内部步骤后，对所有损失进行加权，使用梯度下降来优化元学习器。对于每个蛋白质，ZeroBind训练一个DTI预测任务。图1e给出了ZeroBind中基础模型的架构，其中蛋白质图和分子图被输入到一个骨干GCN中，以学习药物和蛋白质的嵌入。此外，还设计了一个弱监督子图信息瓶颈（SIB）模块来对蛋白质中潜在的结合位点进行建模和发现。SIB模块不仅减少了冗余信息以提高性能，而且还通过识别蛋白质中的关键残基为ZeroBind带来了可解释的见解。图1f引入了一个自适应自注意力模块来衡量蛋白质各项任务的贡献，其中不同的DTI任务对元学习器的贡献不同。ZeroBind支持在零样本和少样本场景中预测DTI。前者使用元学习器直接进行预测，无需使用元测试中蛋白质的样本进行微调，而后者使用特定于蛋白质的模型进行预测，并使用元测试中蛋白质的样本进行微调。

在用于弥散张量成像（DTI）预测的零样本和少量样本环境中，ZeroBind均优于现有的方法。

为了展示ZeroBind的优势，我们将其与多种基准方法进行了比较，并计算了受试者工作特征曲线下面积（AUROC）和精度曲线下面积。

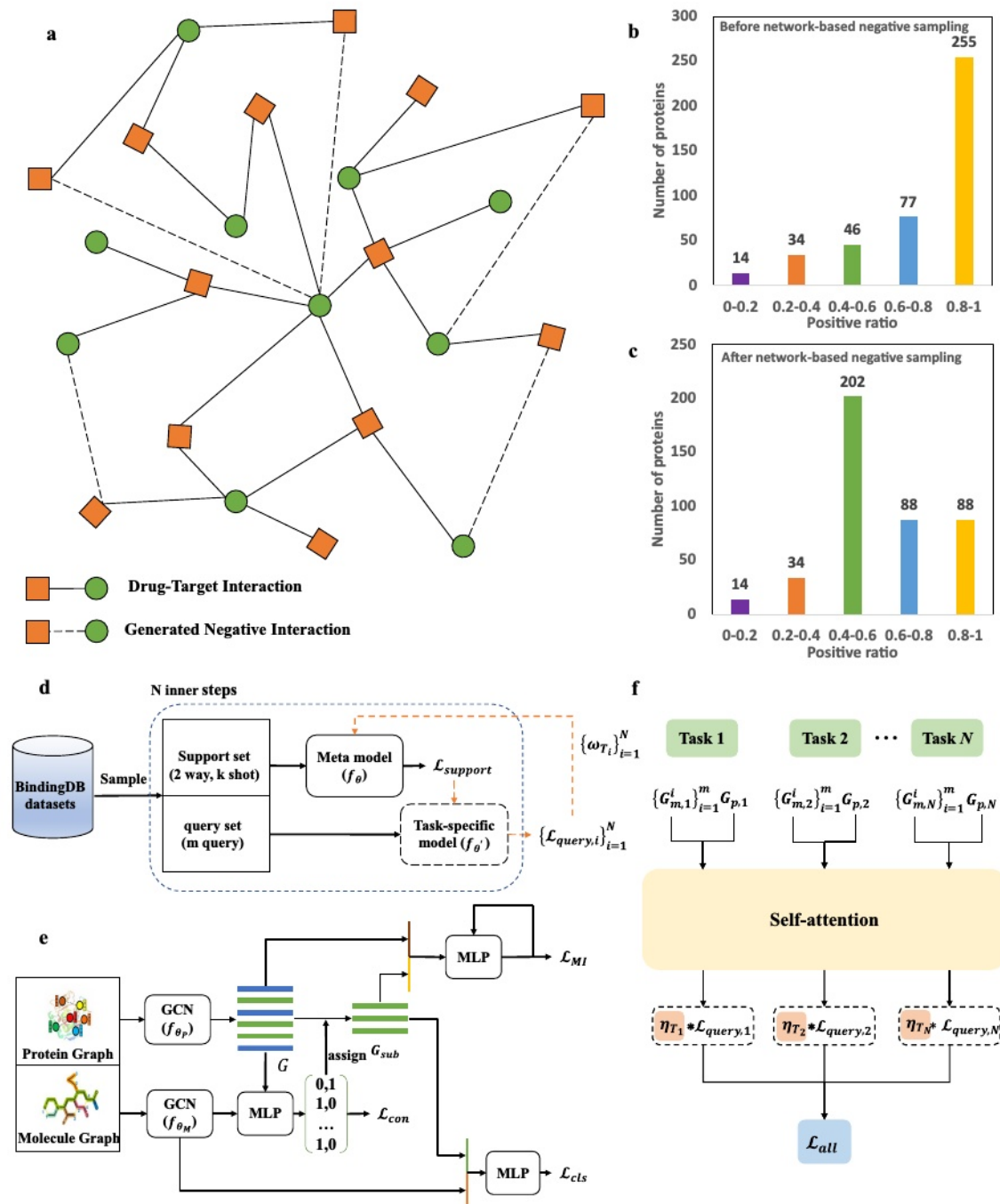


图 1 | ZeroBind 的框架。 **a** 基于网络的负采样策略。由药物和蛋白质靶点组成的二分网络：方形节点表示蛋白质节点，圆形节点表示分子节点，并且只有不同类型节点之间存在边，表示相应的药物-靶点相互作用。实线表示现有的药物-靶点相互作用 (DTIs)，虚线表示具有最短路径距离 ≥ 7 的生成负相互作用。 **b** 基于网络负采样策略前训练集的正比例。 **c** 基于网络负采样策略后训练集的正比例。 **d** 给定支持集和查询集，首先计算 $\mathcal{L}_{support}$ ，并使用支持集利用参数 θ 将基础模型更新为具有参数 θ 的任务特定模型。

对于每个任务，然后特定任务的模型使用查询集 n of the n task 来计算 \mathcal{L}_{query} 。在重复 N 个内部步骤之后，所有损失都由 ω_{T_i} 加权平均，并且进一步执行梯度下降以优化元模型。

e 零结合 (ZeroBind) 中基础模型的架构。对于每个任务，蛋白质图和分子图分别被输入具有参数 θ_p 和 θ_m 的骨干图卷积网络 (GCN)，以获得它们的嵌入。随后，提出了一个 SIB 模块，以弱监督的方式生成蛋白质的 IB 子图作为潜在的结合位点。蛋白质子图嵌入与分子嵌入相连接，并被输入到多层感知机 (MLP) 模块中以识别相互作用。 **f** 任务自适应注意力模块。它采用蛋白质嵌入 G_p 、 k 和分子的连接 n 。

将查询集中所有分子嵌入 G_m ， k 的平均值作为任务 $i=1$ 。

embedding n After n using 自注意力层计算每个任务的权重，记为 η_{T_i} ，总体损失经过平均并纳入 $meta_i = 1$ 中

用于更新模型参数的训练过程。原始数据以“源数据文件”的形式提供。

在三个独立测试集和一个少量样本测试集上计算召回曲线 (AUPRC)。为确保性能比较的有效性,我们对数据集划分和模型训练进行了五次独立实验,使用了五个随机种子,并报告了平均结果以及标准偏差。

所有方法在三个独立测试集上的总体性能如图 2a (补充表 1) 所示。ZeroBind 在 Transductive、Semi-inductive 和 Inductive 测试集上的 AUROC 分别为 0.9521 (± 0.0034)、0.8681 (± 0.0052)、0.8139 (± 0.0035)。我们可以看到,在 Transductive、Semi-inductive 和 Inductive 测试集中,ZeroBind 优于所有基准方法。与最佳基准方法相比,ZeroBind 在 AUROC 上的相对改进在 Transductive 测试集中为 2.86%,在 Semi-inductive 测试集中为 10.29%,在 Inductive 测试集中为 3.38%,相应的 t 检验 p 值分别为 1.02×10^{-6} 、 8.02×10^{-11} 、 3.06×10^{-7} 。此外,与最佳基准方法相比,AUPRC 的相对改进在 Transductive 测试集中为 1.00%,在 Semi-inductive 测试集中为 0.96%,在 Inductive 测试集中为 1.21%,相应的 t 检验 p 值分别为 8.93×10^{-8} 、 5.15×10^{-3} 、 1.56×10^{-5} 。此外,在这三个测试集中,我们观察到所有方法的性能都有所下降,这表明 Transductive 测试集中存在一定的外部捷径学习。

图 2a 显示,由于结合了基于网络的负采样和无监督预训练嵌入以避免潜在的捷径学习,AI-Bind 和我们提出的 ZeroBind 在归纳测试集上的表现优于其他基准方法。ZeroBind 在转导测试集和半归纳测试集上的表现比基准方法更稳定且更好,这表明 ZeroBind 能够有效地学习蛋白质和药物的有用嵌入。此外,ZeroBind 在归纳测试集上的表现优于 AI-Bind。潜在的原因是,ZeroBind 使用元学习框架获取了多个蛋白质中 DTI 预测的一般知识,能够很好地泛化到未见过的蛋白质和药物的 DTI 预测。

我们还在诱导测试集和半诱导测试集上评估了 ZeroBind 与其他基准方法在蛋白质特异性 DTI 预测性能方面的表现 (图 2b)。在排除没有结合或非结合标签的蛋白质后,我们获得了 775 种结合药物的蛋白质,并为每种蛋白质

训练了一个任务模型,在此基础上,我们在组合诱导和半诱导测试集上对 ZeroBind 进行了评估。ZeroBind 在 775 种蛋白质上的平均 AUROC 为 0.78,高于 DeepConv-DTI 的 0.66、GraphDTA 的 0.68、Deeppurpose 的 0.69、AI-bind 的 0.75 和 DrugBAN 的 0.73。在这 775 种蛋白质中,ZeroBind 在 525 种、491 种、305 种、180 种和 346 种蛋白质上优于其他基准方法。对于每种方法,我们根据结合分子的数量展示了该方法优于其他方法的数量 (图 2c)。我们观察到,在三个范围内,ZeroBind 在大多数蛋白质上优于其他基准方法,尤其是对于只有 1-10 个已知结合分子的蛋白质,其次是 AI-Bind。

我们进一步在少样本测试集上对 ZeroBind 与基准方法进行了评估。首先,ZeroBind 和其他基准方法使用训练集作为预训练模型进行训练,然后在每个少样本微调集上进行进一步的微调,并在相应的少样本测试集上进行评估。ZeroBind 模型和其他基准方法在少样本测试集上的性能如图 2d (补充表 2) 所示。在少样本测试中,ZeroBind 优于最佳基准方法,相对提高了 AUROC 1.55% 和 AUPRC 1.62%。潜在原因是元学习框架具有强大的泛化能力,能够快速适应具有少量训练样本的其他蛋白质任务。我们可以看到,微调使零样本预测性能大幅提高,这对于蛋白质只有少数已知结合药物的现实应用场景非常有用。

ZeroBind 以弱监督的方式检测与蛋白质已知结合位点高匹配的子图。

ZeroBind 并不将结合口袋信息用作模型训练的真实标签,而是将全局 DTI 标签用作弱监督标签。为了证明 ZeroBind 中的 1B 子图模块能够检测到蛋白质中的结合口袋,我们使用 Jaccard 相似系数³⁸在 PDBbind 数据集^{14,15}上将预测的结合口袋与真实的结合口袋进行比较。包含口袋残基和三维坐标位置的真实结合口袋信息是从 PDBbind 下载的,其中包含 535 种蛋白质的 14,000 个结合残基。Jaccard 相似性

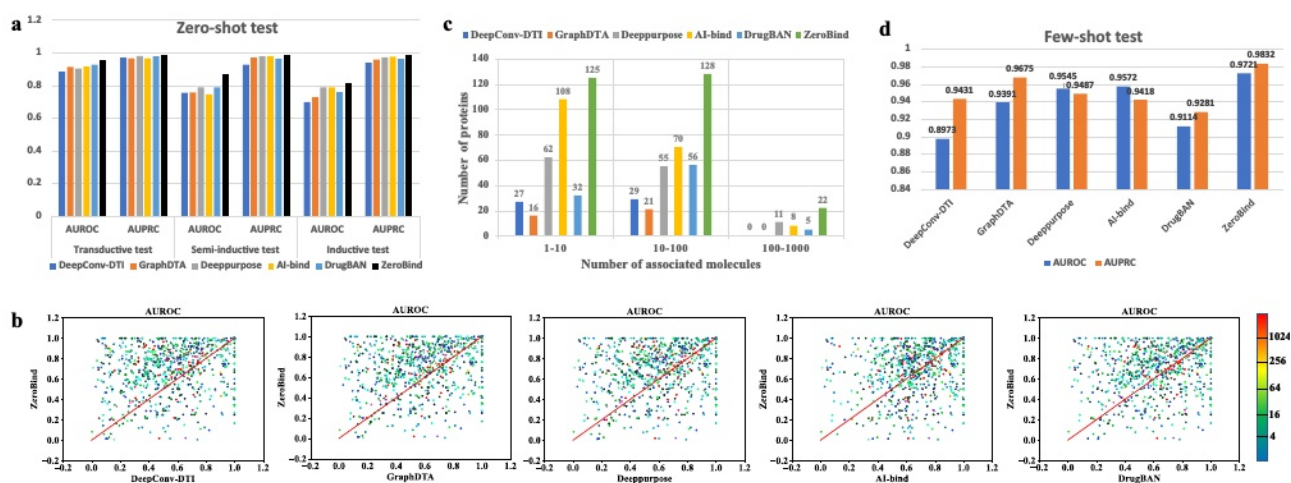


图 2 | 在零样本和少量样本场景中, ZeroBind 与基准方法的性能比较。a 在三个独立测试集上对 ZeroBind 和基准方法进行零样本性能评估。b 在合并归纳和半归纳测试集上对 775 种蛋白质的蛋白质特异性 ZeroBind 和基准方法的受试者工作特征曲线下面积 (AUROC) 比较。点的颜色代表训练蛋白质的数量。c 该方法表现最佳的方法中蛋白质的数量

对组合归纳和半归纳测试集的方法进行了比较。d 在少量测试集上对 ZeroBind 与基准方法进行了少量样本性能比较。补充表 1 和 2 提供了零样本和少量样本预测的数据统计,采用双侧 t 检验且未进行调整。源数据以源数据文件的形式提供。AUPRC 指的是精确率 - 召回率曲线下面积。

系数用于计算两组个体 DTI 之间的相交程度，其公式定义为

$$\delta P, P \cap \delta = \frac{P \cap \delta}{P \cup \delta}$$

贾卡德相似系数 $\delta P, P \cap \delta = \frac{P \cap \delta}{P \cup \delta}$, $P \cup \delta$ 表示集合的并集 $\cup P$

通过 ZeroBind 预测的针对一种 DTI 的结合位点残基，并且 P 表示针对该 DTI 的结合位点中真实结合残基的集合。

此外，我们计算预测的结合位点节点与真实结合位点节点的第一邻域的贾卡德相似系数，记为贾卡德相似系数 $\delta P, P$ 。

邻域 $P = P \cup P_{\text{邻域}}$, $P_{\text{邻域}}$ 表示真实结合口袋的一阶邻域口袋的集合。

图3a、b分别显示了预测结合位点与真实结合位点以及真实结合位点的第一邻域的Jaccard相似系数分布。ZeroBind的平均Jaccard相似系数分别为0.358和0.605。对于邻近位点，大多数位点的Jaccard相似系数都大于0.5。结果表明，尽管预测结合位点与真实结合位点之间存在一定差异，但预测结合位点大多位于真实结合位点周围，表明生成的IB子图在ZeroBind中作为潜在的结合位点具有一定生物学解释性。我们进一步进行了一个实验，随机抽取残基作为潜在的蛋白质口袋，这里用ZeroBind_{random}表示。结果在消融研究中显示，并且我们还计算了随机抽取的结合残基与真实结合位点以及真实结合位点的第一邻域的Jaccard相似系

数。ZeroBind_{random} 分别产生了平均 Jaccard 相似系数为 0.013 和 0.072（所有值均低于 0.2）的结果，这远小于 ZeroBind 的结果。如图 3a、b 所示，我们可以看到随机选择的结合残基与蛋白质中真正的结合位点或其邻位点几乎没有重叠。结果表明，ZeroBind 中的 SIB 模块学习的是潜在的结合位点，而非其他不相关的因素，因为 DTI 结合信息在一定程度上能够引导 IB 子图模块定位潜在的结合位点。

为了进一步证明 IB 子图模块的有效性，我们使用已知的结合位点作为节点分配矩阵 Z 为那些具有已知结合位点的蛋白质训练了一个变体的 ZeroBind 用于归纳测试集。这个变体的 ZeroBind 实现了平均 AUROC 为 0.8278，略高于使用学习到的节点分配矩阵 Z 的 ZeroBind 的 AUROC 0.8032。实验结果进一步验证了 IB 子图模块在 ZeroBind 中的有效性。

在图 3c、d 中，我们进一步将生成的 IB 子图视为丝氨酸/苏氨酸蛋白激酶 N1 蛋白的潜在结合位点。我们可以看到，尽管由于在 IB 子图模块训练中没有已知的结合位点信息，IB 子图引入了一些假阳性结合残基，但 IB 子图中包含的潜在结合残基包含了存放在 BioLip39 中的大部分真实结合残基。

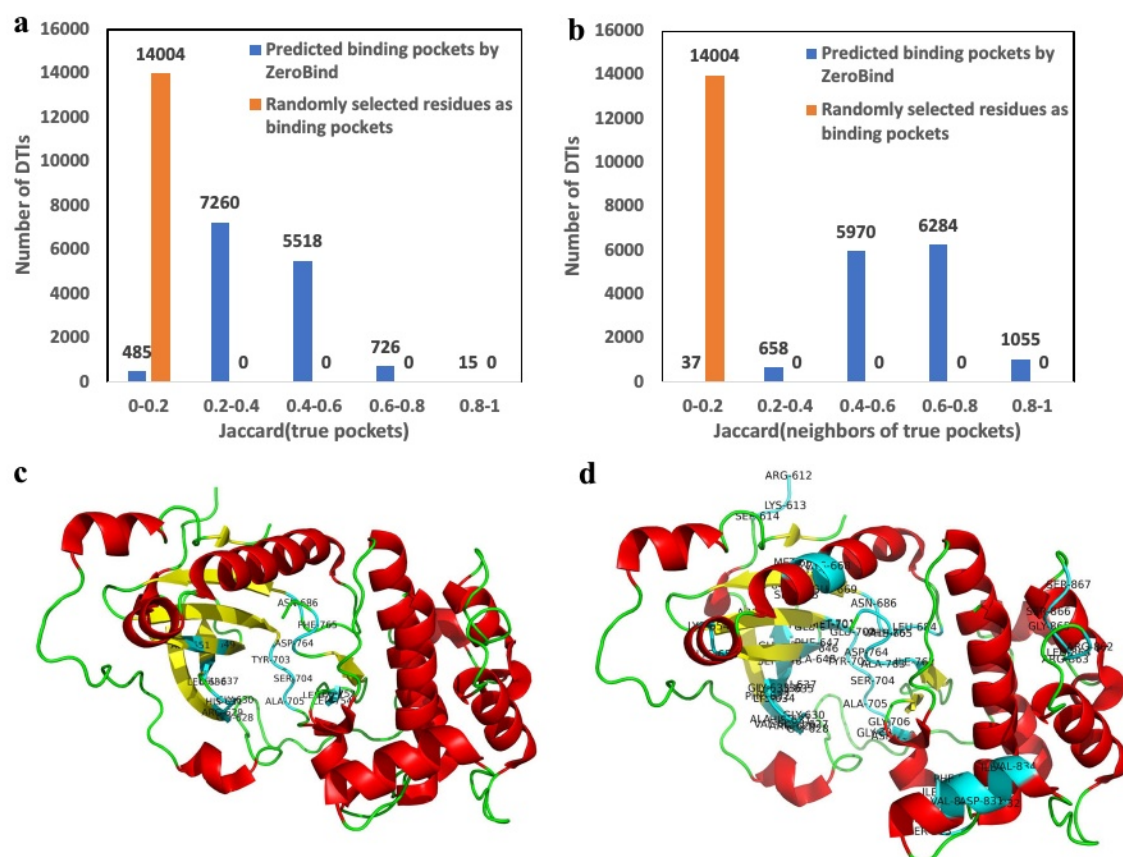


图3 | ZeroBind 能够以弱监督的方式检测蛋白质的结合位点。a 对于各个 DTI 中，预测的结合位点与真实结合位点以及随机选择的结合残基作为真实结合位点的 Jaccard 相似系数的分布。b 对于真实结合位点的第一邻域，预测的结合位点与真实结合位点的第一邻域以及随机选择的结合残基作为真实的第一邻域的结合位点的 Jaccard 相似系数的分布。

结合单个 DTI 的口袋。c 丝氨酸/苏氨酸蛋白激酶 N1 蛋白。通过 BioLip 查询实验验证的 DTI 结合口袋。蓝色部分表示实验验证的结合口袋，连同残基名称和编号。d 通过 ZeroBind 中的 IB 子图预测的潜在结合口袋。红色部分表示蛋白质的螺旋结构，黄色部分表示蛋白质的环结构，绿色部分表示蛋白质的折叠结构。原始数据以“源数据文件”的形式提供。

ZeroBind 能够预测针对新冠病毒 (SARS-CoV-2) 蛋白质的潜在药物。

为了更好地展示 ZeroBind 的有效性, 我们使用自动对接模拟来验证其针对 SARS-CoV-2 蛋白的预测潜在药物。自动对接模拟是一种耗时但可靠的工具, 用于模拟药物-靶点结合过程。为了对公共卫生紧急事件做出快速响应, 我们需要在那些没有太多结合分子的蛋白质上验证 ZeroBind。因此, 我们使用 ZeroBind 来预测训练集中未包含的 10 种 SARS-CoV-2 病毒蛋白的结构与 PubChem⁴⁰ 中 10,000 种药物之间的相互作用, 其中 SARS-CoV-2 蛋白结构的 PDB 编号在补充表 3 中给出。然后, 我们根据 ZeroBind 的预测分数选择排名前十的结合置信度对, 并进一步使用 AutoDock Vina⁴¹ 进行自动对接模拟, 以验证 ZeroBind 预测的有效性。

通过 ZeroBind 预测的前 10 对结合亲和力的平均值为 -7.42 千卡/摩尔 (图 4a), 这些是有前景的药物-靶点对, 也验证了 ZeroBind 预测的可靠性⁴²。如图 9 所示, 对于 SARS-CoV-2 蛋白质和药物的三对结合对, 平均结合亲和力接近 -7.5 千卡/摩尔。此外, 我们在图 4b 中模拟了 ORF8 辅助蛋白与 VZBSCWDKCMOJCR-UHFFFAOYSA-N 药物之间的药物-靶点结合复合物, 其产生了相对较好的结合亲和力评分 -8.4 千卡/摩尔。我们还在图 4c 中模拟了 ORF3a 蛋白与 OLTVRSUIOUTBRQ-UHFFFAOYSA-N 药物之间的药物-靶点结合复合物, 其产生了 -5.8 千卡/摩尔的结合亲和力

评分。如图 4b、c 所示, 我们可以看到预测的药物与 SARS-CoV-2 蛋白质中的口袋结合良好。在未来的工作中, 我们期望 ZeroBind 能验证针对靶蛋白的潜在药物, 尤其是那些没有经过验证的药物的靶蛋白。

关于 ZeroBind 的消融研究

为了展示 ZeroBind 中各个模块的附加值, 我们还进行了一项消融研究来评估其有效性。不同模块的 ZeroBind 的细节如下所示:

- (1) ZeroBind^{MAML}: 在没有元学习策略的情况下直接训练 ZeroBind 的基础模型。
- (2) ZeroBind^{SIB}: 使用蛋白质的所有节点嵌入来识别相互作用, 而不是应用 SIB 模块在蛋白质图上查找 IB 图。
- (3) ZeroBind^{attention}: ZeroBind 没有任务自适应注意力模块来平衡不同任务的重要性。
- (4) ZeroBind^{GN}: ZeroBind 使用 GIN 而非 GCN 作为主干的图神经网络。
- (5) ZeroBind^{random}: ZeroBind 在公式 (4) 中随机设置节点分配矩阵 Z。

结果如表 1 所示。对于 ZeroBind^{MAML}, 我们观察到在半归纳和归纳测试集中有显著的下降, 表明元学习训练策略为未见过的蛋白质和药物提供了强大的泛化能力。我们可以看到, 没有 SIB 时, ZeroBind^{SIB} 的性能较低。

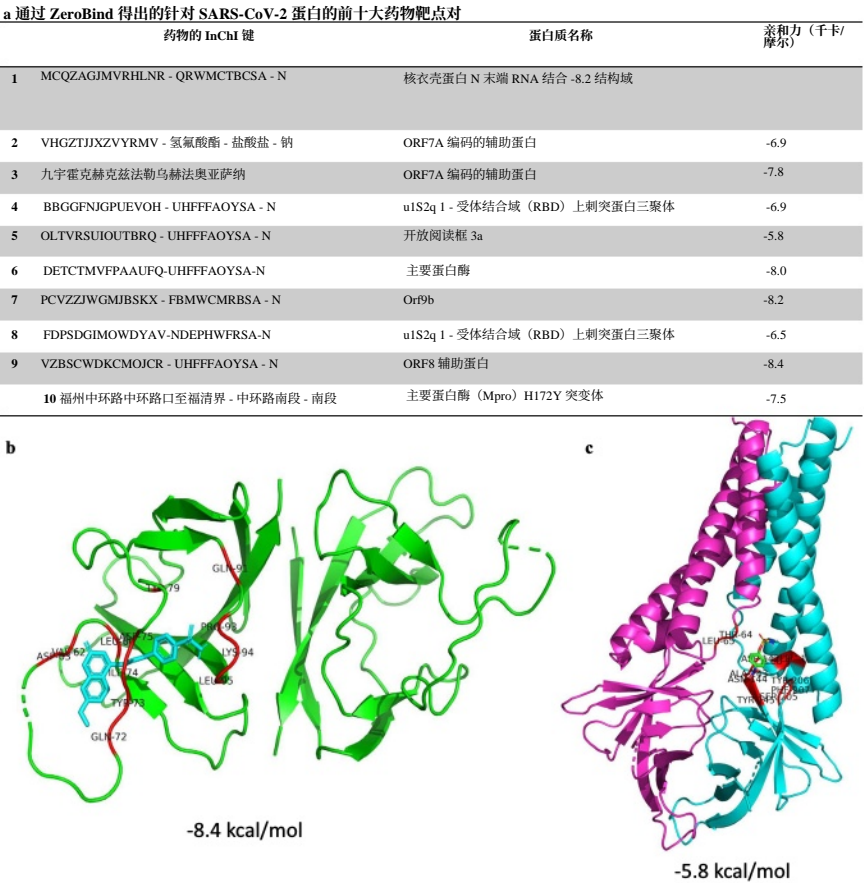


图 4 | ZeroBind 预测与 SARS-CoV-2 蛋白结合的药物。a SARS-CoV-2 蛋白的前 10 大药物-靶点结合对。b 药物 InChI 键 VZBSCWDKCMOJCR-UHFFFAOYSA-N 与 SARS-CoV-2 ORF8 蛋白之间的药物-靶点结合复合物。绿色部分表示蛋白质的主要部分, 蓝色部分表示结合的药物, 红色部分表示潜在的结合位点, 包括残基名称和编号。c 药物-靶点结合

药物 InChI 键 OLTVRSUIOUTBRQ-UHFFFAOYSA-N 与 SARS-CoV-2 ORF3a 蛋白之间的复合物。紫色部分和蓝色部分代表 ORF3a 蛋白的两条主链, 绿色部分代表结合的药物, 红色部分代表潜在的结合位点, 包括残基名称和编号。原始数据以“源数据文件”的形式提供。

用于寻找 IB 子图的模块表明，整个蛋白质的嵌入不如子图嵌入有效。潜在原因是在整个蛋白质图中存在冗余的图信息，进一步表明该分子与结合口袋结合而非整个蛋白质结合。ZeroBind^{attention} 的表现略逊于 ZeroBind，表明有必要使用任务自适应注意力模块来平衡不同蛋白质在不同 DTI 任务中的情况。我们还看到 ZeroBind^{sampling} 的性能下降，原因是基于网络的负采样缺乏由网络生成的非结合注释，而这种注释能够缓解潜在的短视学习。对于 ZeroBind^{GNN}，与 ZeroBind 相比，性能没有显著差异，这表明在 ZeroBind 中，骨干 GNN 对性能的影响小于其他模块。与随机选择的结合口袋相比，ZeroBind^{random} 的表现不如 ZeroBind，进一步验证了 SIB 模块检测潜在结合口袋的有效性。此外，我们可以看到，随机选择口袋的 ZeroBind^{random} 比没有 IB 子图模块的 ZeroBind^{SIB} 表现更差，这表明最初不准确的结合口袋引导模型无法定位真正的结合口袋，从而导致错误的 DTI 预测。

讨论

蛋白质与药物分子的相互作用是一个重要的研究课题，尤其是在面对训练集中未见过的蛋白质和药物分子时。同时考虑蛋白质和分子的信息是一个尚未充分探索的解决此问题的主意。在本研究中，我们将药物 - 靶点相互作用（DTI）预测表述为一个元学习任务，并提出了一个名为 ZeroBind 的元学习框架来解决 DTI 中未见过的蛋白质和药物的泛化问题。具体而言，元学习任务被定义为特定蛋白质的结合药物预测。结果表明，在零样本和少量样本场景中，ZeroBind 优于现有方法。此外，通过 SIB 学习的子图与蛋白质中的结合位点高度吻合，而随机选择的残基作为结合位点几乎与真正的结合位点不重叠。此外，我们在一个真实世界的场景中验证了 ZeroBind 的性能，即它为 SARS-CoV-2 预测药物 - 靶点结合。

由于对图神经网络（GNNs）研究的迅速发展，蛋白质和分子能够以比以往研究中更自然的形式进行编码，而非序列形式。此外，元学习策略还提供了一种更精确的方式来描绘特定于蛋白质的亲和力（DTI）任务空间，这也与实际药

物实验中蛋白质的实验流程一致。然而，ZeroBind 也存在一些局限性，例如元学习训练的困难，其训练过程复杂且容易不稳定。

IB 子图方法为模型提供了可解释的能力，以理解表示学习。在 ZeroBind 中，使用弱监督的 IB 子图方法自动检测潜在的结合位点，该方法不使用结合位点注释作为标签。据我们所知，目前还没有已发表的方法使用 IB 子图或其他基于子图的方法来识别蛋白质中的潜在结合位点，现有的基于子图的方法都集中在药物分子上。在本研究中，由于结合位点数据远远少于 DTI 数据，并且 AlphaFold2 预测的蛋白质结构仍然没有已知的结合位点注释，我们不将真实的结合位点信息纳入模型训练。在基准数据集中的蛋白质中，只有 535 种蛋白质有一部分已知的结合位点。因此，很难使用基于局部结合位点标签的 SIB 模块对所有蛋白质进行训练，以检测结合位点。相反，我们使用 ZeroBind 的 SIB 模块使用全局 DTI 结合标签进行训练，这有可能引导 SIB 模块在蛋白质中定位潜在的结合位点。正如我们的实验所示，具有真实结合口袋的零结合（ZeroBind）表现更好。如果我们能收集到更多的真实结合口袋数据，预计零结合未来的更新会将真实结合口袋信息纳入训练过程，以更准确地适应 DTI 问题。

此外，ZeroBind 中的基础模型是 GCN，并且对于蛋白质 - 分子结合，已经存在更先进的神经架构，例如 EquiBind⁴³ 中使用的 SE(3) 等变 GNN。在未来的工作中，我们期望在 ZeroBind 中研究更先进的 GNN。

方法

数据集生成与扩充

BindingDB¹² 是一个公开的药物 - 靶点相互作用数据库，其中存有药物（类药物分子）与靶蛋白之间的结合亲和力数据。目前，它包含了超过 260 万个实验测定的蛋白质 - 药物复合物的结合亲和力，涉及超过 8000 个蛋白质靶点和超过 110 万个小分子。

为了创建 ZeroBind 的训练和测试数据集，我们应用了几个筛选和预处理步骤来创建一个高质量的基准数据集。首先，对于“目标类型”属性，数据点被筛选为“单一蛋白质”，对于“标准类型”属性，筛选出动力学常数 K_d 、 IC_{50} 和 EC_{50} 。此外，所有目标蛋白质应该是人类或类人类蛋白质，因此使用以下方法对其进行筛选：

表 1 | 在 ZeroBind 上对消融研究的性能评估

计量	模型	传导式测试	半感应测试	归纳测试
曲线下面积（AUC）	零绑定MAML-	0.8632 ± 0.0071	0.7835 ± 0.0063	0.6153 ± 0.0031
	ZeroBind ^{SIB}	0.8556 ± 0.0056	0.8018 ± 0.0121	0.7122 ± 0.0042
	零绑定注意力	0.9057 ± 0.0078	0.8352 ± 0.0020	0.7585 ± 0.0026
	零绑定采样	0.9066 ± 0.0018	0.8280 ± 0.0031	0.7852 ± 0.0043
	ZeroBind ^{GNN}	0.9412 ± 0.0050	0.8652 ± 0.0046	0.8025 ± 0.0027
	零绑定随机	0.8865 ± 0.0017	0.8254 ± 0.0082	0.7562 ± 0.0031
	零绑定	0.9521 ± 0.0038	0.8681 ± 0.0065	0.8139 ± 0.0045
	零绑定	0.9521 ± 0.0038	0.8681 ± 0.0065	0.8139 ± 0.0045
平均精度均值（AUPRC）	零绑定MAML-	0.9540 ± 0.0044	0.9115 ± 0.0010	0.8547 ± 0.0027
	ZeroBind ^{SIB}	0.9385 ± 0.0117	0.9375 ± 0.0028	0.9242 ± 0.0058
	零绑定注意力	0.9789 ± 0.0050	0.9742 ± 0.0065	0.9408 ± 0.0061
	零绑定采样	0.9782 ± 0.0027	0.9645 ± 0.0063	0.9655 ± 0.0057
	ZeroBind ^{GNN}	0.9831 ± 0.0043	0.9850 ± 0.0058	0.9817 ± 0.0062
	零绑定随机	0.9635 ± 0.0056	0.9375 ± 0.0151	0.9288 ± 0.0071
	零绑定	0.9896 ± 0.0013	0.9880 ± 0.0062	0.9872 ± 0.0020
	零绑定	0.9896 ± 0.0013	0.9880 ± 0.0062	0.9872 ± 0.0020

每次实验均执行五次，并报告平均值以及标准偏差。加粗部分表明该方法在比较的方法中是最佳的。

将“智人”作为“目标源生物体”属性。在排除没有 SwissProt 名称的蛋白质以及 ^{RDKit} 无法处理的分子之后, 收集了 150 万对蛋白质 - 药物。我们使用 AI-Bind⁹ 中的阈值, 将动力学常数 K_{on} , K_{off} , IC_{50} 和 $EC_{50} < 1000$ nM 视为正样本, $> 10^6$ nM 视为负样本。

为了证明 ZeroBind 的有效性, 我们通过 cd-hit 和分子骨架分割对蛋白质序列相似性进行聚类, 构建了三个独立的测试集用于模型评估: (1) 直推测试集, 其中具有相同骨架的分子和相同簇的蛋白质在训练集中, 但它们的相互作用在训练集中不存在; (2) 半直推测试集, 其中相同簇的蛋白质在训练集中, 但具有相同骨架的分子不在; (3) 直推测试集, 其中具有相同骨架的分子和相同簇的蛋白质不在训练集中。

我们首先使用 cd-hit, 这是一个广泛使用的蛋白质序列聚类程序, 将 1603 种蛋白质聚类为 1101 个簇, 聚类阈值为 0.4。通过比较分子骨架, 我们将分子按照 0.95 的训练比例分为训练分子集和测试分子集, 并确保这两个集合没有重叠的骨架。由于基于元学习的框架需要足够的数据, 我们首先将具有关联分子数量 < 20 的簇划分为测试簇, 其余簇作为训练簇。然后, 我们使用训练簇中 95% 的蛋白质和训练分子集中的简化分子线性输入规范 (SMILES) 来构建训练集。训练簇中剩余的 5% 的蛋白质和训练分子集中的 SMILES 用作转导测试集。最后, 测试簇和分子集中的蛋白质和 SMILES 进一步用作归纳测试集, 其余数据用作半归纳测试集。在模型训练期间, 应用交叉验证方法进行模型优化。

此外, 我们将半归纳测试集和归纳测试集结合起来构建了另一个少样本测试集, 以评估 ZeroBind 的少样本学习能力。然后, 我们随机选择每种蛋白质的 5 个正向和 5 个负向 DTI 作为少样本微调集, 每种蛋白质的其余 DTI 作为少样本测试集。没有足够正向或负向 DTI 的蛋白质被排除在少样本微调集和少样本测试集之外。训练集和四个测试集的详细信息见表 2。

我们观察到训练集中存在数据不平衡的情况, 大多数蛋白质的阳性比例很高。AI-Bind⁹ 表明, 注释不平衡导致网络学习拓扑捷径, 而非药物 - 靶点相互作用的结合模式。与 AI-Bind⁹ 类似, 我们使用基于网络的训练数据集负采样作为数据扩充来缓解注释不平衡。具体而言, 我们构建了一个二分药物 - 靶点网络。二分网络如图 1a 所示。我们使用迪杰斯特拉算法来找到网络中任意节点对之间的最短路径距离, 并将网络中具有最短路径距离 ≥ 7 的节点对视为非结合对。经过此处理, 注释不平衡略有缓解, 如图 1b、c 所示。

在本研究中, 996 种蛋白质的 3D 结构以 PDB 格式从 RCSB 蛋白质数据库 (<https://www.rcsb.org>) 下载而来, 其余 635 种蛋白质的 3D 结构则通过 AlphaFold 进行预测。此外, 我们从 PDBbind 数据库中下载了包含口袋残基和口袋 3D 坐标位置的真实结合口袋信息, 该数据库包含了 14336 种 DTI 的结合口袋。

蛋白质和药物的图谱构建

在本节中, 我们首先介绍药物图和蛋白质图的构建, 然后给出药物-蛋白质相互作用 (DTI) 任务及其衍生任务、零样本 DTI 预测任务和少样本 DTI 预测任务的形式化定义。

表 2 | 训练集和四个测试集的详情

数据集	蛋白质的数量	药物数量	互动次数
训练集	426	248,997	392,032
传导式测试装置	392	16,946	17,314
半感应测试装置	1575	159,182	228,320
归纳测试集	688	7261	10,165
少样本测试集	1253	127,665	160,948

在使用 ^{RDKit} 从 SMILES²⁷ 字符串构建一个分子之后, 我们应用开放图基准 (OGB) 数据集⁴⁷中使用的编码格式来获得分子表示。具体而言, 原子化学特征包括原子序数、手性、原子形式电荷、所连接的氢原子数量、自由基电子数、杂化类型和芳香性。几何特征包括原子度, 以及用于编码节点表示的环中与否的二进制值。边特征包括键类型、键构型以及键是否共轭的二进制值。

定义 1: 我们将一种药物定义为一个图, 记为 $G_m = V_m, E_m$, 其中 V_m 是表示一个分子的原子的节点集, E_m 是表示哪些节点对相连接的边集。

定义 2: 我们从蛋白质的三维结构中定义一个蛋白质图 $G_p = V_p, E_p$, 其中 V_p 是残基的节点集, E_p 是表示哪些对残基节点相连接的边集。在本研究中, 如果三维结构空间中两个残基之间的欧几里得距离小于 8 埃 (Å)⁴⁸, 则在蛋白质图中这两个残基节点用一条边相连。

对于没有已知三维结构的蛋白质, 我们使用 AlphaFold246⁴⁹ 预测其结构作为补充。蛋白质图的节点特征使用预训练的 ESM-249⁵⁰ (一种通用的蛋白质语言模型) 初始化, 以纳入先验知识。蛋白质图的边特征是节点对之间的欧几里得距离。构建蛋白质图的整个过程如图 5 所示。

在这项工作中, 我们针对每种蛋白质训练一个 DTI 任务。

定义 3: 我们将用于元学习的任务集 T 中的 DTI 任务 T_p 定义为 $T_p = M_p$, 其中 p 是蛋白质集 P 中的蛋白质, M_p 是其相应结合分子和非结合分子的药物集。

定义 4: 我们定义从 DTI 任务继承而来的零样本 DTI 预测任务, 这意味着元模型在元学习测试集中直接用于预测, 而无需使用元学习测试集中蛋白质的样本进行微调。

定义 5: 我们将从 DTI 任务继承而来的少样本 DTI 预测任务定义为, 在对元测试中蛋白质的样本进行微调后, 使用特定于蛋白质的模型进行预测。

少量样本的 DTI 预测时间表和零样本 DTI 预测时间表如图 5c 所示。

在 ZeroBind 中的元学习设置

元学习框架最初是为了在分布式环境中学习多个相关任务中的通用知识, 并利用这些通用经验快速适应额外的任务并提高预测性能而提出的。元学习框架主要有两种类型: (1) 基于梯度的方法: 经典的基于梯度的方法^{MAML}⁵¹ 使用元学习器通过求和多个任务损失并在任务间更新参数来学习良好的初始化。因此, MAML 可以实现高精度和高速度的泛化; (2) 基于度量的方法: 原型网络⁵² 是一种经典的基于度量的元学习算法, 它直接训练每个类别的向量表示 (即原型)。一旦训练出一个好的特征提取器, 新样本的类别就由其向向量空间中最近的原型决定。

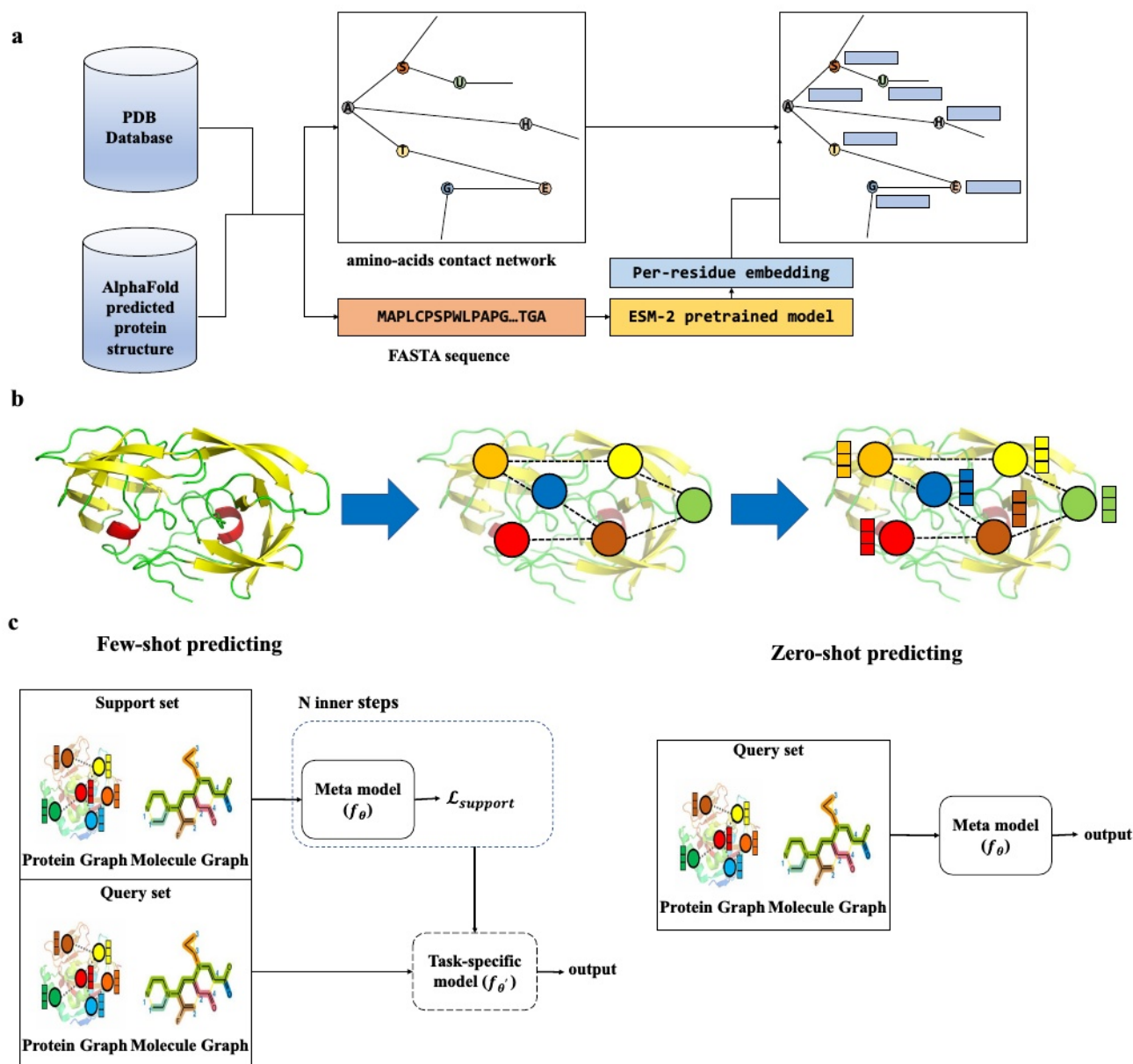


图 5 | ZeroBind 的数据处理。a 从蛋白质三维结构构建蛋白质图谱的过程。我们不是使用肽键作为边，而是用截断距离 < 8 埃 (Å) 连接两个残基，其中边和

节点特征是使用 ESM-2 预训练模型提取的。b 从蛋白质三维结构构建蛋白质图的一个示例。c 少样本 DTI 预测过程和零样本 DTI 预测过程。

在我们的研究中，我们应用基于梯度的方法 MAML 作为训练策略。此外，我们遵循 MAML++³⁷ 进行一些改进，以稳定 MAML 的训练过程。

在 ZeroBind 的细节中，应用了多种技术，包括多步损失优化 (MSL)、学习每层每步的学习率和梯度方向 (LSLR)、子图信息瓶颈 (SIB) 和任务自适应注意力，以提高 ZeroBind 的性能。我们基于 MAML 构建零样本学习框架。首先，我们随机初始化网络参数 θ 。给定从任务分布 T 中采样的训练任务（蛋白质），MAML 旨在学习良好的初始参数，以便能够快速适应额外的任务。对于一对蛋白质和药物，我们采样一个双向、 k 次、 m 查询的训练任务。在这三个超参数中，由于分类任务中只有两种类型的标签（结合和非结合），因此设置双向。在原始的 MAML 框架中，模型首先对每个标签的 k 个数据样本进行几个内部步骤的训练，以学习任务特定的模型，然后在 m 个数据样本上进行测试。在获得每个任务的损失后，MAML 应

用梯度下降，使用所有任务的损失总和或平均值。在这里，我们将训练任务和测试任务分别称为支持集和查询集。然而，MAML 在训练期间存在不稳定性，并且对神经网络架构和大量的超参数敏感。因此，我们遵循 MAML++³⁷ 训练过程，该过程利用了多步损失优化 (MSL) 以及 MAML 中每一层每一步的学习率和梯度方向 (LSLR)。

具体而言，多步损失优化会在每个内部步骤之后计算查询集的损失，并在所有内部步骤优化完成后，使用支持集对基础网络进行优化。更正式地说，我们对模型参数进行如下优化：

$$\theta = \theta - \alpha \nabla_{\theta} \sum_{b=1}^B \sum_{i=0}^N v_i \mathcal{L}_{T_b}(f_{\theta_i}) \quad (1)$$

其中， α 是学习率， \mathcal{L}_{T_b} 表示在每次内步优化后查询集的损失， v_i 表示步骤一的重要性由内部步骤的数量所决定。

盒子 1

零绑定训练过程

Require: $\{\mathbf{G}_m^r, \mathbf{G}_p^r, \mathbf{Y}^r\}$: training data;
 While not done do:
 Sample batch of tasks $\mathbf{T}_\tau \sim \mathbf{p}(\tau)$
 For all \mathbf{T}_τ do
 Sample k examples as support sets of \mathbf{T}_τ $\{\mathbf{G}_{mn}^r, \mathbf{G}_{pn}^r, \mathbf{Y}_n^r\}_{n=1}^k \in (\mathbf{G}_m^r, \mathbf{G}_p^r, \mathbf{Y}^r)$
 Sample m examples as query sets of \mathbf{T}_τ $\{\mathbf{G}_{mn}^r, \mathbf{G}_{pn}^r, \mathbf{Y}_n^r\}_{n=1}^m \in (\mathbf{G}_m^r, \mathbf{G}_p^r, \mathbf{Y}^r)$
 $\theta_{\text{base}} \leftarrow \theta$
 For $i=1$ to inner steps do:
 Calculate the inner step loss weight ω_i
 $\{\mathbf{y}_{in}^r\}_{n=1}^k = \text{BaseModel}(\{\mathbf{G}_{Pn}^r\}_{n=1}^k, \{\mathbf{G}_{Mn}^r\}_{n=1}^k; \theta_{\text{base}})$
 $\mathcal{L}_i^r \leftarrow \text{Eq. (11) with } \{\mathbf{y}_{in}^r\}_{n=1}^k$
 θ_{base} optimize with $\{\mathbf{y}_{in}^r\}_{n=1}^k$
 $\{\mathbf{y}_{in}^r\}_{n=1}^m = \text{BaseModel}(\{\mathbf{G}_{Pn}^r\}_{n=1}^m, \{\mathbf{G}_{Mn}^r\}_{n=1}^m; \theta_{\text{base}})$
 $\mathcal{L}_i^r \leftarrow \text{Eq. (11) with } \{\mathbf{y}_{in}^r\}_{n=1}^m$
 End for
 $\mathcal{L}^r = \{\omega_i\}_{i=1}^N \cdot \{\mathcal{L}_i^r\}_{i=1}^N$
 End for
 $\{\eta_i\}_{i=1}^N \leftarrow \text{Eq. (14)}$
 $\theta \leftarrow \theta - \alpha \nabla_{\theta} \sum_{\mathbf{T}_\tau \sim \mathbf{p}(\tau)} \eta_i \cdot \mathcal{L}^r$
 End while

“盒子 2 号”

BaseModel 前向处理过程

Require: $\{\mathbf{G}_p, \mathbf{G}_m\}$: training data;
 $\mathbf{G}_p = \text{GNN}_p(\mathbf{G}_p; \theta_p)$
 $\mathbf{G}_m = \text{GNN}_m(\mathbf{G}_m; \theta_m)$
 $\mathbf{Z} \leftarrow \text{Eq. (4) with } (\mathbf{G}_p, \mathbf{G}_m)$
 $\mathbf{G}_{p\text{sub}} \leftarrow \text{Eq. (5) with } (\mathbf{Z}, \mathbf{G}_p)$
 For $i=1$ to num steps do
 $\mathcal{L}_{\text{MI-pro}} \leftarrow \text{Eq. (8) with } (\mathbf{G}_p, \mathbf{G}_{p\text{sub}}; \varphi_2)$
 $\mathcal{L}_{\text{MSE}} \leftarrow \text{Eq. (6) with } \mathbf{Z}$
 $\mathcal{L}_{\text{cls}} \leftarrow \text{Eq. (10) with } (\mathbf{G}_m, \mathbf{G}_{p\text{sub}}; \mathbf{Y})$
 $\mathcal{L}_{\text{Base}} \leftarrow \text{Eq. (11) with } (\mathcal{L}_{\text{cls}}, \mathcal{L}_{\text{MSE}}, \mathcal{L}_{\text{MI-pro}})$

LSLR 模块将学习率设置为每一层在每次内部步骤的可学习参数。在不增加太多计算量的情况下，每一层在每一步会自动学习到不同的学习率，这可能有助于缓解过拟合。

总结在 ZeroBind 中的元学习过程（框 1 和框 2），模型首先使用每个任务的支持集更新为特定任务模型，然后计算该任务查询集的损失。重复 N 个内部步骤后，所有查询集的损失通过 $\eta_i \cdot g^N$ 加权平均，并用于优化元模型 $i=1$ 。

通过梯度下降法。经过使用足够数量的样本进行训练后，所学习的模型对于未见过的分子和蛋白质具有良好的预测 DTI 的能力，或者能够迅速适应具有少量训练样本的其他蛋白质任务。ZeroBind 的元学习过程如图 1d 所示。

ZeroBind 中基础模型的架构

图 1e 展示了 ZeroBind 中基础模型的架构，它主要由三个模块组成：一个用于获取分子和蛋白质嵌入的 GNN 模块、一个用于在蛋白质中寻找最具预测性的子图作为结合位点的 SI

B 模块，以及一个将蛋白质子图表示和分子表示连接起来以评估相互作用的密集模块。

基于神经网络的蛋白质和药物表示学习。对于给定的蛋白质-药物对，构建药物和蛋白质图之后，我们首先使用随机初始化的参数将分子原子特征嵌入到向量空间中。我们将分子的节点嵌入表示为 $X_m \in \mathbb{R}^{N_m \times D_m}$ ，也表示为神经网络的初始节点嵌入，记为 Gömlp ，其中 N_m 是节点数量， D_m 是节点嵌入的维度。在这里，我们将基于图卷积网络（GCN）的分子骨架表示为 GCN_m ，用于从分子图学习图嵌入。然后， GCN_m 的第 l 层输出可以表示为：

$$\mathbf{G}_m^{(l)} = \text{RELU}(\hat{\mathbf{A}}_m \mathbf{G}_m^{(l-1)} \mathbf{W}_m^{(l-1)} + \mathbf{b}^{(l-1)}) \quad (2)$$

在 \mathbf{A} 处 \mathbf{G}_m 表示一个图的邻接矩阵， \mathbf{G}_m

\mathbf{G}_m 表示

第 l 层的节点嵌入以及 $\mathbf{W}_{\mathbf{G}_m^{(l-1)}}$ 可学习的权重矩阵。

对于蛋白质图，我们使用预训练的通用蛋白质嵌入 ESM-2 初始化蛋白质图的节点嵌入 $\mathbf{G}_p^{(0)}$ 。与分子图类似，我们使用基于 GCN 的主干 GCN_p 从蛋白质图学习图嵌入。然后， GCN_p 的第 l 层输出可以表示为：

$$\mathbf{G}_p^{(l)} = \text{RELU}(\hat{\mathbf{A}}_p \mathbf{G}_p^{(l-1)} \mathbf{W}_p^{(l-1)} + \mathbf{b}^{(l-1)}) \quad (3)$$

在 ZeroBind 中进行子图学习以识别蛋白质中的潜在结合位点。考虑到一个分子与蛋白质中的结合位点结合，而非整个蛋白质，在对蛋白质主链进行 GCN 处理后，我们应用一个模型不可知的 SIB 模块³³来识别与 DTI 任务相关的最具可解释性且包含最关键信息的子图，其中学习到的子图对应于蛋白质图上的结合位点。SIB 模块的提出是为了在信息瓶颈（IB）原则下识别一个名为 IB 子图的压缩子图。IB 子图还可以消除有噪声和冗余的图信息。在我们的 DTI 任务中，DTI 分数

这在很大程度上由蛋白质口袋的特征所决定，这些特征是蛋白质图谱的子图信息瓶颈，或者是蛋白质中潜在的药物结合位点。

SIB 模块包含一个用于节点分配的子图生成器和一个密集层，以确保生成的子图与 IB 子图相符。子图生成器是一个多层感知机（MLP），它以蛋白质节点嵌入和分子嵌入作为输入，并输出节点分配矩阵 Z ，其公式如下：

$$Z = \text{Softmax}\left(\text{MLP}\left(\text{concatenate}\left(G_p^{(l)}, G_m^{(l)}\right); \phi_1\right)\right) \quad (4)$$

$$G_{\text{sub}} = Z^T G_p^{(l)} [0] \quad (5)$$

其中，多层感知机（MLP）的输出是一个 $n \times 2$ 的矩阵， n 为节点数量。在对多层感知机层的输出应用 Softmax 之后， Z 的每一行都是相应的节点属于 G_{sub} 的概率。

G_p ，记为 $p \times h_1 \times 2 \times G_{\text{sub}} \times h_1$ ， $p \times h_1 \times 2 \times G$ 子图 $j h_1$ ，其中 G_{sub} 表示

发送生成的蛋白质图 G_p 和 G 的子图 子表示

蛋白质图的剩余子图。经过使用足够数量的训练样本进行训练后，所学习的节点分配矩阵 Z 应该趋近于 0 或 1。因此，生成的 IB 子图嵌入可以通过取 Z^T 与节点嵌入的乘积的第一行来获得。

为了稳定

培训 子图生成的过程
将 $p \times h_1 \times 2 \times G_{\text{sub}} \times h_1$ 与 $p \times h_1 \times 2 \times G$ 分离 对于子集 $j h_1$ ，我们应用一种连通损失如下：

$$\mathcal{L}_{\text{MSE}} = \text{MSE}\left(\text{diag}\left(\text{Norm}\left(Z^T A Z\right)\right) - \text{diag}\left(I_2\right)\right) \quad (6)$$

其中 Norm 是行归一化， A 是蛋白质图 GP 的邻接矩阵， I_2 是一个 2×2 的单位矩阵。最小化 \mathcal{L}_{MSE}
能够鼓励 $p \times h_1 \times 2 \times G_{\text{sub}} \times h_1$ ， $p \times h_1 \times 2 \times G$ 子集 $j h_1$ ！ $\frac{1}{2} \times 0, 1 = \frac{1}{2} \times 1$ ，

这会导致一个独特的 Z 来稳定训练过程，并且在子图中形成一个紧凑的拓扑结构。

形式上，子图信息瓶颈通过优化以下目标函数来寻找 IB 子图：

$$\max_{G_{\text{sub}}} I(Y, G_{\text{sub}}) - \beta I(G, G_{\text{sub}}) \quad (7)$$

其中 β 是一个权重， Y represents 是标签 相关的 \boxtimes with G 以及 G 表示原始图。 $I(Y, G_{\text{sub}})$ 和 $I(G, G_{\text{sub}})$ 子 分别
表示 Y, G_{sub} 和 G, G_{sub} 之间的互信息，并且 G_{sub} 表示 IB 子图嵌入。

为了优化目标函数，第一项的下限被定义为具有子图嵌入的真实标签和预测标签之间的分类损失的反数。通过最小化分类损失， $I(Y, G_{\text{sub}})$ 的下限达到最大值。对于第二项，DONSKER-VARADHAN 表示 s_4 的

KL 散度被用于近似 $I(G, G_{\text{sub}})$ 的上限。 $I(G, G_{\text{sub}})$ 的近似可以表述为：

$$\max_{\phi_2} \mathcal{L}_{\text{MI-pro}}(\phi_2, G_{\text{sub}}) = \frac{1}{N} \sum_{i=1}^N \text{Dense}(G_i, G_{\text{sub}}, \phi_2) - \log \frac{1}{N} \sum_{i=1, j \neq i}^N e^{\text{Dense}(G_i, G_{\text{sub}}, \phi_2)} \quad (8)$$

其中， Dense 是由几个多层感知机（MLP）层组成的密集网络，其输入为 G 和 G_{sub} 的嵌入的连接。

为了最小化 $I(G, G_{\text{sub}})$ ，应用了最大优化 的若干内部步骤来最小化 $I(G, G_{\text{sub}})$ 的上限。

在获得生成的蛋白质 IB 子图的嵌入之后，我们将分子嵌入和蛋白质 IB 子图嵌入连接起来，作为输入提供给多层感知机分类层。DTI 分类损失可以表述为：

$$\hat{y} = \text{MLP}(\text{concatenate}(G_m, G_{\text{sub}}); \theta_{\text{cls}}) \quad (9)$$

$$\mathcal{L}_{\text{cls}} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \quad (10)$$

总的损失可以表述为：

$$\min_{\theta_M, \theta_P, \phi_1, \phi_2, \theta_{\text{cls}}} \mathcal{L}_{\text{base}} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{MI-pro}} \quad (11)$$

$$s.t. \phi_2^* = \arg \max_{\phi_2} \mathcal{L}_{\text{MI-pro}}$$

其中， λ_1 和 λ_2 是用于平衡不同损失重要性的权重， θ_M, θ_P 是 GCN_m 和 GCN_p 的参数， ϕ_1 是 IB 子图生成器的参数， ϕ_2 是密集网络的参数， θ_{cls} 是 MLP 层的参数。

在此，我们总结了训练过程， \mathcal{L}_{cls} 通过利用预测值与真实值（DTI 标签而非结合口袋标签）之间的交叉熵来衡量估计分布与原始数据分布之间的差异。 \mathcal{L}_{MSE} 如公式 (6) 所示，鼓励节点分配矩阵 Z 接近 1 或 0。时间

LMI，作为

如公式 (8) 所示，是对公式 (7) 中引入的 $I(G, G_{\text{sub}})$ 的一种近似。为了使公式 (7) 最大化，需求 随着增加

寒理子图信息瓶颈，即 $I(Y, G)$ 的下限，我们首先在密集网络中对 $\mathcal{L}_{\text{MI-pro}}$ 的负值执行梯度下降。目标是使 LMI 最大化。

并达到 $I(G, G_{\text{sub}})$ 的上限，从而满足方程 (11) 的约束。在此之后，我们使用梯度下降法来 \boxtimes optimize 方程 (11)。通过优化 \mathcal{L}_{cls} ，我们增加了

I 的下限 Y, G_{sub}

并且通过优化 $\mathcal{L}_{\text{MI-pro}}$ 来降低 $I(G, G_{\text{sub}})$ 的上限，这种方法旨在使公式 (7) 最大化，并寻找用于检测蛋白质中潜在结合位点的子图信息瓶颈。

在我们的任务中，我们的目标是在弱监督训练过程中找到蛋白质结合口袋，因为蛋白质结合口袋数据的数据数量远远少于结合数据的数据。因此，我们设计了 IB 子图，以在 DTI 结合标签而非蛋白质中的结合口袋标签上进行优化来寻找蛋白质结合口袋。考虑到蛋白质口袋是由蛋白质和分子共同决定的，我们在生成公式 (4) 中的节点分配矩阵 Z 时，将分子嵌入和蛋白质嵌入的连接作为输入。

任务自适应自注意力

在传统的元学习时间表中，会采样批大小的任务。

并且在应用小批量梯度下降时对其进行处理时采用相同的权重。然而，多个任务的相同权重无法反映不同任务对元模型优化的贡献的重要性。我们进一步设计了一个任务自适应自注意力模块（图 1f），以自动学习不同任务的重要性。由于每个任务都是特定蛋白质的 DTI 预测，我们将蛋白质子图嵌入的连接以及查询集中所有分子的嵌入平均值作为任务嵌入。自注意力模块如下：

$$h_{T_b} = \text{concatenate}\left(G_p, \text{Mean}\left(\left\{G_m^i\right\}_{i=1}^m\right)\right) \quad (12)$$

$$\begin{aligned} Q &= W_Q \{h_{T_b}\}_{b=1}^B \\ K &= W_K \{h_{T_b}\}_{b=1}^B \\ V &= W_V \{h_{T_b}\}_{b=1}^B \end{aligned} \quad (13)$$

$$\{h_{T_b}\}_{b=1}^B = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (14)$$

$$\mathcal{L}_{all} = \sum_{b=1}^B \eta_{T_b} \mathcal{L}_{query,b} \quad (16)$$

其中, T_b 是 DTI 任务之一, h_{T_b} 表示任务嵌入, W_Q , W_K 和 W_V 是可学习的参数矩阵, Q , K , V 是生成的注意力矩阵, d_k 是任务嵌入的维度, 用于在训练期间规范 QK^T 的方差并稳定梯度值, η_{T_b} 是生成的权重, 用于平衡不同任务的重要性。

MetaDTA36 使用多头交叉注意力网络来捕获支持配体和查询配体之间的关系, 而非不同的蛋白质之间的关系, 其中它不使用任何蛋白质信息。相比之下, ZeroBind 利用任务自适应注意力模块来学习蛋白质在不同任务中的重要性, 并且它关注不同蛋白质之间的共享结合模式。

实验设置

我们将图卷积网络作为基础图神经网络。在我们的实验中, GCN_P 的 GCN 层数设置为 4, 我们将嵌入维度设置为 1280、512、256、256、256, GCN_M 的 GCN 层数设置为 3, 我们设置的嵌入维度为 256、256、256。我们将内循环中的更新步长设置为 5。应用 $L_{M_{\text{pro}}}$ 的最大优化的 20 个内步。我们将 λ_1 和 λ_2 设置为 0.05。 $L_{M_{\text{pro}}}$ 优化的学习率设置为 0.01。元学习率采用模拟退火算法。我们使用 Pytorch 来实现模型, 并在 GPU 上运行它。

基准方法

我们将 ZeroBind 与多个基准进行比较, 以评估其优势, 并在三个独立测试集和一个少量样本测试集上计算受试者工作特征曲线下面积 (AUROC) 和精确率-召回率曲线下面积 (AUPRC)。

1. **DeepConv-DTI¹¹**: DeepConv-DTI 训练一个卷积神经网络 (CNN) 来学习相邻残基的嵌入, 并训练一个多层感知机 (MLP) 来学习分子指纹, 然后将两者连接起来进行药物靶点相互作用的预测。
2. **GraphDTA¹⁰**: GraphDTA 利用多种类型的图神经网络模型来嵌入分子嵌入, 并利用卷积神经网络模型来嵌入蛋白质残基特征。
3. **Deeppurpose⁸**: Deeppurpose 是一个用于药物-蛋白质相互作用预测的先进深度学习库, 具有多种骨干网络。在此, 我们使用卷积神经网络 (CNN) 作为骨干网络, 用于学习药物和蛋白质的嵌入。
4. **AI-bind⁹**: AI-bind 使用预训练的 mol2vec 和 protvec 模型来初始化学分子嵌入和蛋白质嵌入, 并通过连接分子嵌入和蛋白质嵌入来进行药物-靶点相互作用 (DTI) 预测。
5. **DrugBAN⁵⁵**: DrugBAN 采用具有域自适应功能的深度双线性注意力网络 (BAN) 框架, 以明确地学习药物与靶点之间的局部成对相互作用。

报告摘要

关于研究设计的更多信息可在本文所链接的《自然》系列报告摘要中获取。

数据可用性

在线网络服务器可在 <http://www.csbio.sjtu.edu.cn/bioinf/ZeroBind/> 免费获取。基准数据集是从原始数据库 BindingDB (https://www.bindingdb.org/bind/downloads/BindingDB_All_2D_202311_sdf.zip) 中收集的, 可在 <http://www.csbio.sjtu.edu.cn/bioinf/ZeroBind/datasets.html> 免费获取, 同时还有 SARS-CoV-2 测试数据集, 本研究中使用的实验蛋白质结构数据是从 RCSB PDB 数据库 (<https://www.rcsb.org/downloads/>) 下载的, 而由 AlphaFold 预测的结构是从 AlphaFold 蛋白质结构数据库 (<https://www.alphafold.ebi.ac.uk/>) 下载的。所有 PDB 和 AlphaFold 代码都可以在 GitHub (<https://github.com/myprecioushh/ZeroBind>) 上找到。本文提供了原始数据。

代码可用性

ZeroBind 的源代码可在 GitHub 上获取 (<https://github.com/myprecioushh/ZeroBind>), 同时还有使用文档和设置环境。

参考文献

1. Peska, L., Buza, K. & Koller, J. Drug-target interaction prediction: a Bayesian ranking approach. *Comput. Methods Prog. Biomed.* 152, 15–21 (2017).
2. Bagherian, M. et al. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Brief. Bioinforma.* 22, 247–269 (2021).
3. Abbasi, K., Razzaghi, P., Poso, A., Ghanbari-Ara, S. & Masoudi-Nejad, A. Deep learning in drug target interaction prediction: current and future perspectives. *Curr. Med. Chem.* 28, 2100–2113 (2021).
4. Deng, J., Yang, Z., Ojima, I., Samaras, D. & Wang, F. Artificial intelligence in drug discovery: applications and techniques. *Briefings Bioinformatics* 23, bbab430 (2022).
5. Thafar, M., Raies, A. B., Albaradei, S., Essack, M. & Bajic, V. B. Comparison study of computational prediction tools for drug-target binding affinities. *Front. Chem.* 7, 782 (2019).
6. Cheng, A. C. et al. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* 25, 71–75 (2007).
7. Alonso, H., Bliznyuk, A. A. & Gready, J. E. Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev.* 26, 531–568 (2006).
8. Huang, K. et al. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* 36, 5545–5547 (2020).
9. Chatterjee, A. et al. Improving the generalizability of protein-ligand binding predictions with AI-Bind. *Nat. Commun.* 14, 1989 (2023).
10. Nguyen, T. et al. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics* 37, 1140–1147 (2021).
11. Lee, I., Keum, J. & Nam, H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* 15, e1007129 (2019).
12. Gilson, M. K. et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 44, D1045–D1053 (2016).
13. Davies, M. et al. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* 43, W612–W620 (2015).
14. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind database: Collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J. Med. Chem.* 47, 2977–2980 (2004).

15. Wang, R., Fang, X., Lu, Y., Yang, C.-Y. & Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* 48, 4111–4119 (2005).
16. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082 (2018).
17. Sachdev, K. & Gupta, M. K. A comprehensive review of feature based methods for drug target interaction prediction. *J. Biomed. Inform.* 93, 103159 (2019).
18. Wu, Z., Li, W., Liu, G. & Tang, Y. Network-based methods for prediction of drug-target interactions. *Front. Pharmacol.* 9, 1134 (2018).
19. Wang, H., Zhou, G., Liu, S., Jiang, J.-Y. & Wang, W. Drug-target interaction prediction with graph attention networks. Preprint at <https://arxiv.org/abs/2107.06099> (2021).
20. Öztürk, H., Ozkirimli, E. & Özgür, A. A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction. *BMC Bioinforma.* 17, 1–11 (2016).
21. Perlman, L., Gottlieb, A., Atias, N., Ruppin, E. & Sharan, R. Combining drug and gene similarity measures for drug-target elucidation. *J. Comput. Biol.* 18, 133–145 (2011).
22. Mei, J.-P., Kwok, C.-K., Yang, P., Li, X.-L. & Zheng, J. Drug-target interaction prediction by learning from local information and neighbors. *Bioinformatics* 29, 238–245 (2013).
23. Thafar, M. A. et al. DTiGEMS+: drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *J. Cheminformatics* 12, 1–17 (2020).
24. Cheng, F. et al. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* 8, e1002503 (2012).
25. Luo, Y. et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8, 1–13 (2017).
26. Chen, H. & Zhang, Z. A semi-supervised method for drug-target interaction prediction with consistency in networks. *PLoS ONE* 8, e62975 (2013).
27. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36 (1988).
28. LeCun, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551 (1989).
29. Elman, J. L. Finding structure in time. *Cogn. Sci.* 14, 179–211 (1990).
30. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30, 6000–6010 (2017).
31. Jiménez, J., Skalic, M., Martínez-Rosell, G., & De Fabritiis, G. K deep: protein-ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J. Chem. Inf. Modeling* 58, 287–296 (2018).
32. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* 20, 61–80 (2008).
33. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. 31st Conf. Neural Inf. Process. Syst. 31, 1025–1035 (2017).
34. Nguyen, T. M., Nguyen, T., Le, T. M. & Tran, T. GEFA: early fusion approach in drug-target affinity prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 19, 718–728 (2022).
35. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* 2, 665–673 (2020).
36. Lee, E., Yoo, J., Lee, H. & Hong, S. MetaDTA: meta-learning-based drug-target binding affinity prediction. *ICLR2022 Machine Learning for Drug Discovery* (2022).
37. Antoniou, A., Edwards, H. & Storkey, A. How to train your MAML. *Proc. ICLR* 2019 (2019).
38. Murphy, A. H. The Finley affair: a signal event in the history of forecast verification. *Weather Forecast.* 11, 3–20 (1996).
39. Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* 41, D1096–D1103 (2012).
40. Kim, S. et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 47, D1102–D1109 (2019).
41. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461 (2010).
42. Gurung, A. B., Bhattacharjee, A. & Ali, M. A. Exploring the physico-chemical profile and the binding patterns of selected novel anticancer Himalayan plant derived active compounds with macromolecular targets. *Inform. Med. Unlocked* 5, 1–14 (2016).
43. Stark, H., Ganea, O. E., Pattanaik, L., Barzilay, R. & Jaakkola, T. EquiBind: geometric deep learning for drug binding structure prediction. *Int. Conf. Mach. Learn.* 2022, 20503–20521 (2022).
44. Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. Greg Landrum (2013).
45. Sussman, J. L. et al. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr. Sect. D: Biol. Crystallogr.* 54, 1078–1084 (1998).
46. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).
47. Hu, W. et al. OGB-LSC A large-scale challenge for machine learning on graphs. 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks (2021).
48. Xie, Z. W. & Xu, J. B. Deep graph learning of inter-protein contacts. *Bioinformatics* 38, 947–953 (2022).
49. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123–1130 (2023).
50. Hospedales, T., Antoniou, A., Micaelli, P. & Storkey, A. Meta-learning in neural networks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 5149–5169 (2021).
51. Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *Pr. Mach. Learn. Res.* 70, 1126–1135 (2017).
52. Snell, J., Swersky, K. & Zemel, R. Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* 30 (2017).
53. Yu, J. et al. Recognizing predictive substructures with subgraph information bottleneck. *IEEE Trans. Pattern Anal. Mach. Intell.* <https://doi.org/10.1109/TPAMI.2021.3112205> (2021).
54. Donsker, M. D. & Varadhan, S. S. Asymptotic evaluation of certain Markov process expectations for large time. *I. Commun. Pure Appl. Math.* 28, 1–47 (1975).
55. Bai, P. Z., Miljkovic, F., John, B. & Lu, H. P. Interpretable bilinear attention network with domain adaptation improves drug-target prediction. *Nat. Mach. Intell.* 5, 126–136 (2023).

致谢

这项工作得到了中国国家重点研发计划（编号：2020AAA0107600，X.P.）、国家自然科学基金（编号：6172 5302 和 6207 3219 至 H.B.S.，6190 3248 至 X.P.）以及上海市科学技术委员会（20S11902100 至 X.P.，22S11104100 至 H.B.S.）的支持。

作者贡献

Y.W. 参与了手稿的撰写、数据整理和准备，实现了预测模型并进行了实验，运行了对接模拟，并撰写了手稿。Y.X. 参与了数据准备和手稿撰写。J.Y.、Y.Y. 和 H.B.S. 为模型和实验的设计以及手稿撰写提供了指导。X.P. 构思了这个项目，设计了实验，参与了数据分析并撰写了手稿。

利益冲突

作者声明不存在利益冲突。

附加信息

补充信息 在线版包含补充材料，可在 <https://doi.org/10.1038/s41467-023-43597-1> 获取。

通信和材料请求应寄给潘晓勇。

同行评审信息 《自然·通讯》感谢 Thin Nguyen 以及另一位匿名评审员对本研究同行评审工作的贡献。一份同行评审文件可供查阅。

转载和许可信息可在<http://www.nature>

出版商注：施普林格自然集团对于已出版地图中的管辖权主张以及机构归属保持中立。

[.com/reprints](https://www.nature.com/reprints) 上查阅。开放获取 本文根据知识共享4.0国际许可协议授权，允许在任何媒介或格式中使用、分享、改编、分发和复制，只要您对原始作者和来源给予适当的引用，提供知识共享许可链接，并说明是否进行了更改。本文中的图片或其他第三方材料包含在本文的知识共享许可中，除非在材料引用中另有说明。如果材料未包含在本文的知识共享许可中，且您的预期用途不受法定法规的允许或超出许可用途，您需要直接向版权所有方获得许可。要查看本许可副本，请访问<http://creativecommons.org/licenses/by/4.0/>。

©作者（们）2023 年