这篇是IJCAI2022上讲药物和基因蛋白相互作用关系预测的论文，我觉得可能会和我们现在在做的晶体和气体相互作用有一些共性的地方，所以就拿出来分享一下。

首先，药物和基因蛋白相互作用关系的预测可以加快药物的开发和精准的治疗。现有的很多方法都非常依赖于许多高质量的领域知识或者是基于transductive形式的推理，因此比较难以泛化到知识缺乏的场景或者是未知的基因和药物上。所以作者提出了一种不需要很多的知识并且是inductive的表示学习方法。在contributions里面，他们提到了三点，第一点点是他们提出了一种子图生成的图表示学习方法，并且没有用到额外的领域知识和全局信息，可以泛化到未知的药物和基因上；第二点是通过communicative message passing实现了对于多关系药物和蛋白作用的准确预测；第三点是他们从现有的数据集中构造了两个新的benchmark进行药物和基因相互作用关系的预测，并通过实验证明了他们所提出的算法的有效性和可解释性。

然后接下来是方法部分，首先在他们的图里面有两类节点，一类是药物，另外一类是基因，基因和药物之间存在着多种关系。然后它先这里他建立了一个关系矩阵，药物和基因之间如果存在关系的话，就会在关系矩阵中有一个数字对关系进行表示，然后不同的数字就代表着不同的关系类别。在训练或者预测的时候，会先固定一对药物和基因节点，目标是预测他们之间的相互作用关系属于哪一个类别，首先它们围绕这一对药物和基因节点进行子图提取的操作，使用的random walk with restart的方法，提取h跳内的子图。然后对于这张子图的表示，矩阵N来表示节点的embedding，但是边的embedding和关系的embedding是分开的，边的embedding用的是矩阵E，关系的embedding用的是矩阵R，邻接矩阵除了节点与节点的邻接矩阵之外还有一个节点和边之间的邻接矩阵Ane以及边和关系之间的邻接矩阵Are。

然后为了区分子图中不同的节点，它先给每个节点都打了一个数字的标签，然后这个标签的计算方式就是2i+j，i是这个节点的跳数，j是代表这个节点是药物还是基因，如果是药物j就为0，如果是基因j就为1，最后将这些标签转化为one-hot编码作为节点的初始表示。然后关系的初始表示是关系类型的one-hot编码，因为之前提到关系也有很多类型。边的初始表示和关系的表示是一样的，然后将节点和边的表示投影到相同的维度。

后面就是消息聚合机制，把两个节点和他们之间的边concat起来，然后通过一些函数算出这条边对应的注意力分数，然后根据注意力分数将周围的边聚合起来，再和这层原始的节点表示相加，通过变换得到这一层的节点表示。

对于边的表示，会聚合这条边两边的节点信息，以及这条边所对应的关系信息，然后将聚合得到的边的表示与这一层的边表示以及这条边最初的表示聚合起来作为这一层的边的表示。

在k层之后，就得到了最终节点和边的表示，在进行预测的时候，会将目标的药物和基因的表示concat连接起来，然后通过一个多层感知机算出对应的关系标签，训练的loss就是它与真实标签之间的均方根误差。

在实验方面，我们在原有的两个数据集上构建了新的任务，因为原来的这两个数据集不是专门用来做关系分类的，但它们里面也包含了药物和基因之间的关系，比如第一个DrugBank的数据集里面有激活和弱化两种关系，而DGIdb这个数据集里面有14种关系。为了保证数据集的质量，他去除了重复的节点并且只保留了最少存在五条边的节点。在跑数据的时候，他们也分别设计了transductive的情景以及inductive的情景。在transductive的情景下，每个节点都会出现在训练的图以及测试的图中，但是训练的时候会只选择80%的关系，而剩下的20%的关系会用在测试集里面，并对它们进行预测。在做transductive训练的时候，他选择了一些基于矩阵分解的方法和基于GNN的方法来作为对比，并且也尝试将它们设计的消息传递的机制换成GCN的方法或者Pooling的方法，最终结果都是它们所设计出的模型效果是最好的。

而在inductive情景下，它并不是会用到所有的节点，而只选择80%的药物节点以及它们的关系作为训练集，剩下没有见过的节点会被用到测试集中作为子图中的节点，并且测试集中的这些节点只会保持20%数量的关系，而剩下80%的关系就作为inductive情景下需要预测的关系。然后再inductive的情景下他们的模型也是做到最优的。并且transductive和inductive的结果是比较接近的。

在Ablation study这一块，它们先把消息聚合机制换成其他的GNN方法，这个结果已经在之前的表格里看过了。然后它们也尝试了将子图提取的方法换成另一种enclosing的子图提取方法，并且讨论了跳数和模型准确率之间的联系。在图里可以看到使用random walk with restart的方法会比enclosing的方法更好，而在跳数方面，在3跳的时候效果是最好的。在正文部分，内容是这些，然后在附录里，他还写了关于挖掘药物和基因之间新的关系的实验。他这里做了两个实验，第一个是给定一个药物，叫做Metformin，是一种降血糖的药，然后通过模型预测出10种最有可能与该药物有关的基因联系，在结果里面发现有6种预测的关系在之前的医学文献里可以被找到。而给定一种基因MAPT，通过模型预测与它存在相互作用最强的10种药物，然后有4种关系可以被医学文献所证明。说明了这个模型预测出的结果还是比较可信的。