这是一篇关于化学分子表示预训练的论文，在深度学习中我们可以使用化学分子的向量表示进行分子性能的预测，由于大部分的化学分子都缺少相应的性能标签，所以分子表示的预训练模型常常使用自监督的方法，但是现存的自监督方法在分子表示的预训练中存在一些缺陷，作者认为主要有两点原因，首先第一点是预训练的任务不够好，目前使用自监督方法来进行分子表示预训练可以主要分为两种策略，一种是生成式的，它的想法和NLP中的bert有相似之处，就是在分子图中将一部分的节点或者边或者是一个子图进行遮蔽，然后让预训练模型对这些遮蔽的内容进行还原；第二种是对比式的，它通过节点替换、节点丢弃或者边的扰乱进行图增强，然后让模型学习将这些增强的图数据与原始的分子图向量表示进行匹配。但是作者认为这两种方法在分子图预训练领域不是很实用，因为在自然语言处理中我们隐蔽掉一个词，这句话的语义大概率不会发生变化，在图像处理中使用图像增强操作也大概率不会改变原始图像的内涵，但是在分子图中，我们进行微小的改动就可能会对分子性能产生很大的影响，比如我们把一个分子中的氧原子换成了一个溴原子，或者丢掉一个原子，那么这个物质他原始的分子信息大概率就丢失了，分子性能也会发生很大的改变，因此现存的预训练策略对于化学分子的语义理解是不足的。

第二点是模型它本身的能力不够，由于化学分子他天然地呈现出图的特征，我们可以使用图中的节点表示原子，使用边来表示化学键，因此GNN、GIN等图算法模型在分子表示用的很多，但是这样的模型存在的缺点是它聚合的大多是某个节点附近节点的信息，很难捕获到分子图中的长程范围内的信息。相比之下，transformer模型它可以很好的捕获到全图范围内的信息，但是transformer模型在分子表示中也存在着一些困难，因为transformer处理的数据是序列类型的，它比较难用在图类型的数据上，但也有研究尝试过将图拆成序列然后喂进transformer里，但这样很可能会丢失一些图中的结构信息。

所以基于以上这两点，作者提出了他的思路。首先在预训练策略方面作者做了改进，作者在他的工作里使用的也是生成式的策略，但是不同点在于，他在隐蔽节点的同时，也引入了额外的关于分子的知识，来引导模型往相对正确的方向走；其次在模型方面，他提出了一种叫做线形图的transformer模型，将分子图转化为线性线形图的同时保留了它的结构性息。作者把他的这个方法就做KPGT。

然后，我们就具体看下作者的做法。

首先，他们工作的第一步是，将分子图数据转化为线形的数据并作transformer，作者称之为line graph transformer。首先他们将原本的分子图抽象为一个新的图，具体做法是选取原图中的每一条化学键和它左右的两个原子组成一个新的结点，同时，如果原图中的两个化学键他们之间有个公共原子，那么在新的图中我们也用一条边将他们相连。那么对于这个新图的每个结点的向量表示，作者采用了这样一种方法，它将原来的两个原子的向量做一个相同的线性变换并相加，然后将化学键的向量表示也做一个线性变换，然后将化学键和原子表示在特征维度上连结起来。然后，他介绍了所用到的transformer模型的具体架构，首先将结点向量组成的特征矩阵通过不同的线性变化得到查询矩阵和键值对的矩阵，每个解码块里先采用多头注意力机制，然后进行残差连接的规范化操作得到一个中间矩阵，将这个中间矩阵通过一个线性变换将特征维度扩大了四倍然后通过一个GELU的激活函数，然后又通过一个线性变换将特征维度缩回了原尺寸，再通过残差连接和规范化之后得到这一层的每个结点的向量表示，然后堆叠多层之后得到最终的表示。但是，这样的一个解码器作者认为还不够好，因为它只是单纯地把每个结点从图上拿下来然后直接送到transformer中，没有考虑整个图的结构信息，因此他在这里引入了两种可以表示图结构信息的encoding方法。

首先是path encoding。它的具体做法是，对于图中的任何两个结点，他先找到它们之间的最短路径，然后将这条路径上的每个结点的向量表示进行不同的线性变换再相加并取平均，然后进行另外一个线性变换生成一个标量值，这个值代表了两个结点之间的路径信息，作者将这些标量组成一个与路径有关的注意力矩阵。

然后第二个是distance encoding。它先计算出了两个结点之间的距离，然后通过一个映射又转化为一个标量，那么和前面类似，作者将不同节点所计算出的标量组成一个和距离有关的注意力矩阵。

最后作者将这两个注意力矩阵融入了原来的注意力矩阵得到一个新的注意力矩阵，然后再使用这样的注意力矩阵进行transformer训练，作者认为，这样一个新的注意力矩阵融合了更为丰富的分子图的结构信息，所以可以使得模型的能力更强大。

然后，除了改进transformer之外，作者还提出了一种新的预训练方法。他这里采用的也是生成式的自监督方法，但是并不能采用传统的生成式的自监督方法，因为正如刚开始提到的那样，对分子图的局部信息进行微小的改动就可能会对原来分子的信息产生很大影响。另外分子图中，相邻原子之间的相互关联性是很弱的，如果我们不增加额外的信息的话，预训练的时候唯一的限制就是原子的化合价，比如我们把分子中的一个氧原子mask掉然后让模型去猜，那么所有和氧原子一个化合价的原子都会成为被模型选择的答案，但这些其他原子所组成的分子和原来的分子之间特性是会相差很大的。所以作者在预训练阶段加入了一些额外的知识来丰富分子图的信息，从而提高模型的预测能力。

具体来说，它这里提到了两个和分子有关的知识概念，分别是molecular descriptors和fingerprints，大概意思就是和分子特性有关的分子描述符和分子指纹，他们包含了丰富的分子特性，比如说亲水性，亲脂性等物理化学特性以及分子的结构，并对这些特性进行了量化。而这些分子描述符和分子指纹都是前人通过一些方法设计好的，可以直接拿来用。因此，作者利用每个分子的分子描述符和分子指纹设计了一个单独的结点，他这里用到了200个分子描述符和512个分子指纹作为K结点的初始特征，称之为knowledge node，也就是图中的K结点，然后将它与其他节点相连，并加入到预训练任务中，那么在预训练的时候模型就获得了更为丰富的信息，从而引导对于隐蔽结点的预测，而且除了对结点进行隐蔽并预测之外，他们也对K结点的一部分特征进行隐蔽并回归预测。

那么以上就是作者对整个模型的设计，然后作者在多个分子领域内的数据集上进行了预测。他首先在8个数据集上进行了分类任务，然后在分类任务中也细分成了不进行fine-tune的实验和进行了fine-tune的实验，上面这部分是不进行fine-tune的实验，下面这部分是fine-tune的实验，性能的评价指标使用了AUROC这样一个评价方式，那么这个AUROC的值越接近于1则表示模型的性能越好，对于每个数据集，每个模型都跑了三遍然后算出指标的平均值和方差，括号内的数字就代表方差，可以看见在这8个数据集上KPGT模型的性能都是最优的。进行了fine-tune之后，KPGT的模型性能依旧是最佳的，同时与没有fine-tune的模型相比，总体上性能也有提升，除了第一个BACE的数据集，作者认为这种性能下降可能是fine-tune过程中的灾难性遗忘问题所导致的，他也认为这是未来值得深入探索的方向。

随后，作者还在另外三个模型上进行回归任务的训练，评价指标用的是RMSE也就是均方根误差，RMSE越小的话，模型的性能就越好，可以看出在微调前和微调后KPGT的性能都比baseline更好。

最后，作者进行了消融实验。首先，他把他们的改进的transformer模型换成了其他的主流模型，在回归任务和分类任务上，性能都出现了下降，证明了他们所设计的模型的强大性。然后他们又对他们的预训练策略进行评估，他们先单纯的将分子描述符和分子指纹通过输入多层感知机进行分类和回归预测，又尝试在其他模型里加入分子指纹，最后发现仍然是KPGT模型的性能最优，证明了他们所设计预训练方法的优越性。最后，作者也探讨了masking rate的影响，发现比例在50%的时候性能是最优的，并且这个比例也比之前的其他方法大很多，这也证明了额外知识引入的有效性。