我这周还是找了一篇交叉的论文，这是arxiv上个月挂出来的一篇用diffusion model来做分子设计的论文，是max welling他们团队做的。因为最近师兄师姐都在提diffusion model，我就想学习一下diffusion model的原理，然后搜了一下刚好有一些用diffusion model做分子设计的，并且这一篇它是基于分子的片段去设计分子，和我们现在的工作还算有一些相近的地方。

它主要的一个思想就是基于一些分散的分子片段，然后通过diffusion model在片段之间填充原子，最后生成一个完整的分子。最后生成的这些分子在多样性以及可合成性上有比其他方法更好的表现。首先片段的概念是一些较小的分子化合物，一般来说由少于20个原子组成。虽然不同结构的分子数量非常多，但是片段的数量相对来说会少很多，所以以片段为基础来设计分子会节约很多的时间。

所以在他们的任务中，他们想要达到的最重要的一个目标是通过片段来设计分子，第二点模型的算法需要满足欧氏空间里的不变性，也就是说输入模型的原子或者片段的坐标发生整体平移、旋转或者映射时，最后生成的效果都应该保持不变，这也就是它们题目里以及文章里提到的equivariant这个单词的概念。其次设计出的分子能够与蛋白质口袋较好地匹配。

然后首先文章在这一块简单的介绍了一下diffusion model，一般的diffusion model由两部分构成，一部分是正向的加噪声的过程以及一个逆向的去噪声过程。而加噪声和去噪声的过程都被假设为马尔可夫过程，也就是当前时间步的特征只会取决于上一个时间步的特征，同时也会假设从上一个时间步采样得到当前时间步是符合高斯分布的。从这个公式我们可以看到，在给定前一个时间步的特征xt-1时当前时间步的特征xt的条件概率分布满足一个多元的正态分布，并且这个正态分布的均值和方差都会受到人为设置的超参的影响。在方差不等于0的情况下，每一步的采样都是会引入噪声的，当时间步足够大的时候，x就会趋向于纯噪声。在前向推断的时候，每个时间步的概率分布都是人为设好的，但是在逆向推断的时候从后一个时间步到前一个时间步的概率分布是无法直接得到的，所以这就是diffusion model要去学的东西，就是去学习逆向推断的一个概率分布，一般的做法是把每个时间步的特征输入到diffusion model里，diffusion model会输出每个时间步的均值与方差，然后逆向推断就是从这个学出来的分布里面采样，最后会去最大化真实数据出现的概率。然后也可以认为diffusion model学习的是一个噪声，正向的过程是加噪声，逆向的过程就是去噪声，最后训练优化的时候就是直接去minimize每个时间步两个噪声的MSE。在训练好一个diffusion model之后，我们可以输入给模型一些原始噪声，就可以通过去噪的过程，生成最终我们需要的东西。

那么具体到这个分子生成的任务上的话，首先需要对分子进行特征表示，分子或者片段中的每个原子的特征由两部分构成，一部分是该原子的三维坐标r，另一部分是原子种类的one-hot编码h，将这两个向量拼起来组成一个原子的表示。然后因为坐标的数据是连续的，但是原子种类的数据是离散的，所以在训练以及采样的时候它们会稍微存在一点区别。在加噪声的过程中，坐标向量以及原子特征向量都会平等的加噪声，那么离散的原子特征会变成连续的，原子的特征会一直保持连续的状态直到去噪过程的最后一步，在去噪过程的最后一步取所有原子种类中概率最大的那个当作预测的原子，从而把连续的特征转变为离散的特征。

然后这篇文章里对于模型的训练过程没有详细的介绍，而是直接讲扩散模型训练完成后怎么进行从片段生成分子。

然后首先需要注意的是，他并不是从片段的数据库里任意拿出一些配体进行组装，它的这些片段还是从原来的一个分子里面切割出来的，比如说它会从原来的一个比较大的分子切成三部分，然后保留两端的片段，这些被保留的片段会被记作u，扩散模型要做的是基于这些保留的片段去预测片段中间这个linker的化学组成以及空间结构。

在去预测linker的化学组成以及空间结构之前，首先它们会去预测linker的长度。它们的做法是把这些片段组成一个全连接图然后在上面做GNN。虽然这些片段不是相连的，但是会人为地把它构造成一张全连接图，也就是让u中的任意两个节点都彼此相连，然后会使用类似GCN的算法去学一个linker的长度，但是在这个GCN里面他只会利用到原子种类部分的特征而不会用到坐标的特征。在学习得到linker size之后，就会以预测的linker size生成一定数量的初始原子，然后生成这些原子的初始表示，那么为了保持分子的欧式不变性，它会先根据周围的环境算出一个质心坐标，然后根据这个质心坐标重新标定各个片段的坐标，然后在均值为0，方差为1的正态分布里采集初始的原子特征，这个初始的原子特征可以被视作加了非常多噪声的特征，然后diffusion model做的就是一个去噪的过程，一步步地生成合理地元素种类以及它们的坐标。去噪过程中在刚才讲diffusion model中已经提到过，它实际上是把当前时间步的特征输入到一个神经网络里，然后去学一个分布或者说去学一个噪声，在他们的工作中使用的是一个GNN去学习一个噪声，这里图的构建方法和之前预测linker size的做法是类似的，会先把片段和这些需要被预测的元素人为地构造成一张全连接图，所以这个GNN的输入是当前时间步linker中的原子以及片段中的原子的特征，通过GNN得到的向量的坐标部分会减去原来的坐标，这就是坐标部分的噪声，而元素部分的特征就是把GNN输出的特征当作噪声，显然这样的一个噪声采样过程它是满足欧式不变性的，因为它坐标部分的噪声它计算的是两个坐标之间的相对距离，不会受到平移或者旋转的影响。这样在t个时间步之后，预测出最后的整个分子。

然后到这一步的话，一个用diffusion model来生成分子的流程就差不多做完了。但是我们在生成分子的时候肯定需要去考虑这个分子的实际用途，比如在这个任务里面，它希望生成的分子能够和蛋白质口袋能够比较好地匹配，所以为了考虑蛋白质口袋的信息，他们还做了一个额外的实验，也就是把蛋白质口袋的中各个原子的坐标特征以及元素种类特征在加到diffusion model里面，也就是在原来的环境向量u里面加入蛋白质中的原子特征，然后在这种情况下学到的linker，能够更好地与蛋白质进行匹配，但是这个只能在同时有蛋白质和以及药物分子的数据集中来做。

在实验部分的话，首先可以看一下他用到的一些数据集，第一个是zinc-250K，首先它需要把这些分子切成片段，然后他这里自己提了一种切片的方法把一个分子切成了三部分，它的做法就是在两条没有形成环的单键上将分子切开，然后这样的切割方法产生的片段会有很多，所以它会根据一些指标来过滤掉一些无效地切片方法，最终它的数据集扩大到了四十几万条，验证集和测试集分别是400条。然后第二个数据集也使用同样的方法将分子切割成片段。然后这两个数据集都是将一个分子切割成两个片段和一条linker，在第三个数据集中它将一个分子切割成了多个片段和一条linker，然后他这里使用的切割方法和之前的不太一样，用的是别人工作中的切割方法。然后第四个数据集是含有蛋白质的数据集，它的数据集中包含了药物分子和蛋白质在连结状态下的原子排布情况。

然后可以来看一下他们的实验结果，他这里用到了一些比较特殊的评价指标，QED衡量的是生成的分子是否像一个药物，SA是指这个分子它的可合成性如何，Rings代表linker中环的平均数量，一般linker中环越多代表生成的质量是越高的，valid衡量生成的分子在化合价上是否是合理的，unique和novel评价分子的单一性以及新颖性。他这里的baseline不是很多，就两个baseline，在前三个数据集上，使用扩散模型在大部分评价指标上做到了比较好的效果。然后这张表衡量的是与初始分子的相似性，主要是SC这一列，分别列举了相似度大于0.7、0.8以及0.9的比例，可以看出扩散模型在总体上表现是比较好的。然后最后一个有蛋白质的数据集上的实验，这幅图它主要展示的是预测分子与蛋白质之间的冲突情况，可以如果不加蛋白质的信息的话，它的冲突数量会比较大，而加了蛋白质之后它的冲突数量就少了很多。最后放的是他总体的一个算法流程图，这篇文章主要聚焦的是sampling这一部分，training这一块讲的不是很多。