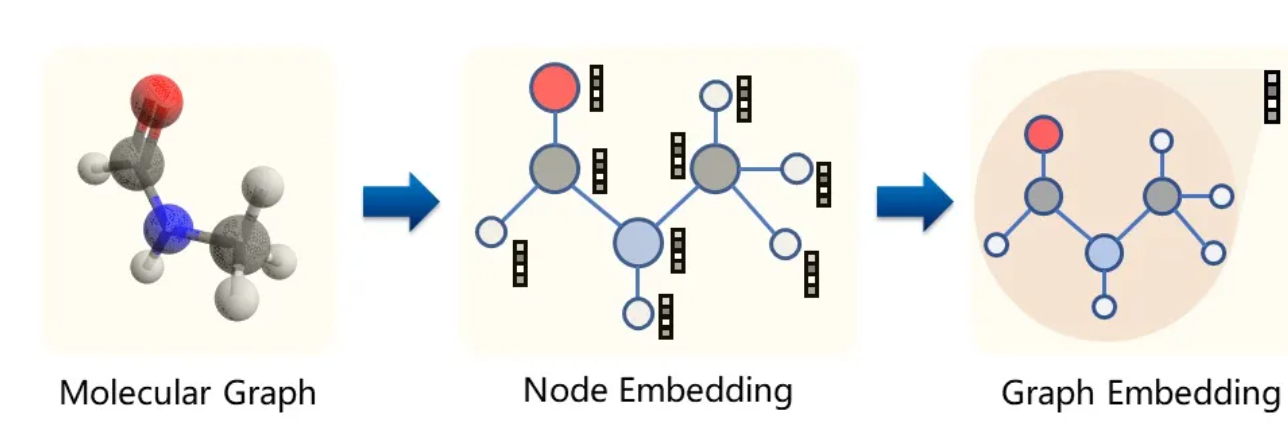
**这周主要解决了周期性的问题(之前写的代码找不到了，所以重写了一遍)，然后说明周期性问题的具体内容，解决方法和处理的进度。**

1. **周期性的具体问题**

**区别于分子的图表示方式，我们所研究的COFs属于晶体，是一种重复性结构，如何表示一个晶体作为模型的输入是我们需要解决的问题。**

这个问题出现在我们利用qm9数据集训练的模型来预测COFs属性时，由于qm9数据集是小分子数据集而COFs是晶体。

作为模型的输入，我们采用了图来表示一个分子。**这种表示分子的方式是基于计算化学中的“局部化学环境”的概念，原子的状态由其周围环境决定。**我们通过消息传递的方式来实现上述方式。如图（这里用一下别人的图）：



我们将原子表示为图中的一个node，并且使用一个由原子本身属性决定初始的多维向量来表示原子的初始状态（初始特征），原子之间是否行成边取决于两个原子之间的距离（截断距离），并且这个距离经过一些转化作为边的初始特征，最终分子被表示为图，后续采用图神经网络。

对于分子，在构造图的时候，原子之间是否成边只需要考虑分子内部的原子；而对于晶体，在考虑原子之间是否城边时不仅需要考虑晶格（可以类比成一个分子）内部的原子，还需要考虑周围的原子。

这里有几个需要解决的子问题：1.晶格外部原子如何表示；2.自动构造边。

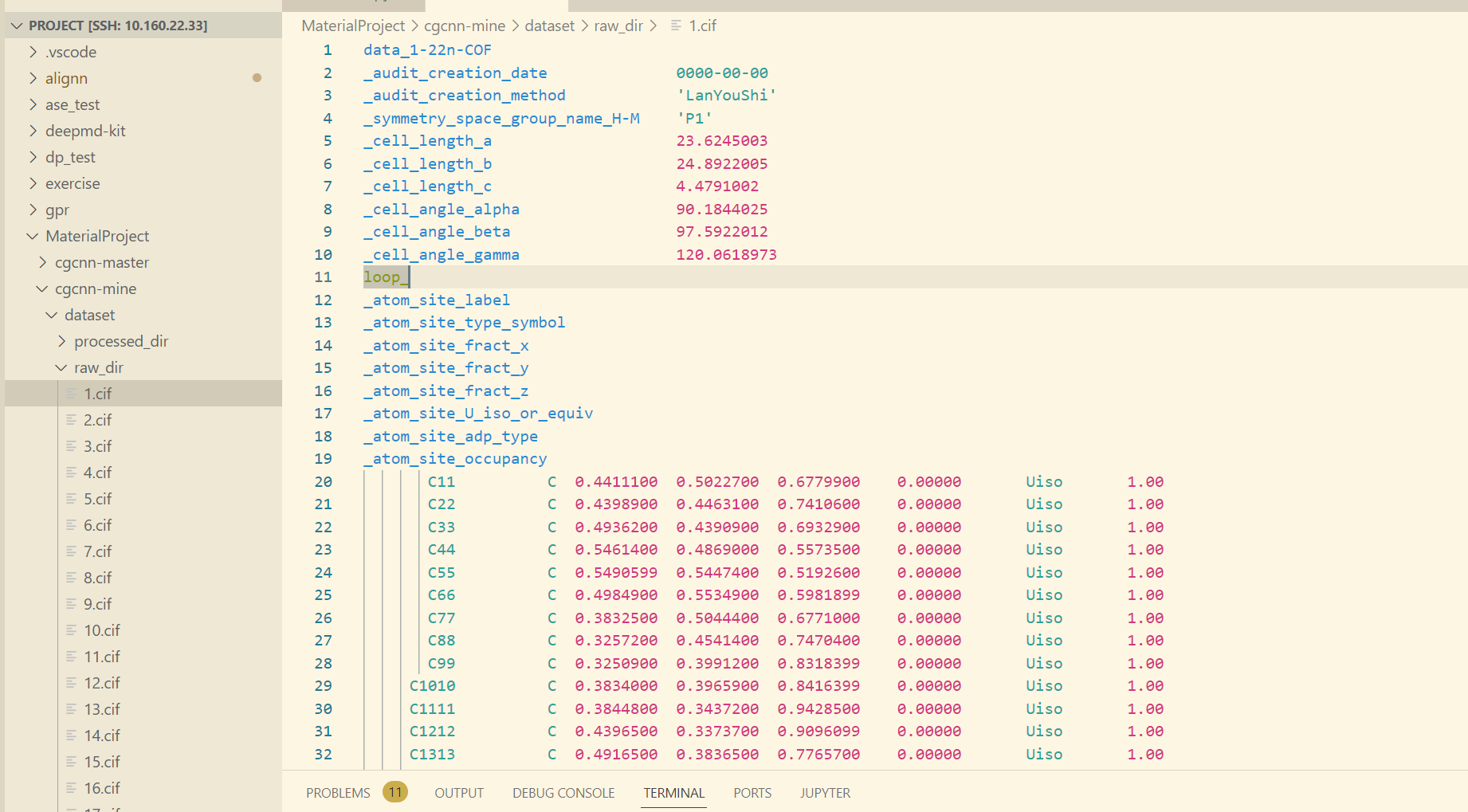
1. **解决方案**

晶体是一种规则的重复结构，不同晶格内相应位置的原子状态相同，所以他们可以由相同的向量表示，问题1解决；pymatgen工具包可以自动获取晶体的邻居，问题2解决。

将晶体表示成一种特殊的图结构

1. **处理进度**

使用了原先找到的COFs数据集，将这些COFs转换为对应的图结构。



如图，将原始表示COFs的cif文件格式转换为图结构。从而可以使用图神经网络去处理。

**体会**

上周组会中老师有提到不知道我最近在做什么，所以我想在这周的周报里说明一下我做的研究：

* 1. **关于周期性的问题**

最早应该是在我大四寒假前夕的某次组会中提出来的，当时我们意识到如果不能用一种合理的方式去表示晶体的话，那么这项工作的质疑性很大。最终这个问题解决是在我本科毕设的最后一章，由于是我临时添加的所以没有做什么实验，在临近毕业答辩的前两周我在一篇论文中看到了别人是如何处理晶体属性预测的，然后将这个方法应用在了我的模型里面。这个问题的解决没有引起大家的响应，是因为周期性的解决只是解决晶体属性预测中的子问题——如何表示一个晶体。但整个晶体属性预测包含如下几个基本部分（还有更多）——晶体的表示，模型结构，模型训练。

**关于晶体的表示**，将晶体表示为一张图是一种比较易于我们这些做机器学习的人理解的方法；还有其他的方法包括使用周期性边界条件（和将晶体表示为一张图本质上是一样的，前一种是在数据上做处理，后一种是在模型中做处理）。

**模型结构**，我们所使用的消息传递框架在分子力场中二体相互作用模型可以找到对应，其本质上是构造一个基于二体相互作用的机器学习模型（其他模型都可以，只是我们一般使用图神经网络）。目前我们已经知道的模型，包括dtnn、schnet、dimenet，是非常优秀的模型。

**模型训练**，这个问题是我们做预测器的关键问题，也是我们目前一直没能解决的问题，这个问题的本质是如何使用小样本训练（高效的利用已有样本）。这个问题在刚开始的时候就已经被提出来了，并且已经有了初始的思路，利用一种主动学习的方法，其中包含两个需要解决的点——1.如何选取候选集进行人工标注；2.人工标注（怎么去做计算，这个是完全的计算化学知识）

1. 第一个问题的解决方法参考了有一篇论文使用最优的样本来进行人工标注，但是目前无法验证。
2. 第二个问题最开始我希望化学系那边能够帮忙，这需要他们提出需要计算的属性以及如何使用计算化学的方法来进行计算。（这个问题我也提出来很多次，一直得不到结果，最后我只能去学习计算化学，到现在还是没有学会，但是这个过程中我深入理解了分子模拟的相关概念）。
   1. **关于物理模型的问题**

嵌入物理模型是最开始老师给我安排的题目，这块的研究我已经了解了很多种嵌入物理知识的方法，但是受限于在解决实际问题时不了解其中有哪些物理知识。对于这个题目的研究我的打算是利用在晶体属性预测（尤其是COFs属性预测），并且已经将知识引入到了模型中，但这些都是很浅的物理知识嵌入。这个问题的后续研究我目前的做法是去学习分子模拟，在学习了更加深入的物理知识后再去思考如何嵌入物理知识。

**最后，我需要给自己说明的是我并没有在不停的更换题目，整个过程我一直是在围绕这COFs属性预测这个题目，包括物理模型、数据库构建和周期性问题。**

**最开始研究物理模型的时候学习了pinn，如何嵌入偏微分方程的方法，后续由于对COFs的物理建模的不了解，没有构建出来偏微分方程，所以这个研究我暂时停下脚步，等后续了解COFs再深入研究。**

**周期性的问题很早就解决了，但是不具有创新性，因为计算化学中已经解决了这个问题，我只是发现了解法，所以我后来没有过多强调这个问题了。**

**数据库的构建，是在模仿国外的“材料基因组计划”，这个是一项偏化学和高性能计算的任务，其目的是在于需要数据来帮我们理解这个问题，最开始的时候我们就讨论没有数据的问题，因为我们甚至不是少数据训练而是无数据训练。所以一起期望化学那边能够多提供一些数据，后来在了解到没有更多现成呢个的数据后只能选择自己算，但是这是一项成本非常高的事情，需要弄清楚我们需要计算哪些属性，以及如何计算。但是那边一直支支吾吾，我也没有办法进行。**

**但是在整个过程我也意识到我对于化学的不了解以及工具使用上的不足，寒假补了点计算化学的知识，尝试着自己去做了一些计算。最近在补习自己的代码能力以及学习一些国外和国内开发的工具，正在学习dpdispatcher、HPC、pymatgen、rdkit、ase等等，工具善其事，必先利其器。**