一种新的分子编码方法辅助机器学习模型进行药物小分子筛选与生成

摘要

我们提出了一种分子编码方法。这一方法将为机器学习和化学这一交叉领域开创全新的工具，以一种非常不同的视角来解决现有分子性质的预测相关问题。并在相应的数据集中取得了很好的效果。机器学习领域存在的一个问题是：我们不仅想知道答案的结果而且还想知道答案的原因，我们使用这种编码方式来研究机器学习模型的可解释性，并且发现这种方法可以找到分子的活性位点。最终我们认为这一编码方式的优势在于以下几个方面：首先，模型具有很高的准确性并在一些任务上领先现有模型。其次，本文提到的编码方式具有普遍的适用性，不仅可以用于卷积神经网络，也可以用于循环神经网络或任何可以读取张量数据的模型。再者，这种编码方法可以获得很好的可解释性，比如本章中使用的类激活映射方法。继而，这种编码方式具有可扩展性，我们不仅可以把元素信息编码进矩阵中，也可以加入成键信息、分子中原子的电荷信息等等。最后，这种编码方式可以用来生成特定属性的小分子。

绪论

如何对分子进行描述是化学家面临的核心命题之一。从球棍模型到价键学说，再到密度泛函理论，它们的提出与发展都是与计算化学所面临的问题息息相关。密度泛函理论的关键在于寻求对体系电子性质的精确描述，一位量子化学工作者将极力避免经验参数的引入而通过对薛定谔方程的严格求解来获取分子性质的答案。毫无疑问，密度泛函理论在化学领域已经取得的巨大成功。但对问题的求解往往需要庞大的计算量，这意味着昂贵的设备与耗时的等待。

机器学习尤其是深度学习是近年来热门领域。应用机器学习算法理解图数据结构则是当前研究的热点。与计算化学所使用的密度泛函方法所追求的目标相反，机器学习是由经验（数据）驱动的：通过对以往数据的不断学习而对新的数据进行预测。在化学领域应用机器学习算法有着天然的优势，我们已经积累了分子与其有关性质的庞大数据库，分子就是天然的图数据结构，除此之外还有药物合成路径，代谢网络等。但与此相应的也伴随着的问题与挑战，例如，分子的结构千变万化，应该如何解决分子由非欧几里得空间的拓扑结构到欧几里得空间的映射问题；如何解决稀疏编码的问题；如何解决庞杂且费时费力的特征工程；是否能在足够大的分子上进行编码又不会使数据大小呈现几何增长。

对于分子而言，每一个原子有其对应的三维空间坐标，且无论分子在空间中如何旋转，分子的性质与结构不变，即分子三维结构的旋转不变性。并且分子基于价键关系其结构千变万化，手动确定特定的描述规则需要大量的人工操作且容易在特定条件下失效。通过图片的方式对分子进行描述则很难反应分子中原子的空间关系且受限于图片尺寸大小。通过点阵的方法则需要考虑编码数据的存储大小（或称为稀疏编码问题）以及分子在空间中的旋转不变性。如何在反应分子结构拓扑关系的同时以较小的存储成本自动化的编码分子，从而输入到机器学习模型中去，成为人们研究的主题。对于分子编码而言，现在的主流方法有分子指纹图谱[68]、SMILES编码[69]、对称函数[70]、权重图[71,72]、库伦矩阵[73]、分子片段集[74]、分子三维坐标[75]、电子密度[76]。

作者在图论的基础上提一种全新的图数据结构，即分层邻接网络。并使用python完成了分子由三维结构数据到分层邻接网络数据算法的实现。然后通过设计神经网络模型，以端到端的方式实现了以有监督的方式预测分子性质，这包括回归与分类相应的问题：例如热力学函数的回归，药物分子活性与否的分类。并且通过迁移学习和多任务联合学习的方法，解决某些小数据集难以训练的问题。我们认为这一图编码方式的主要优势在于，在实现分子结构到数学矩阵的映射的同时，保留了分子结构的层次信息。这使得传统的机器学习方法如卷积神经网络(CNN)，循环神经网络(RNN)，生成对抗网络（GAN）得以直接应用。传统机器学习模型一大问题在于“黑盒”，即我们无法得知模型基于哪些事实或原因做出了判断结果。我们使用算法以热力图的方式找到网络的在特征图上的置信区间以期发现药物分子结构的活性位点。我们相信这些工作将为药物分子设计与高通量分子筛选提供巨大帮助。最后需要强调的是，这一方法不仅针对于分子，也可以应用于所有的稀疏图数据。

方法

分子结构对应的图用邻接列表表示，分子中的每一个原子作为图顶点，原子对应的元素种类作为每个节点的权重，原子与原子之间的键连关系作为边，我们使用广度优先算法遍历该图。其算法计算过程为：

在图上选取一个顶点作为根顶点，并把该根顶点归为一个集合，称该集合为，记作:



其中下标称作层级序数，上标称作顶点序数。

进行第一轮迭代，寻找该根顶点连接的所有顶点，找到的这些顶点称为叶顶点，并将这些叶顶点归为一个集合，称该集合为，记作:



进行第n轮迭代，找到集合所有顶点连接的全部顶点，在找到的这些顶点中，排除层级和对应集合的所有顶点，将剩余顶点归为一个集合（注意，在该集合内部顶点之间可能存在连接关系，故将中顶点排除），称该集合为，记作：



当集合为空时，说明算法已遍历完所有顶点。遍历算法结束。

定义集合为:



集合中子集的个数称为分子的长度n。包含顶点最多的集合，它所包含的顶点个数称为分子的宽度m。

选取图中每一个顶点v作为根顶点,进行上述广度优先算法遍历,得到。选取分子长度最大的集合为实际集合，并分别指定实际集合对应的根顶点，长度，宽度依次为实际根顶点，实际长度，实际宽度。

在实际集合中，根据图的邻接列表，可以得到子集中所有顶点和子集中所有顶点之间的连接关系。因此所有相邻层级之间都可以建立一个二分图(这里我们称之为分子的子邻接网络)。一共可以建立n个二分图。定义分层邻接网络为：



按照分子的实际宽度，对每一个二分图建立其相应的×大小的方阵，方阵定义如下：



其中为方阵第i行第j列的元素，代表第n层级集合中序数为i的顶点。代表第n+1层级集合中序数为j的顶点。可以是一个常数，也可以是一个向量。它包含成键两端原子的相应信息或者它们的成键类型。在这里我们令为：



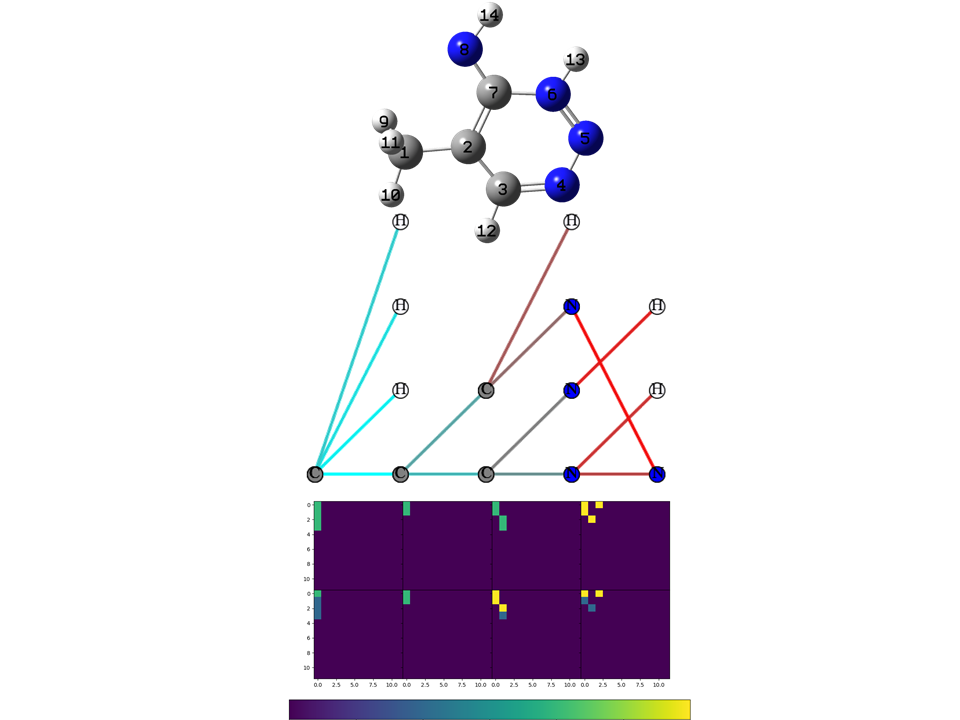
其中，分别为顶点，对应的原子序数或者数值编码。并定义分层邻接矩阵为：



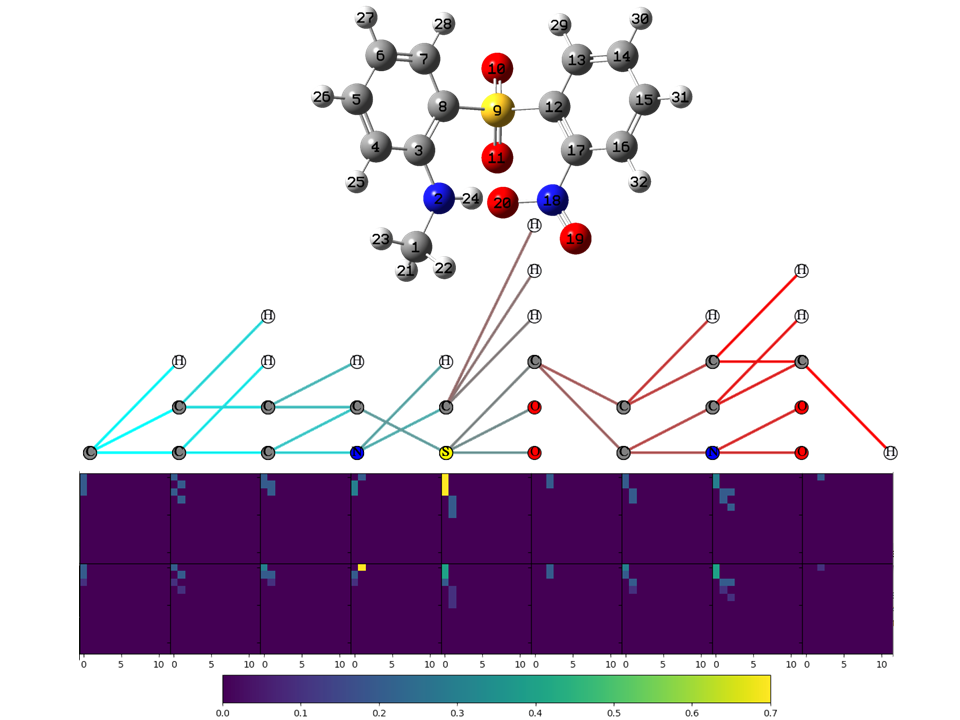
可将视为由构成的分块矩阵(这里我们称之为分子的子分层邻接矩阵)。

最终我们使用python实现了分子图数据编码为矩阵数据的这一过程的转换，并通过python图形库matplotlib 将这一过程可视化表示出来。

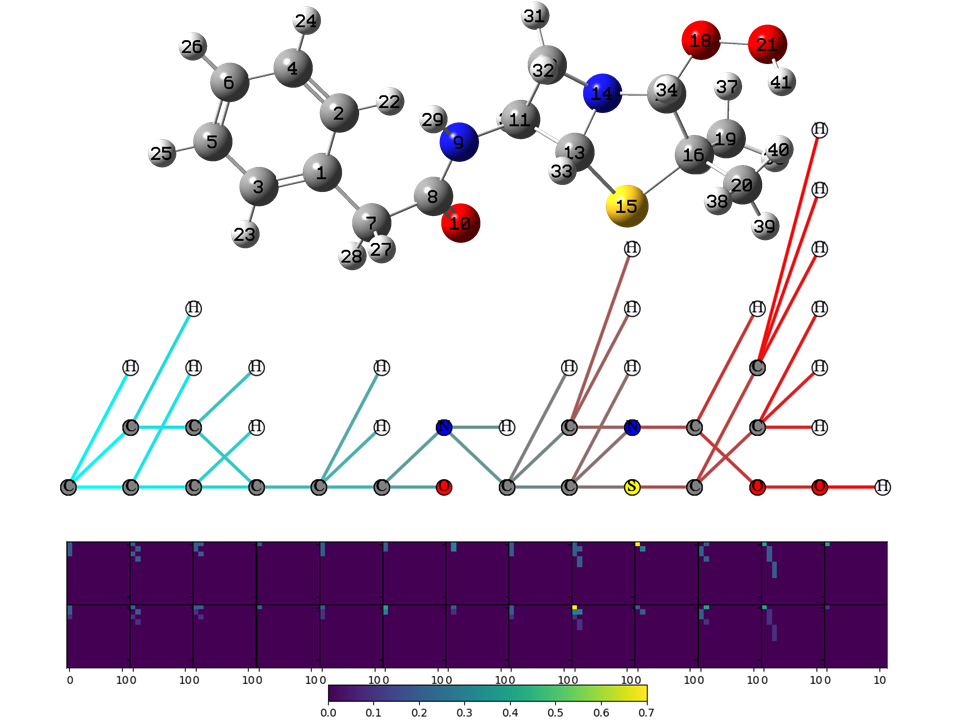
下图为一些分子编码结果示例。图最上方为分子三维结构，中间为分层邻接网络，最下方为分层邻接矩阵。对于分层邻接矩阵，示意图中上下分为两行矩阵，每一行举证由子分层邻接矩阵拼接而成。第一行分层邻接矩阵的每一个像素代表分层邻接网络每一对键连关系的前一个顶点对应的原子种类。相应的，第二行分层邻接矩阵的每一个像素代表分层邻接网络每一对键连关系的后一个顶点对应的原子种类。原子种类按其原子序数对应大小以颜色明暗来表示。图6.1中算法以1号碳原子出发遍历整个分子。划分为分层邻接网络后编码为分层邻接矩阵。原子序数大的分子优先排序，对于相同元素，它所连接的顶点排序越低，则该元素排序越低，以此来尽可能减少分层邻接网络中的连线交错，用上述规则来确定分子编码的唯一性。图6.3中的分子为青霉素。



**图6.1 分子编码示例图**

.

**图6.2 分子编码示例图**

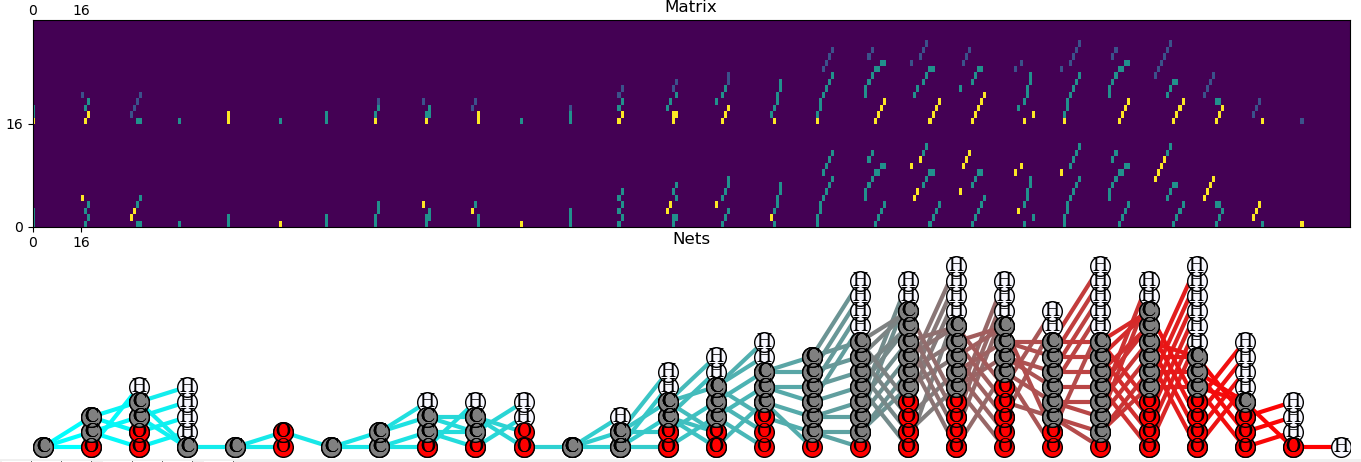


**图6.3 分子编码示例图**

性质分析

子分层邻接矩阵的性质与变换有如下特点：如果多个分子与同一个分子相连，则它们在子分层邻接矩阵中的同一行或者同一列。换言之，如果在子邻接矩阵中的元素位于不同行或者不同列则表示对应原子没有连接关系。而同一行或者同一列出现多个非零元素表明这些顶点对应的原子们与同一原子成键。此外，正向对角线上存在元素时表明在分层邻接网络中相应的连线不相交且不与同一个顶点相连，而反向对角线上存在元素时则表明相应的元素相交(对角线可以为矩形对角线)。

由于上述特性，分子中被划分到同一邻接网络的部分，如果这个部分中的原子或基团与该部分其他原子或基团间没有关系则会在子邻接矩阵中被划分到不同行不同列中。这是分层邻接矩阵相较于邻接网络最大的不同，不同部分可以在子邻接矩阵中的不同区域分布，这助于机器学习模型学习分子表示。如图6.4所示，该分子具有174个原子，尽管分层邻接网络错综复杂，但是在分层邻接矩阵中可以划分出不同原子的对应关系。



**图6.4 分子编码示例图**

因为对于分层邻接网络而言，同一层的顶点交换位置（连接边也跟随改变）图依然等价。所以对于分层邻接矩阵而言，对序数为n的一个子邻接矩阵作行变换中的交换两行，那么对序数为n+1的子邻接矩阵对应列作初等列变换（或者对对序数为n的一个子邻接矩阵作列变换中的交换两列，那么对序数为n-1的子邻接矩阵对应行作初等行变换中的交换两行）后图形仍然等价。因而，对数序为奇数的子分层邻接矩阵进行初等行变换中的交换任意两行，那么对序数为偶数的子分层邻接网络进行初等行变换中的交换任意两列则可保持图等价。

网络中每一层的n个顶点排列顺序的不同可以产生n!种排列方式。排列方式的不同会产生分层邻接网络以及分层邻接矩阵不同种表示方式。我们希望对于同一个分子，其分层邻接矩阵的表示是唯一的，所以我们需要确立每一层顶点的排列原则。

其规则如下，在分层邻接网络的每一层中，每一个顶点的排序按照它们在元素周期表中的顺序从低到高排列。对于相同的元素集合，它们的排列顺序依照它们所连接的上一层的顶点顺序来排列顺序，这样可以在按照原子序数排列的情况下尽可能减少子邻接网络中的边相交。

每一层顶点的排列原则并不是唯一的，也可以不考虑顶点对应的原子种类而完全按照分层邻接网络中的最小相交数目来排列。

按照一定规则可以减少该排列种类的数目，例如对于不同元素，每一层的顶点按照它们对应的原子序数从大到小排列。这样可以减少网络的多样性，从而增加子邻接矩阵网络的复用率。这样做的目的在于有利于机器学习模型发现编码数据中的规律性。选取分子长度最大的根节点进行遍历也是出于这种考虑。而对于相同元素则应相互交换位置，以二分图的边最小相交的方式进行排列。

值得注意的是奇元环图的等价问题，集合 包含的顶点之间存在可能连接关系，这种情况当且仅当图含有奇环时存在。这意味着分层邻接网络结构中将会缺少分子中奇元环上的某一键连关系。这是由于分层邻接网络由若干个二分图构成的，而根据图论：一个图是偶图（二分图）当且仅当它不包含奇环。

为了解决这一问题，我们可以使用占位顶点的方法，使得分子图与邻接网络形成一一映射关系。在图遍历的过程中，如果层l中顶点va vb相连，则在下一层级中添加vh。在这两层所对应的二分图中，建立va vh vb vh新的连接关系.

在分层邻接矩阵中顶点权重用其对应的原子序数来表示，而占位符顶点权重可以用0.5数值表示。

此外，我们需要与邻接网络的差异，正如我们在邻接网络中所叙述的，邻接网络以分子中所有原子为横，纵坐标，然后寻找横纵坐标对应原子的连接关系，如果相连则该矩阵对应元素为1，否则为零。这样虽然分子整个结构的信息储存在了矩阵中，但矩阵本身却无法反应分子任何结构层次信息。其次，邻接矩阵为对角线上元素为零的对称矩阵，而分层邻接矩阵不是对称矩阵，对角线上的元素也不一定为0。

为什么要选取分子长度m最大的L为实际L选取分子长度m最大的L为实际L可以增加分子的层次感，可以增大子分层邻接网络的复用率。确定了分子的根顶点实际上也就确定了分子的取向。通常而言，对于分子的编码，应当保持分子结构的旋转不变性。当编码的所有分子采取这种根顶点选取方式时，实际上相当于确立了所有分子的取向。这是十分重要的。

通常而言，选取的根顶点使分子的长度最大时对应分子的宽度最小。在一个分子结构数据库当中，选取其中宽度最大的分子，其宽度为所有分子的统一宽度。同样的，在模型训练的分子结构数据库中应选取其中长度最大的分子，其长度为所有分子编码为分层邻接矩阵的统一长度，不足的部分用0补齐。此外值得注意的一点是，如果机器学习模型最后部分采用的是全局最大池化，则输入数据大小尺寸可以不固定，也就是说可以对可变长度的分子进行处理。

结果与讨论

选取数据集中单分子作为训练和预测对象

BACE

**表7.3 BACE预测结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Task | Accuracy | TP | TN | Balanced Accuracy | ROC |
| BACE | 0.81419 | 0.79231 | 0.82635 | 0.80933 | 0.88017 |

BBBP

**表7.3 BBBP预测结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Task | Accuracy | TP | TN | Balanced Accuracy | ROC |
| BBBP | 0.85602 | 0.92881 | 0.60227 | 0.76554 | 0.87820 |

HIV

**表7.3 HIV预测结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Task | Accuracy | TP | TN | Balanced Accuracy | ROC |
| HIV | 0.83852 | 0.53623 | 0.84719 | 0.69171 | 0.74904 |

TOX21

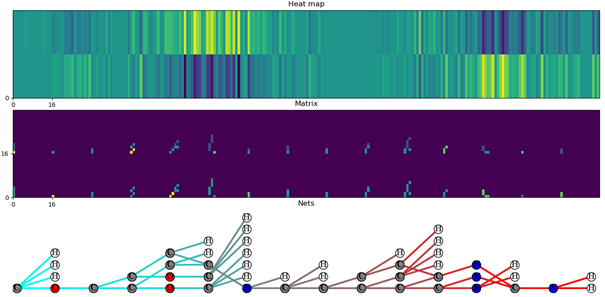
**表7.4 TOX21中12种毒理性预测结果**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Task | NR-AR | NR-AhR | NR-AR-LBD | NR-ER | NR-ER-LBD | NR-aromatase |
| AUC-ROC | 0.8324 | 0.9005 | 0.8535 | 0.8110 | 0.7314 | 0.7683 |
| Balanced Accuracy | 0.7892 | 0.8219 | 0.7601 | 0.7215 | 0.6852 | 0.6958 |
|  |  |  |  |  |  |  |
| Task | NR-PPAR-gamma | SR-AER | SR-ATAD5 | SR-HSE | SR-MMP | SR-p53 |
| AUC-ROC | 0.7293 | 0.7382 | 0.7760 | 0.7001 | 0.8491 | 0.7872 |
| Balanced Accuracy | 0.6408 | 0.6727 | 0.7054 | 0.6658 | 0.7713 | 0.7296 |

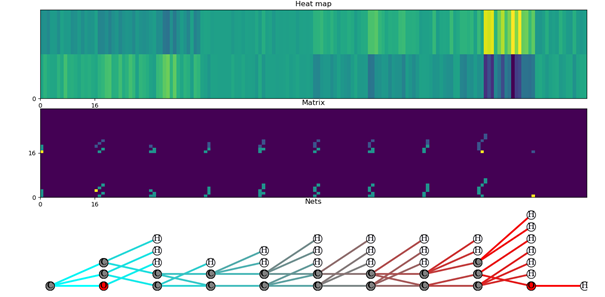
ClinTox

**表7.3 ClinTOX预测结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Task | Accuracy | TP | TN | Balanced Accuracy | ROC |
| ClinTox | 0.73897 | 0.74903 | 0.5000 | 0.62452 | 0.69636 |



我们使用类激活映射（Class Activation Mapping, CAM）[84–86]的方法来生成分子活性位点热力图,我们可以非常直观的看到机器学习模型的决策依据。如图5.4图中分为三个部分，依次为分层邻接网络，分层邻接矩阵，以及分子BBBP活性热力图。热力图分为上下两个部分，下方为模型对分子可以穿越脑血屏障的位点判断，上方为模型对分子无法穿越脑血屏障的位点判断，颜色越亮代表对这一判断认可程度越高，反之越低。模型认为该分子可以穿越脑血屏障与结果一致。从热力图中可以发现模型之所以认为该分子具有BBBP活性主要是出于末尾的氨基区域所做出的结论。然而甲基部分不利于该分子具有BBBP活性。



如图5.5分为三个部分，依次为分子NR-AhR活性热力图，分层邻接矩阵，以及分层邻接网络。热力图分为上下两个部分，下方为模型对分子无活性部分的判断，上方为模型对分子有活性部分的判断，颜色越亮代表对这一判断认可程度越高，反之越低。模型认为该分子具有NR-AhR活性预测结果与实际结果一致。可以发现模型之所以认为该分子具有NR-AhR活性主要是出于末尾的羟基所做出的结论。

值得注意的是，我们在这里使用的特征可视化方法可以获得分子活性特征的一维信息，适合于药物小分子的活性位点判断。我们下一步的工作是对这一方法进行扩展，使得这一方法这一找到具体到哪一些原子起到作用。

如果我们将分子编码后的每一个分层邻接矩阵视为分子的一个基本单元的话，我们可以用一个编号来代表该基本单元。分子即可变为一串序列。我们可以用语言模型来生成分子。

构成分子的基本单元在分子数据库中被分子大量复用。我们可以使用机器学习的方法来学习这些基本单元组合成分子的统计学规律，进而生成分子

总结

我们提出了一种新的编码方法并设计神经网络对大量分子结构数据进行有监督的学习，然后对分子的热力学性质和药物活性进行预测最后得到了很好的预测效果。这表明我们设计的编码方式行之有效。这一编码方式具有别的方法所不具备的优势，即我们可以寻求模型的可解释性以发现药物分子的活性位点。

一种新的分子编码方法辅助机器学习模型进行药物小分子筛选与生成

在此研究中我们提出一种分子编码方法：序列二分图网络。以这种分子编码方法我们设计了卷积神经网络来进行药物分子筛选，并且使用循环神经网络来生成新的分子。我们发现基于这种编码方法，机器学习模型可以有效的筛选药物小分子，并且可以产生合理的分子结构。这种端到端的方法避免了特征工程，可以为机器学习领域带来新的工具

在机器学习领域将分子的图数据所在的非欧式空间映射到欧式空间。尽管图论中经典的关联矩阵、邻接矩阵等数据类型可以完成图结构信息到矩阵的映射，但是它只反应了分子中原子的键连关系，分子在空间中基团与基团的结构层次信息却被打乱了，本文设计的机器学习编码方式在保留分子中原子的键连关系的同时，分子中存在的层次关联信息也被反应了出来。

**A new molecular coding method assists machine learning model to screen and generate small drug molecules**

**Abstract**

In this study, we propose a new molecular coding method: sequence bipartite graph network of molecular structure. Using this molecular coding method, we designed a convolutional neural network to screen drug molecules, and used a recurrent neural network to generate new molecules. This molecular coding method allows convolutional neural networks and recurrent neural networks to effectively screen drugs and automatically design molecules. This end-to-end approach avoids feature engineering and can bring new tools to the field of machine learning.

Keywords：Molecular coding; Sequence bipartite graph network; Machine learning; Small molecule screening; Small molecule generation

**Introduction**

机器学习的流行

如何对分子进行描述

前人的工作，基于smiles作为文本生成任务，但往往会生成有无化学意义的分子。

自己的贡献

我们的目标是使机器学习模型能够依据期望的性质来自动化设计分子，

分子反向设计，分子图生成

Our goal is to enable machine learning models to automatically design molecules based on the desired properties.

本文以一种新的方式看待分子结构

基于图的方式可以生成更合理的分子。

最佳的根顶点使得分子最长，最窄。

每一层的排序使得二分图最小交错（不同连接部分在矩阵中不同行不同列）。

使用占位符表示奇数环

性质

二分图相交与矩阵对角线

矩阵的初等变换

QM9

ZINC

生成分子

生成目标分子

一种分子编码方法在药物筛选与分子设计中的应用

看待分子图结构方式的改变

从二分图的角度描述分子的基本单元，

分子被拆分并重新表述为序列

保留分子键连关系的同时，保持分子结构与结构的层次关系

这种编码方式，，分子相当于具有三个色彩通道的图片

**Methods**

机器学习模型中的运算是对一系列张量数据进行的数学操作。将分子所对应的图数据映射为矩阵是机器学习模型学习及预测分子性质的重要步骤。尽管图论中经典的关联矩阵、邻接矩阵可以完成分子的图数据结构信息到矩阵的映射，但是它只反应了分子中原子的键连关系，分子在空间中结构与结构的层次却被打乱了，本文设计的机器学习编码方式在保留分子中原子连接关系的同时，分子中存在的层次关联信息也被反应了出来。

本节先从一些最基本的分子结构开始。如图所示，我们可以将一些小分子用加权二分图表示。然后将加权二分图依照一种规则转化为矩阵。



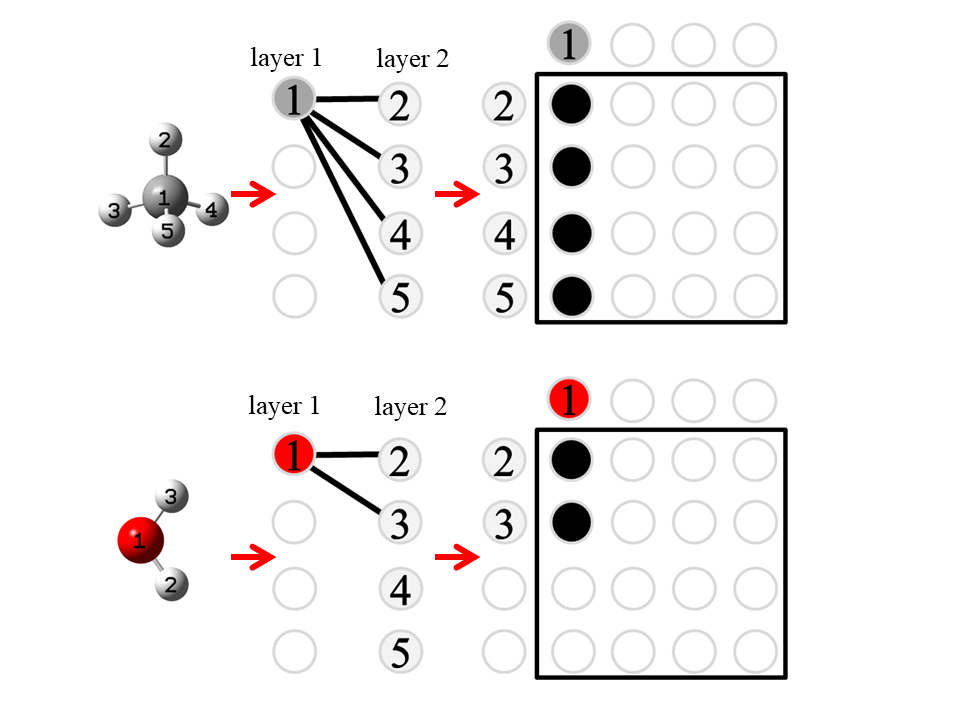
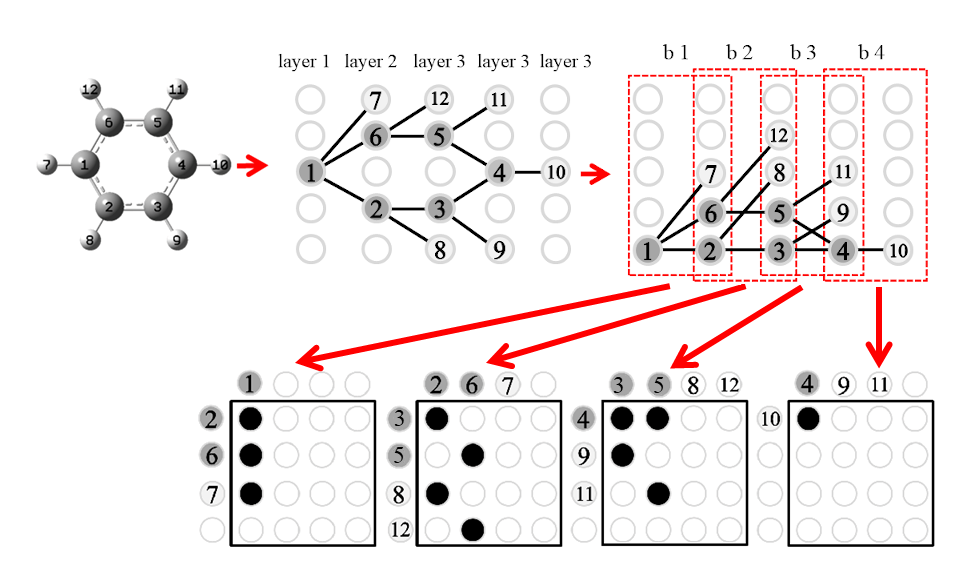


图1

对于这种简单的原子，我们从一种非氢原子出发，将该出发点定义为根顶点，寻找所有的连接原子，将根顶点作为集合1，将所有被连接原子作为集合2，绘制二分图。但是对于苯环分子，由于被连接原子还连接了其它原子，我们需要序列二分图来表示这种连接。继而按照上述方式转化为矩阵。（注意，这里由图到矩阵没有采用邻接矩阵的方法）



图二

为了能够在算法上实现这一过程，我们定义算法如下：

分子结构对应的图用邻接列表表示，分子中的每一个原子作为图顶点，原子对应的元素种类作为每个顶点的权重，原子之间的连键作为边，我们使用广度优先算法遍历该图。其算法计算过程如下：

在图上选取一个顶点作为根顶点，并把该根顶点归为一个集合，称该集合为，记作:



其中下标称作层级序数，上标称作顶点序数。

进行第一轮迭代，寻找该根顶点连接的所有顶点，找到的这些顶点称为叶顶点，并将这些叶顶点归为一个集合，称该集合为，记作:



进行第n轮迭代，找到集合所有顶点连接的全部顶点，在找到的这些顶点中，排除层级和对应集合的所有顶点，将剩余顶点归为一个集合（注意，在该集合内部顶点之间可能存在连接关系，故将中顶点排除），称该集合为，记作：



当集合为空时，说明算法已遍历完所有顶点。遍历算法结束。

定义集合为:



集合中子集的个数称为分子的长度n。包含顶点最多的集合，它所包含的顶点个数称为分子的宽度m。

选取图中每一个顶点v作为根顶点,进行上述广度优先算法遍历,得到。选取分子长度最大的集合为实际集合，并分别指定实际集合对应的根顶点，长度，宽度依次为实际根顶点，实际长度，实际宽度。

在实际集合中，根据图的邻接列表，可以得到子集中所有顶点和子集中所有顶点之间的连接关系。因此所有相邻层级之间都可以建立一个二分图(这里我们称之为分子的子邻接网络)。一共可以建立n个二分图。定义分层邻接网络为：



按照分子的实际宽度，对每一个二分图建立其相应的×大小的方阵，方阵定义如下：



其中为方阵第i行第j列的元素，代表第n层级集合中序数为i的顶点。代表第n+1层级集合中序数为j的顶点。可以是一个常数，也可以是一个向量。它包含成键两端原子的相应信息或者它们的成键类型。在这里我们令为：



其中，分别为顶点，对应的原子序数或者数值编码。定义分层邻接矩阵为：



可将视为由构成的分块矩阵(这里我们称之为分子的子分层邻接矩阵)。

|  |  |
| --- | --- |
| **Algorithm 1** | |
| 1：  2：  3：  4：  5：  6：  7：  8：  9：  10：  11：  12： | **input:** graph and  **output:**  **while do**  **if then**  **end if**  **return** | |

输入图

输出集合

初始点作为第一层

广度优先搜索遍历图上顶点

找到上一层所有顶点的连接顶点，建立集合C

查看顶点是否属于集合C

如果不是的话新建一层，并把该顶点放入该层集合

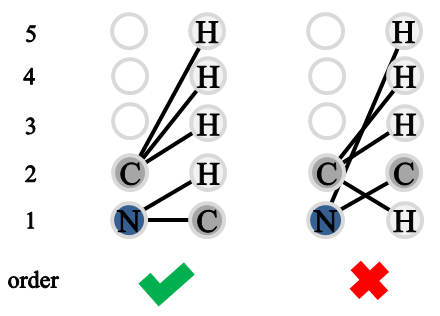
如果是的话把该点放入当前层

广度优先搜索结束终止遍历

讨论：

**每一层原子的排序**

我们需要以相同的方式确定每一层顶点的排列方式。排序方式为：含有较大原子序数的顶点排在前位，反之排在后位。而对于含有相同原子序数的顶点，这些顶点连接的上一层顶点排的顺序作为它们的排列依据。上一层连接的顶点排的顺序越向前，它在该层的排列越向前。这样可以减少交错。



图三

**二分图与对应矩阵的性质**

通过观察可以发现：一个顶点与多个顶点相连，不为零的元素位于矩阵同一行或者同一列；不共顶点的连线不相交，矩阵中的非零元素位于正向对角线上；不共顶点的连线相交，矩阵中的非零元素位于反向对角线上。

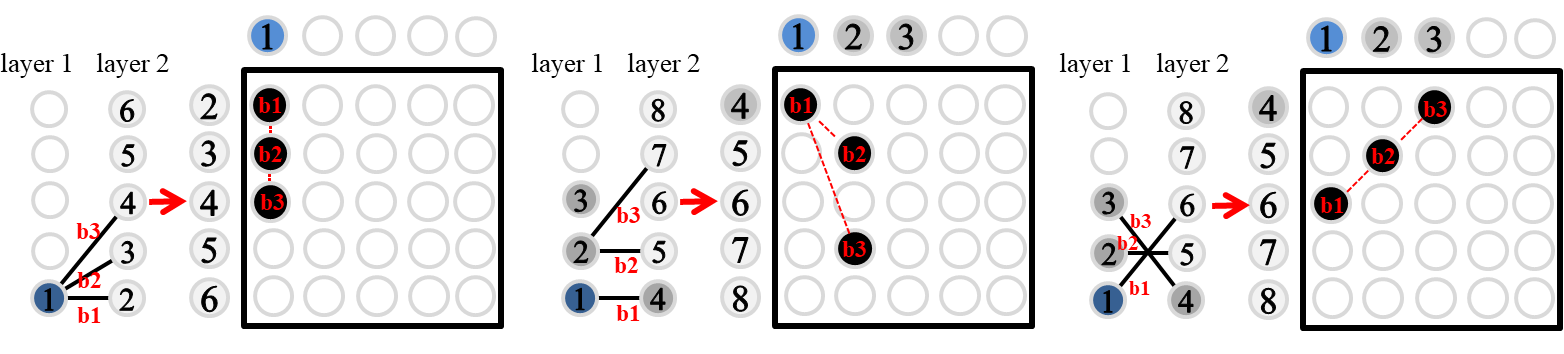


图4

被划分到同一二分图中的分子片段，如果没有连接关系将被划分到不同行不同列当中。

设计每一层顶点的排列方式，如果不同顶点之间能够尽可能避免连线的交错，我们可以在矩阵中区分分子划分在该区域的不同部分。

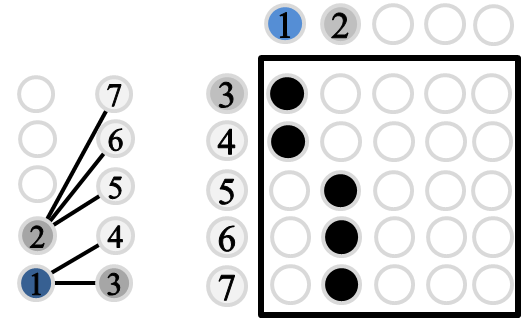


图5

**图等价**

1. **矩阵初等行变换的意义**

序列二分图某一层顶点交换顺序，相当于对应矩阵做初等变换。

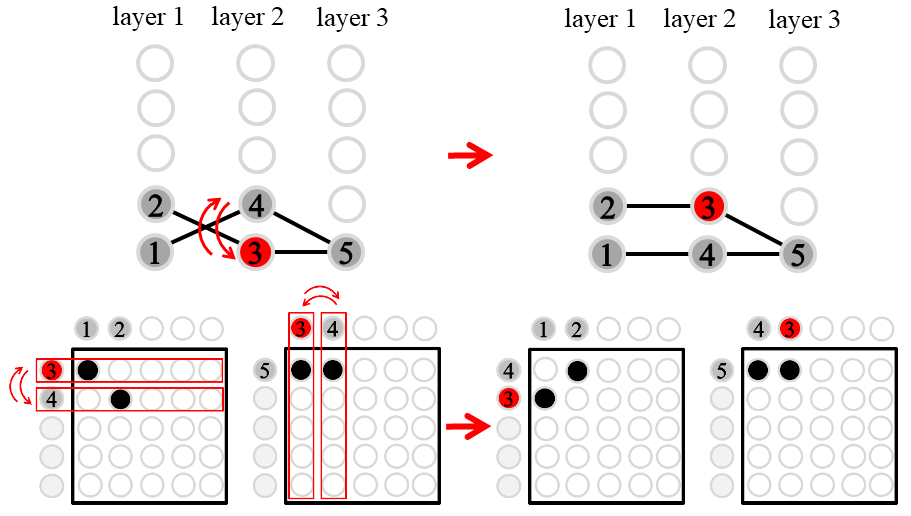


图6

1. **图论问题**

一个图是偶图（二分图）当且仅当它不包含奇环。对于连续二分图结构而言，如果把所有的键表示出来，也需要该分子不包含奇圈。然而在分子结构中奇圈结构普遍存在，例如3元环与五元环。为了解决这一问题，我们可以使用占位符的方法。

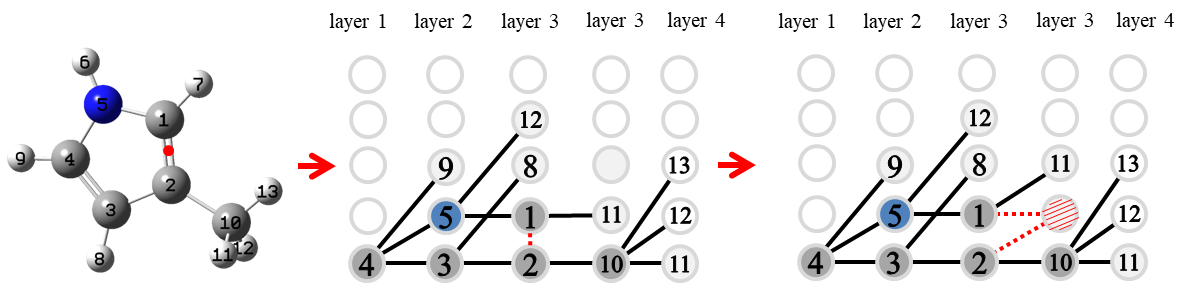


图7

当某一层顶点内部两个顶点存在连接关系，我们在下一层建立占位顶点与它们相连。

分层邻接网络与矩阵的性质以及图等价策略

**小分子分解为序列二分图**

**二分图序列**

**转换为矩阵序列**

**图等价**

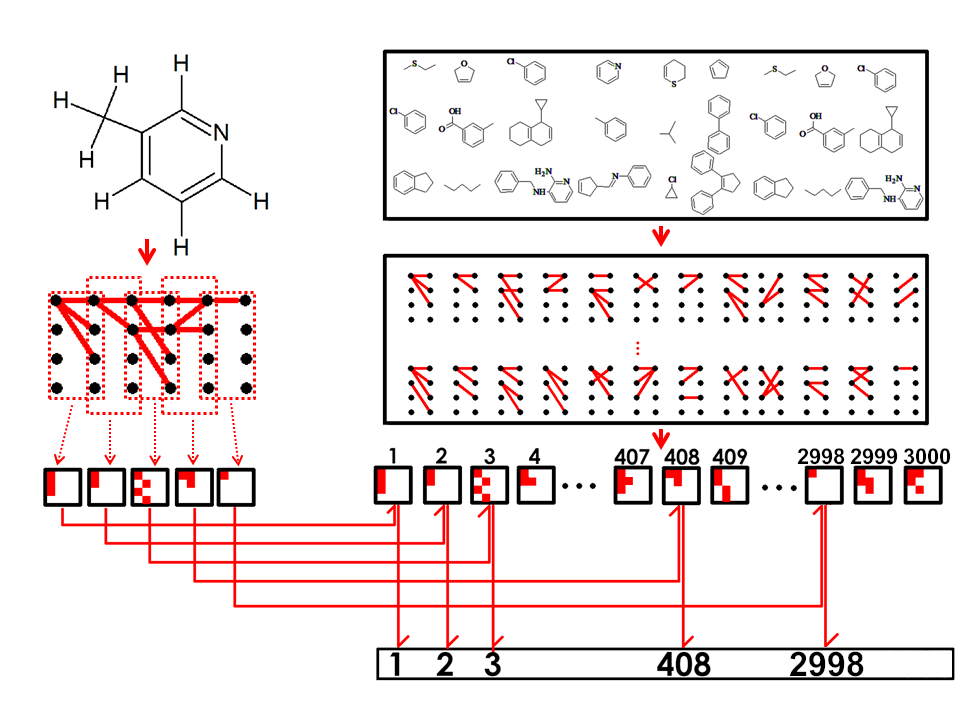
**每层原子的排序**

**二分图对应矩阵的性质**

**序列二分图对应序列矩阵的潜在性质**

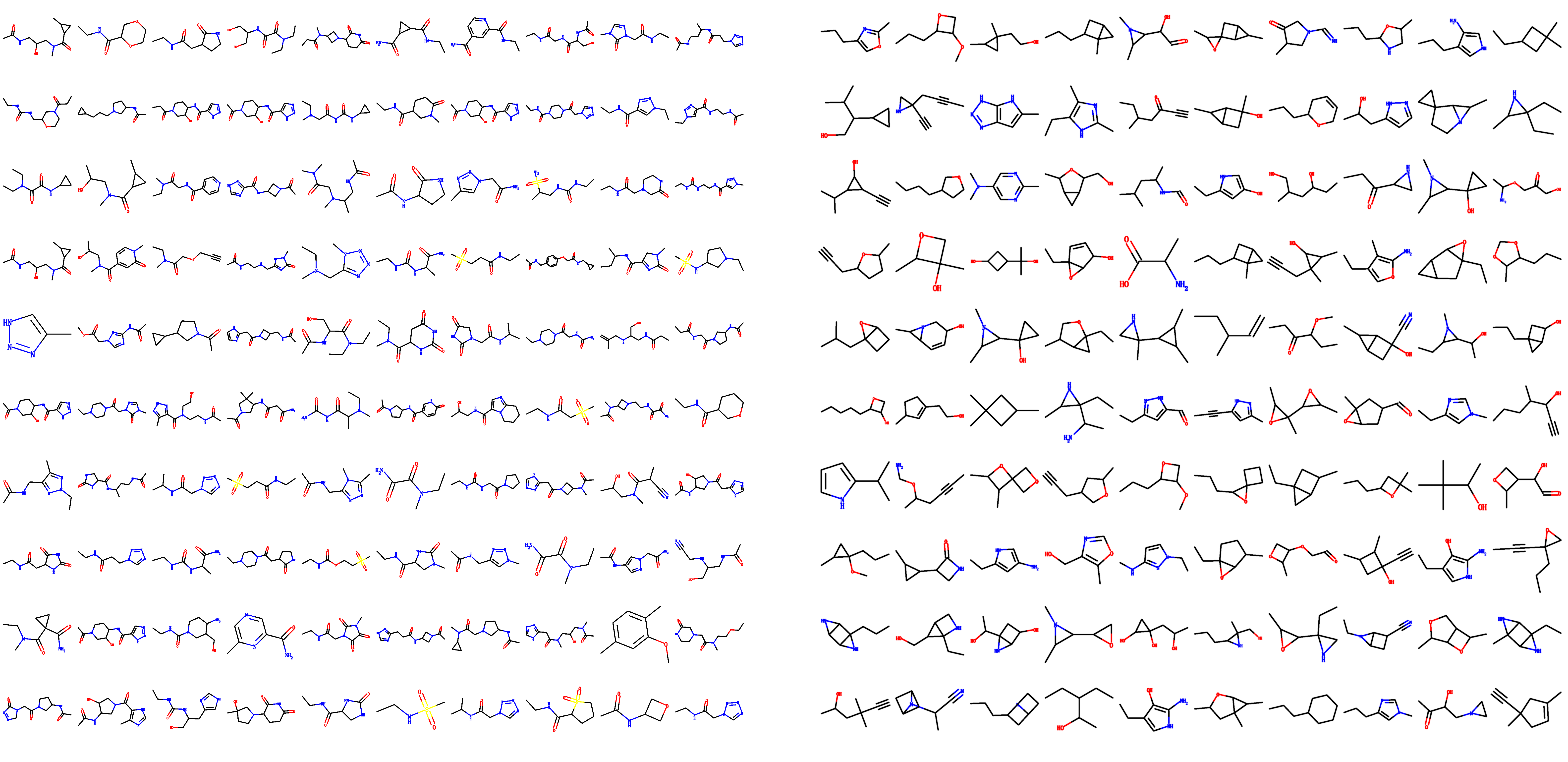
**多层信息的填入**

**卷积时物理含义明确**



从ZINC数据库中下载10万个分子，将它们转换为矩阵序列。一共具有1283382个矩阵元素，去除重复之后有8329种矩阵。换言之，这10万个分子由这8329种矩阵排列组合而成。

这些矩阵。



生成的随机分子，

调控RNN模型生成指定分子

隐藏层h0输入分子的性质

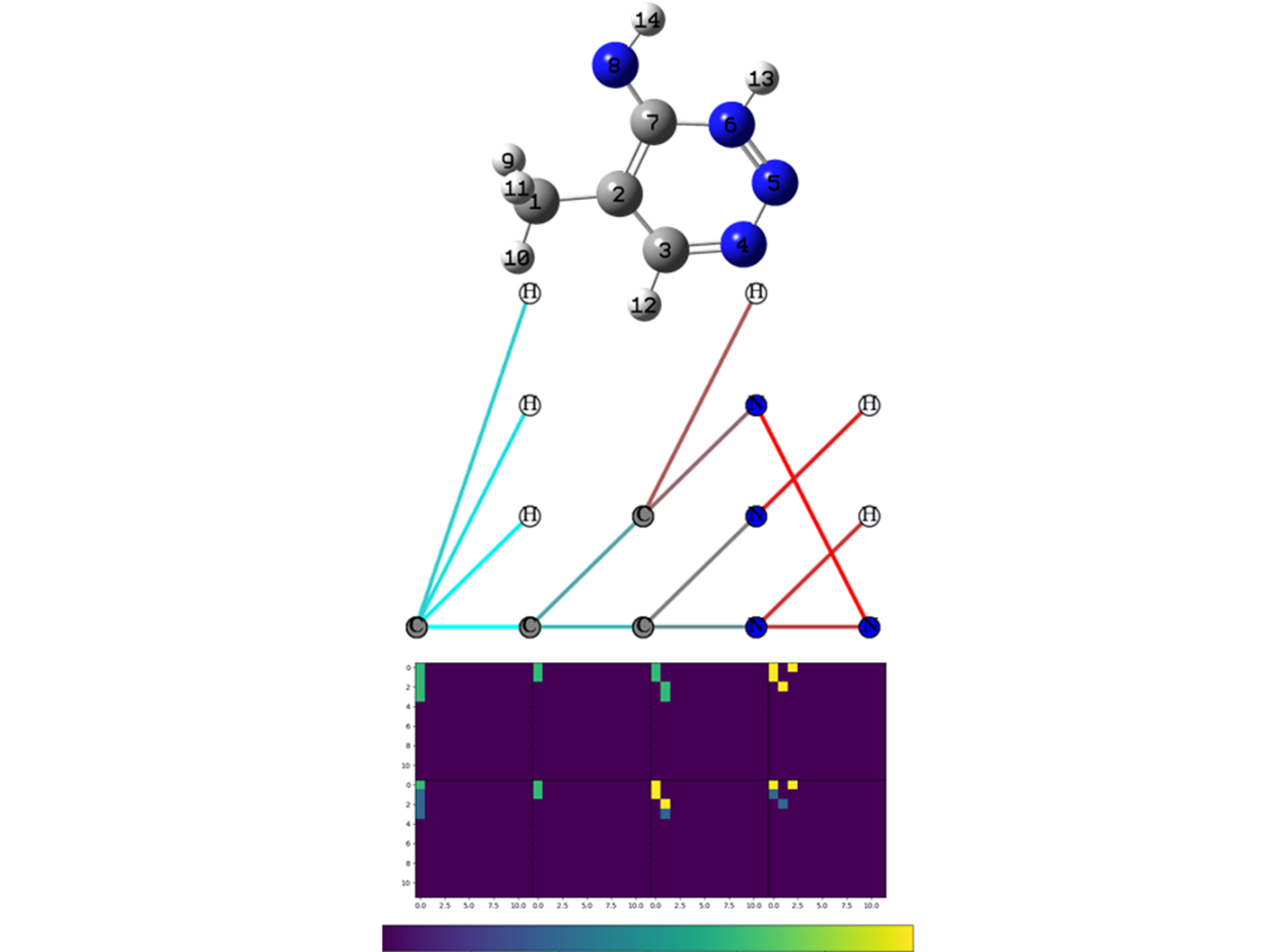
对既有结构的改造以满足某种性质

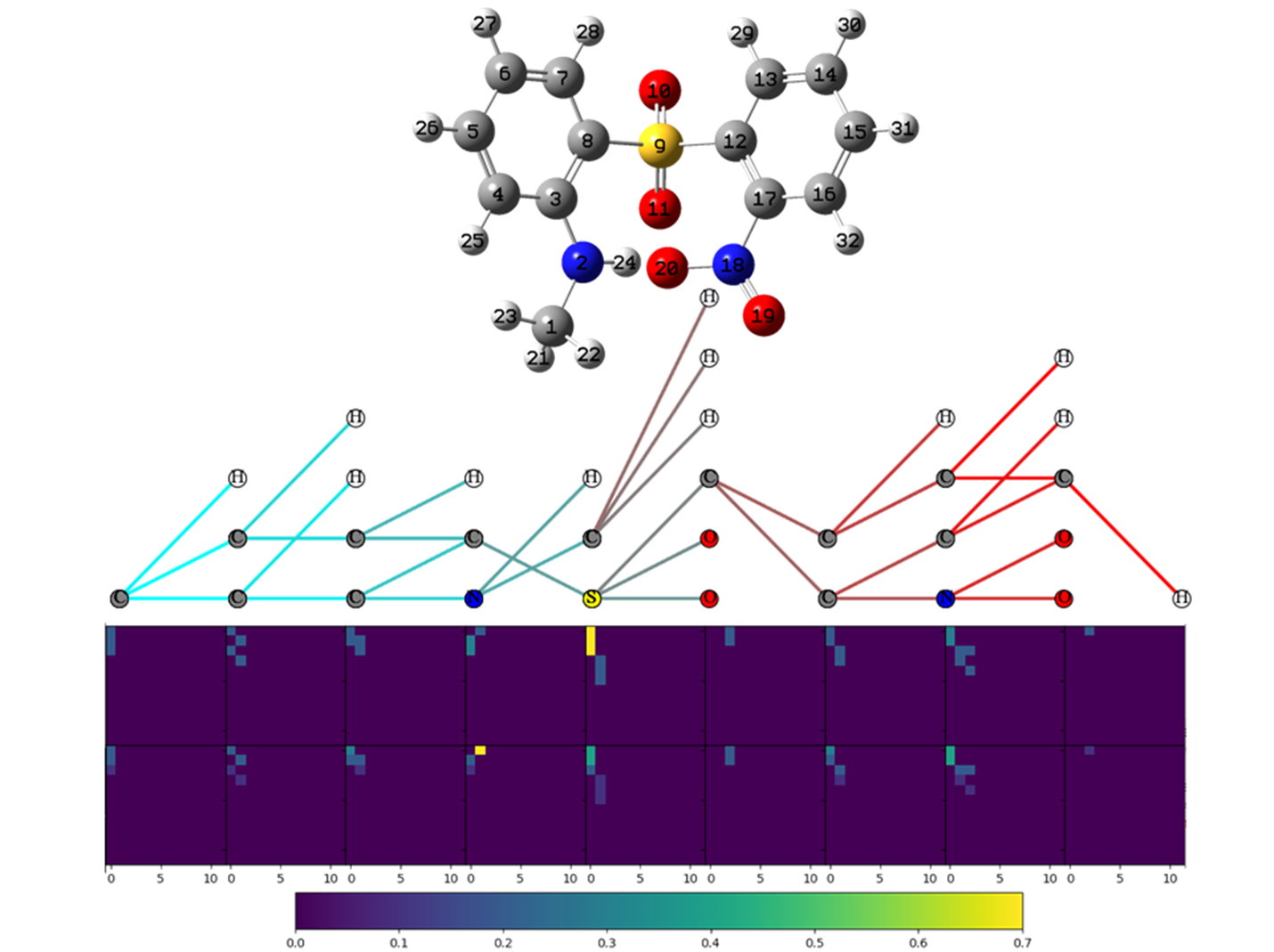
多种性质共同调控

**进一步工作：**

**基于连接的连接**

**最终我们使用python实现了分子图数据编码为矩阵数据的这一过程的转换，并通过python图形库matplotlib 将序列二分图与对应矩阵可视化表示出来。**





**分子筛选**

**txtCNN模型**

**分子生成**

**charRNN模型**

**Results and Discussion**

**数据集**

**训练结果**