**A new molecular coding method assists machine learning model 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 recurrent neural network to generate new molecules. This molecular coding method allows recurrent neural networks to effectively generate molecules with specified properties. 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 generation

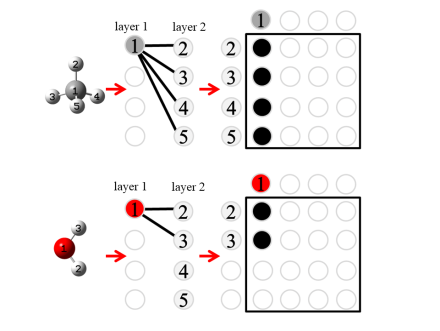
**Introduction**

In the past, designing molecules to meet specific properties or functions required professional knowledge and long-term experience. Machine learning methods are expected to accelerate this process. Our goal is to enable machine learning models to automatically design molecules based on the desired properties. Molecule is different from image. It is a kind of graph structure in non-Euclidean space. To make machine learning models understand and design molecules, we first need to characterize the molecular structure. The characterization of molecules is divided into representing molecules as strings and graphs. The traditional SMILES method is based on people's pre-set rules to describe the molecular structure as a string. In recent years, graph representation methods have gradually increased. Compared with string, graph can describe molecules more directly and expressively.

**Methods**

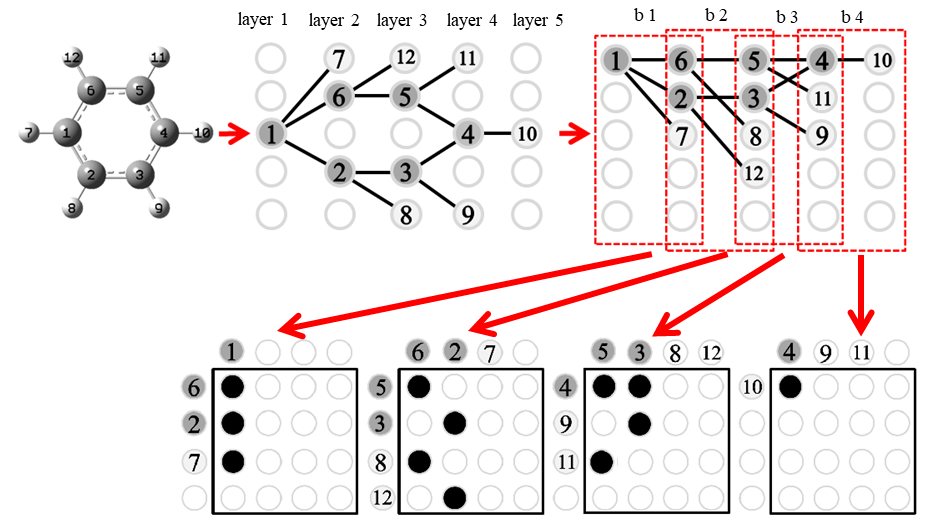
The molecular is a weighted, undirected, sparse graph. The vertices and edges in the graph are regarded as the atoms and chemical bonds of molecules. The vertices and edges in the graph are regarded as atoms and chemical bonds in the molecule. In graph theory, bipartite graphs have interesting properties. A bipartite graph is a graph whose vertices can be divided into two disjoint and independent sets U and V such that every edge connects a vertex in U to one in V. They only consider the connection relationship between vertices in two sets. For simple graphs such as methane or water molecules, we can easily find a way to convert them into bipartite graphs, and then convert the bipartite graph to a matrix according to formula 1.

(1)



**Fig. 1** Structures

However, this method is no longer feasible for slightly more complex structures. After the vertices are divided into two sets, the connection relationship of some vertices in the set cannot be expressed. Therefore, we consider disassembling the molecule, and then draw it on multiple bipartite graphs, and finally convert each bipartite graph into a matrix in order. In addition, we sort the vertices in each set of the bipartite graph according to their corresponding atomic numbers. As shown in Figure 2, the topological information of the benzene molecule is completely preserved during this transformation.

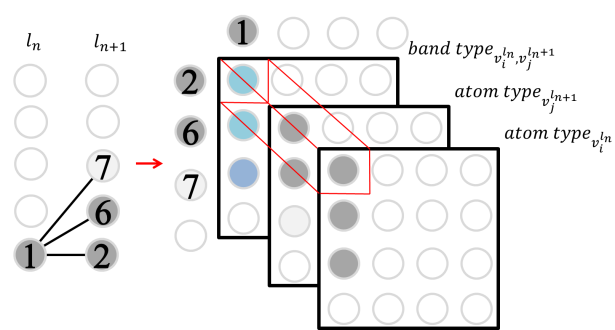


**Fig. 2** Structures

We can add more details about molecular information in the matrix, as shown in formula (2) (3). The vector x in formula (3) contains the atom type and bonding information in the molecule. Each bipartite graph corresponds to a three-dimensional tensor, as shown in Figure 3.

(2)

(3)



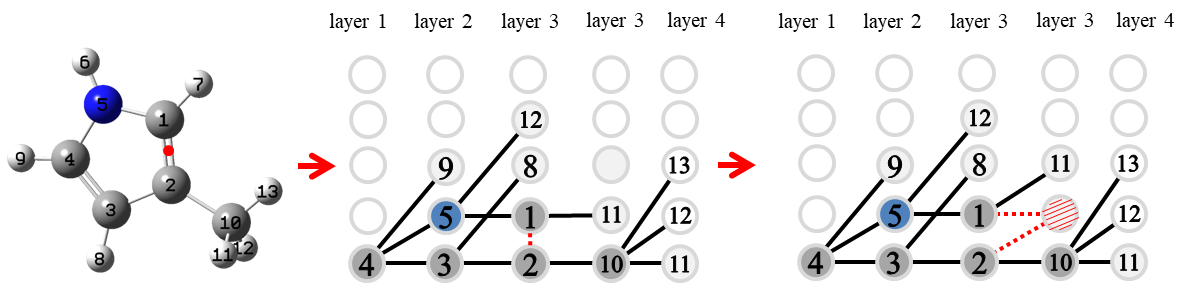
**Fig. 2** Structures

Therefore, how to map molecular graph to multiple bipartite graphs according to structural relationships is the key to this encoding process. The breadth first search algorithm is a classic graph traversal algorithm. The algorithm searches for some vertices, and then finds all unseen vertices connected by these vertices . This brings us inspiration. Set V and set can be used as two sets of a bipartite graph. The bipartite graph is established by observing the connection relationship between the vertices of these two sets. The process is shown in Algorithm 1

|  |
| --- |
| **Algorithm 1 Create Layers algorithm** |
| 1: **Input:** graph, vertex in G  2: **Output:**  3: Set as the root vertex , .  4: Find all adjacent vertices of vertex in and put these adjacent vertices into set ：.  5:  6: **Repeat**  7:  8: **For** set **do**  9: Find all adjacent vertices of and put these adjacent vertices into set .  10: .  11:  12: **until**  13: **return** |

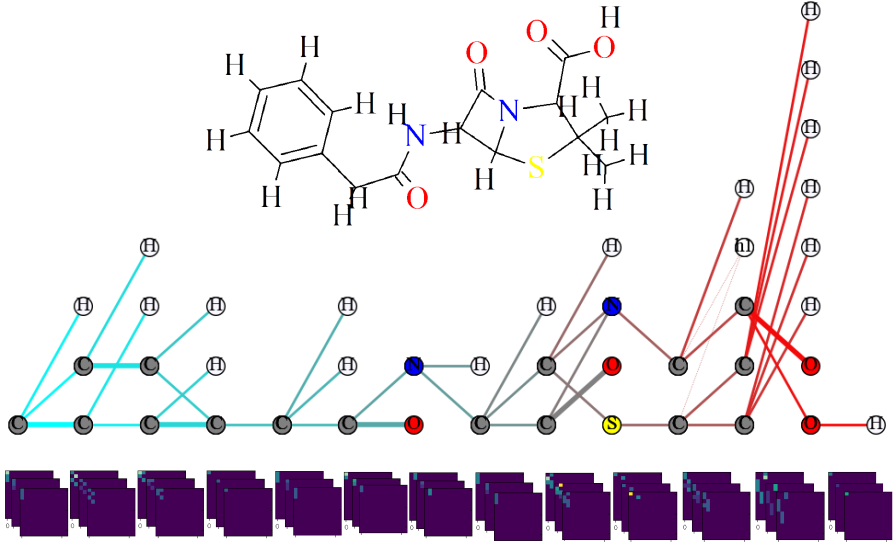
In graph theory, a graph is a bipartite graph if and only if it does not contain odd cycles. For this mapping method, if we want to represent all the bonds, we also need that the molecule does not contain odd circles. However, odd-circle structures are common in molecular structures. To solve this problem, we can use the placeholder method to make the graphs equivalent. Algorithm 2 shows this process. When it is found that there are two vertices connected in the same layer, add placeholder vertices in the next layer, and add the connection relationship between these two vertices and placeholder vertices, As shown in Figure 3.

|  |
| --- |
| **Algorithm 2 Add Placeholder Layers algorithm** |
| 1: **Input:** graph, vertex in G,  2: **Output:**  3: Get from **Create Layers()**  4:  **For**  set **do**  5: **If** and is connected in **then**  6: Add placeholder to , add vertex and edges to graph.  7: Calculate the number of layers:  8: Count the number of vertices in each layer  9:  10: **Return** |

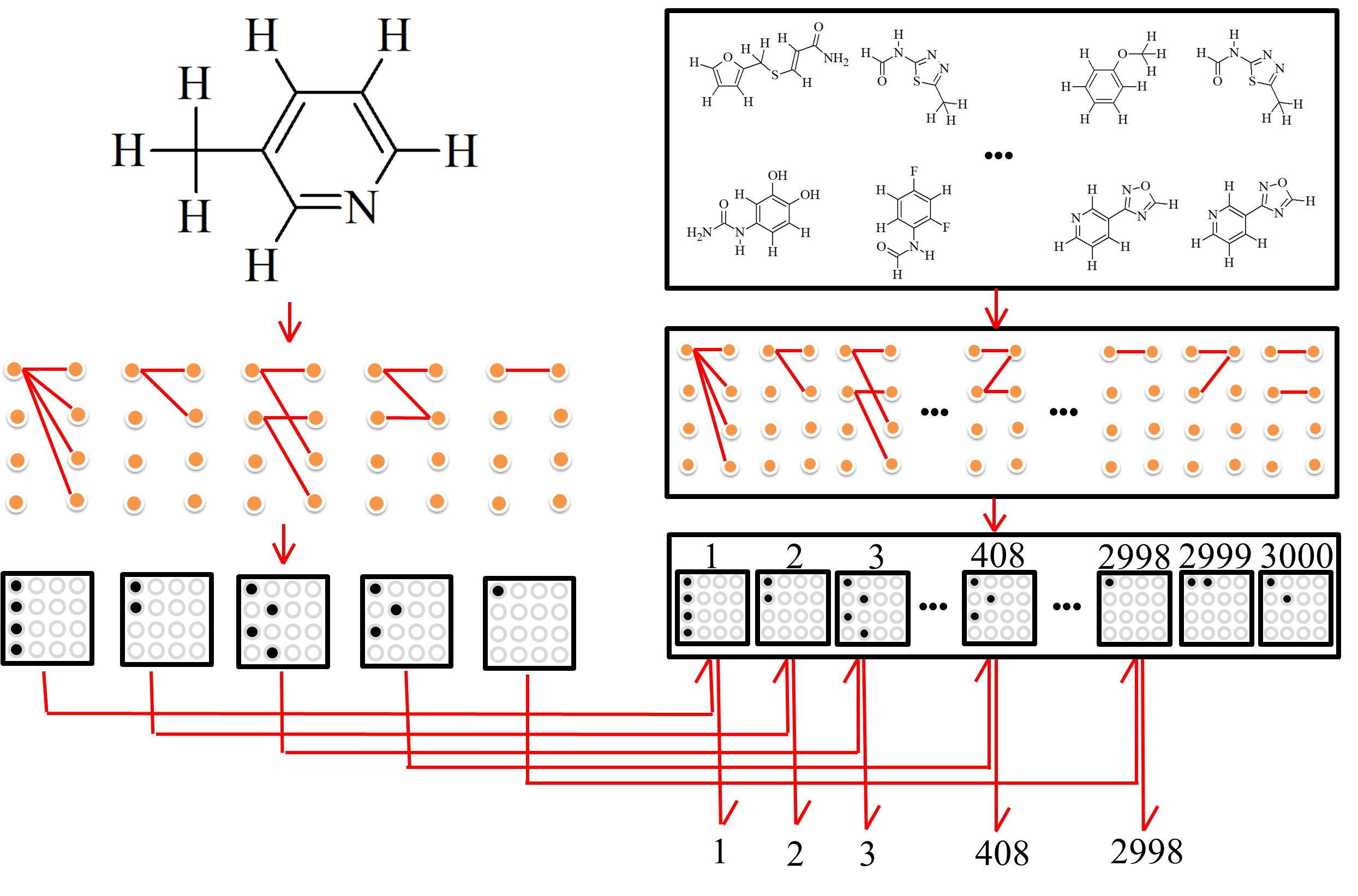


Selecting different vertices will lead to different vertex traversal orders, and different vertex traversal orders will result in different bipartite graph sequences. We believe that for this encoding method, the best traversal method is to divide the graph into as many levels as possible.

|  |
| --- |
| **Algorithm 3 Select Root Vertex** |
| 1: **Input:** graph  2: **Output:**  2:  3:  **For** **do**  4: Get from **Add Placeholder Layers ()** and addto  5:  6: **Return** |



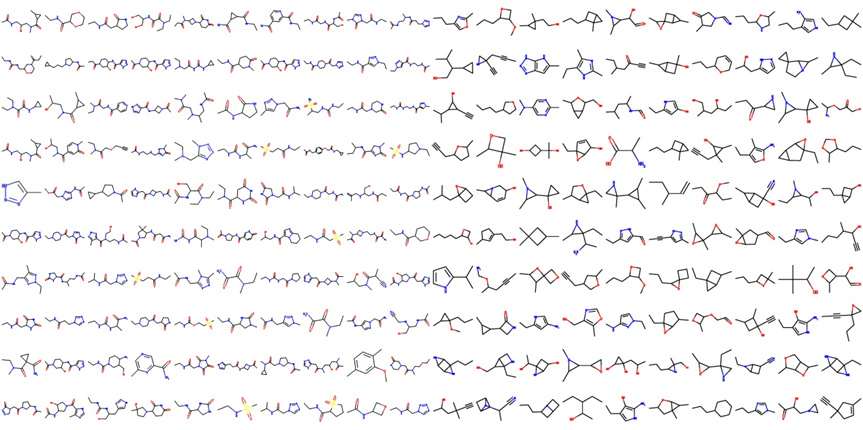
These tensors are regarded as the building blocks of molecules, and they are reused in the molecular library. We sort these building units according to the frequency of occurrence in the molecular library, and use the ordinal number as their index. When a molecule is broken down into tensors in order, the numerical value of the molecule can be encoded by finding their corresponding index.



通过序列模型

条件生成

在ZINC 以及QM9数据集的分子生成的结果



Conclusion

在这篇文章中我们提出了对分子图编码的方法。它可以将分子图结构映射到矩阵上，同时保持它的层次信息。

快速

不仅在序列生成领域，这种方法可以应用于多种机器学习任务。

但是，对于稍微复杂的结构这种方法不再可行。顶点被分为两个集合后，集合中的一些顶点具有的连接关系无法被表示出来。因此我们考虑将分子进行拆解，进而将它绘制在多个二分图上，最后把每一个二分图按顺序转换为矩阵

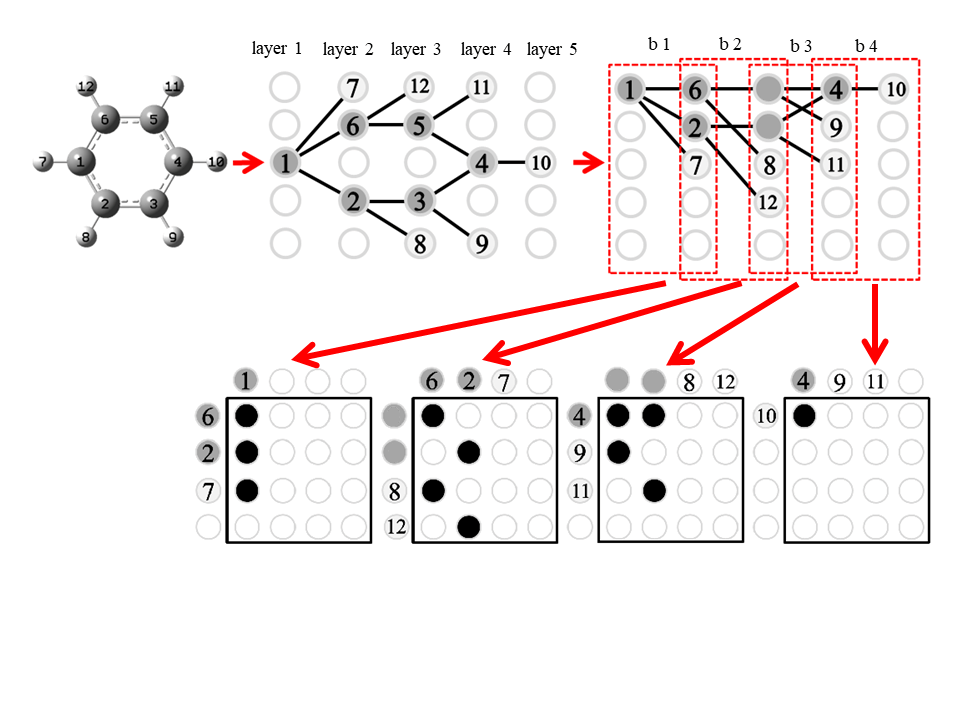
在以往，对分子进行设计使其符合特定性质或功能需要专业的知识、长期的经验以及反复试错，机器学习方法有望加速这一过程。我们的目标是反向设计分子。分子与在欧式空间的图像不同，它是一种非欧空间的图结构数据。使机器学习模型理解以及设计分子，我们首先需要对分子的结构进行表征.对于分子的表征大致分为将分子表示为字符串或者图。传统的SMILES方法是基于人们预先设定的规则将分子结构描述为字符串。近年来图表示方法逐渐增多。相较于字符串，图对分子的描述更加直接，更具表现力.

**Methods**

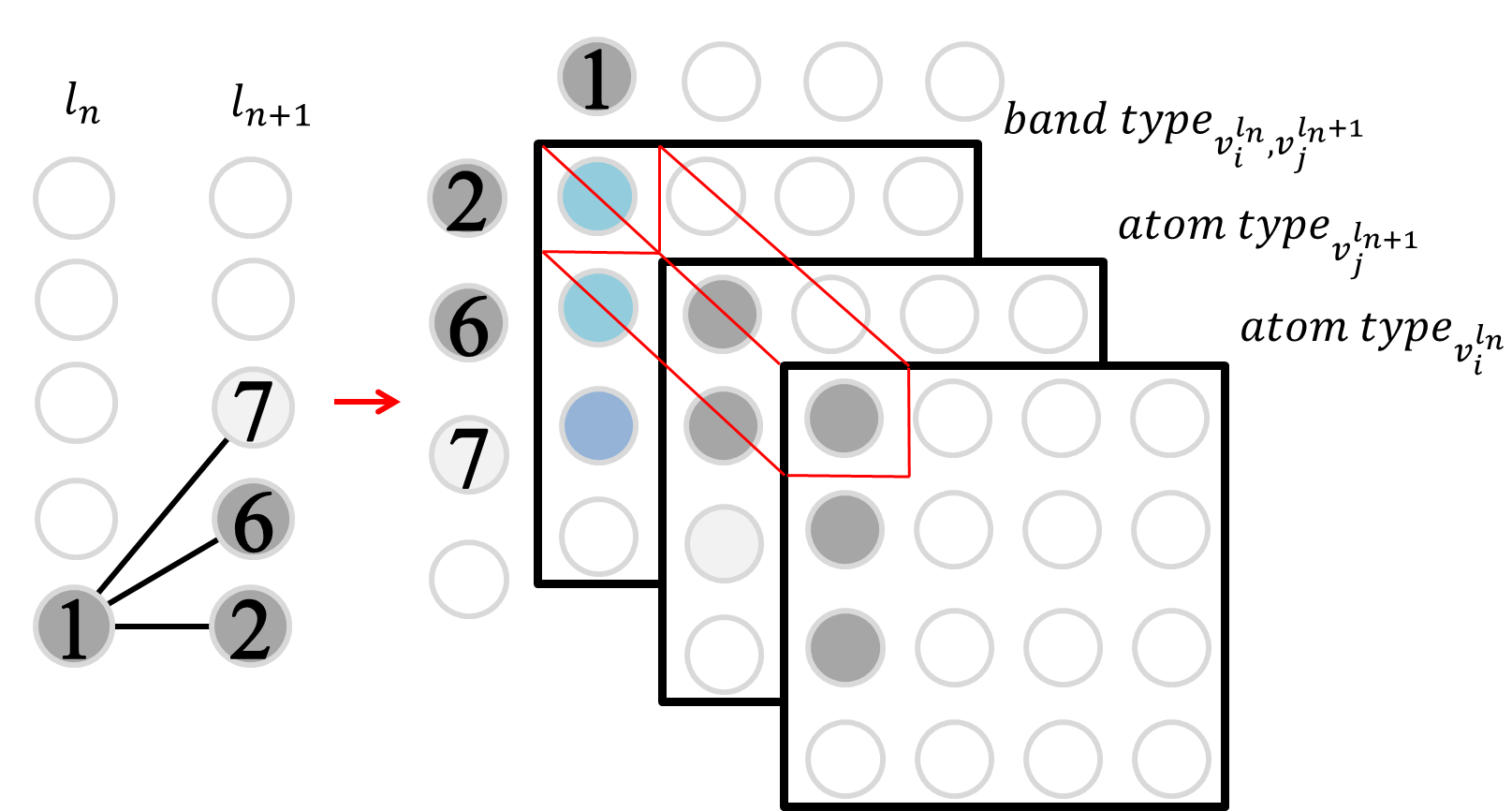
分子是一种加权的，无向的，稀疏图。图中顶点和边被视为分子中的原子和化学键。我们尝试用一种新的方式对图结构进行表示。在图论中，二分图有着有趣的性质，它只考虑两个集合间顶点的连接关系。对于甲烷或者水分子这样简单的结构而言，我们可以轻易寻找到一种方式将其转换为二分图。接着依照公式1将二分图转换为矩阵。

(1)

但是，对于稍微复杂一些的结构这种方法不再可行。顶点被分为两个集合后，集合中的顶点具有连接关系无法被表示出来。因此我们考虑将分子进行拆解，进而将它绘制在多个二分图上，对于每一个二分图再进行其到矩阵的转换。另外，我们对二分图的每一个集和中的顶点按照它们对应的原子序数进行排序。如图二所示，对于该苯分子，图的拓扑关系在这一变换过程中被完整的保留了下来。



我们可以在矩阵中增加更多关于分子信息的细节,如公式(2)(3)所示。公式（3）中向量x包含了分子中原子类型和成键信息。每个二分图对应一个三维的张量，如图3所示



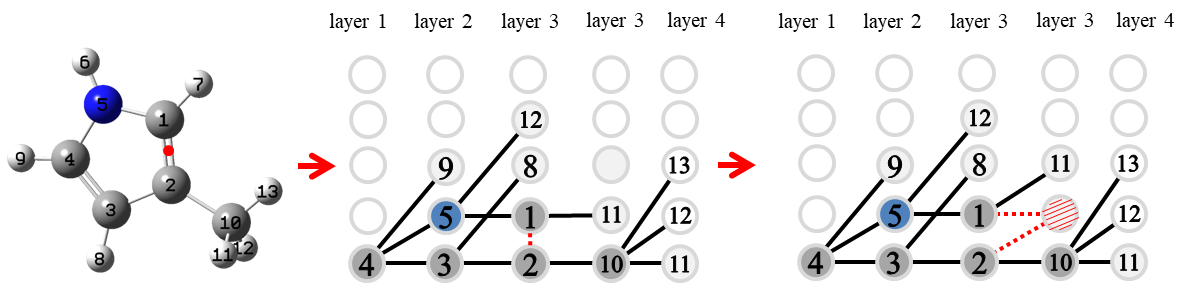
那么，如何将分子图按照结构关系映射到多个二分图中是完成这一编码过程的关键。广度优先算法是一种经典的图遍历算法。它会搜寻一些顶点，然后寻找这些顶点所连接的全部未见到顶点，通过不断迭代遍历整张图。这给了我们重要启发，集合与集合可以作为一个二分图的两个集合。通过观察这两个集合间顶点的邻接关系建立二分图。如算法1以所示

This continuous bipartite graph structure can be obtained by the idea of breadth first search. The algorithm is designed as follows

|  |
| --- |
| **Algorithm 1 Create Layers algorithm** |
| 1: **Input:** graph, vertex in G  2: **Output:**  3: Set as the root vertex , .  4: Find all adjacent vertices of vertex in and put these adjacent vertices into set ：.  5:  6: **Repeat**  7:  8: **For** set **do**  9: Find all adjacent vertices of and put these adjacent vertices into set .  10: .  11:  12: **until**  13: **return** |

In graph theory, a graph is a bipartite graph if and only if it does not contain odd cycles.

在图论中，一个图是二分图当且仅当它不包含奇环。对于这一映射方法而言，如果我们想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。为了解决这一问题，我们可以使用占位符的方法使图等价。

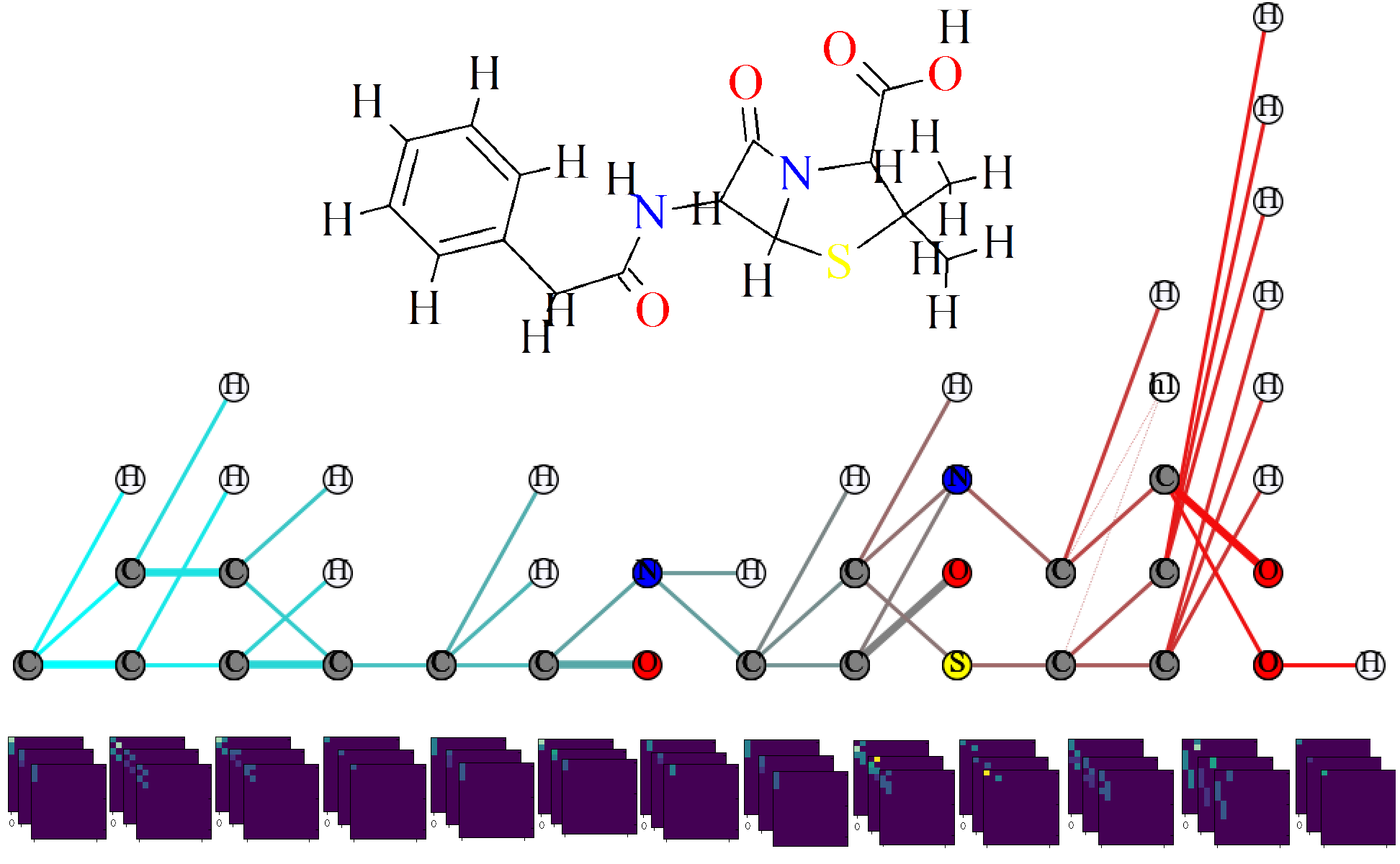


|  |
| --- |
| **Algorithm 2 Add Placeholder Layers algorithm** |
| 1: **Input:** graph, vertex in G,  2: **Output:**  3: Get from **Create Layers()**  4:  **For**  set **do**  5: **If** and is connected in **then**  6: Add placeholder to , add vertex and edges to graph.  7: Calculate the number of layers:  8: Count the number of vertices in each layer  9:  10: **Return** |

选取不同的顶点会得到不同的顶点遍历顺序，顶点遍历顺序的不同会得到不同的二分图序列，我们认为对于该编码方法而言，最好的遍历方法是使图尽可能划分出更多的层级。

|  |
| --- |
| **Algorithm 3 Select Root Vertex** |
| 1: **Input:** graph  2: **Output:**  2:  3:  **For** **do**  4: Get from **Add Placeholder Layers ()** and addto  5:  6: **Return** |

对于L中的ln ,ln+1,建立矩阵序列



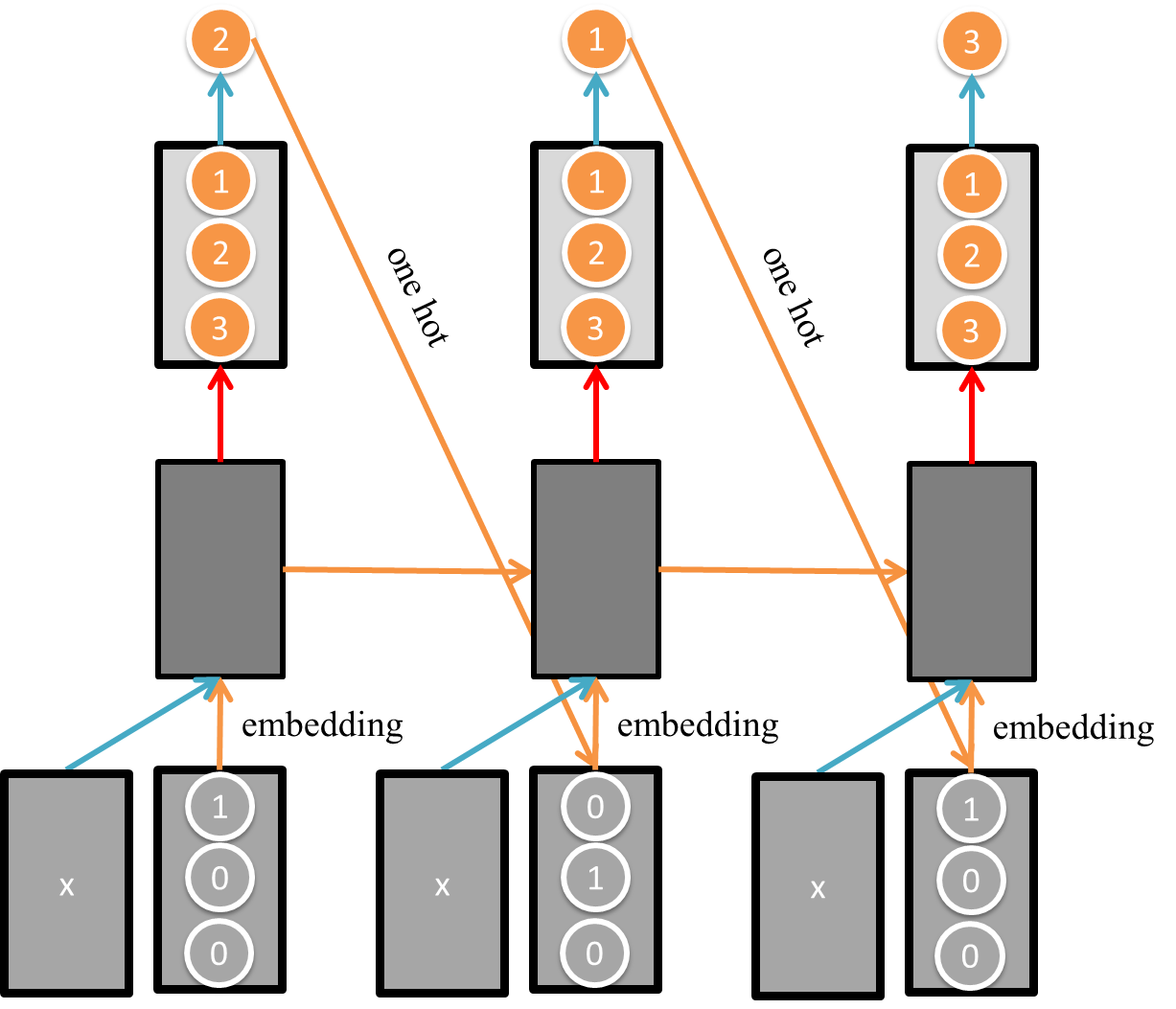
算法对青霉素分子结构编码

Edge

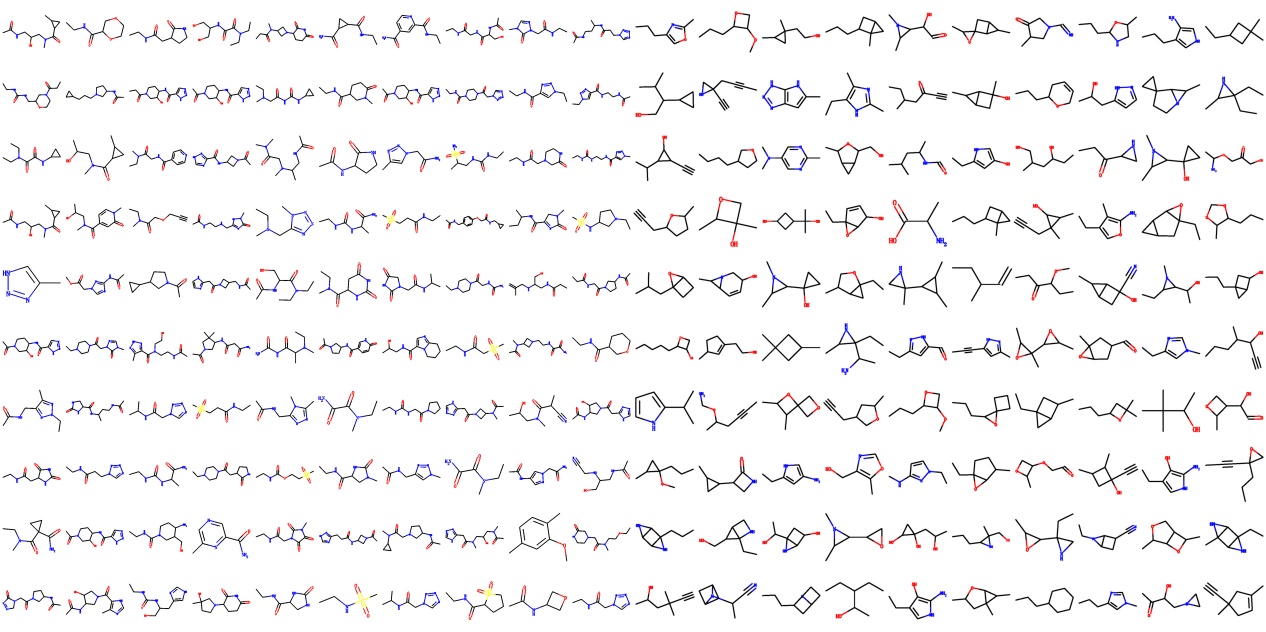
建立占位顶点使得图等价

**Results and Discussion**

这些张量被视为分子的构造单元，它们在分子库中被大量复用。我们将这些构造单元按照在分子库中出现的频次排序，并将序号作为它们的索引。当一个分子按顺序被拆解为张量时，找到它们对应的索引就可以实现对分子的数值编码。

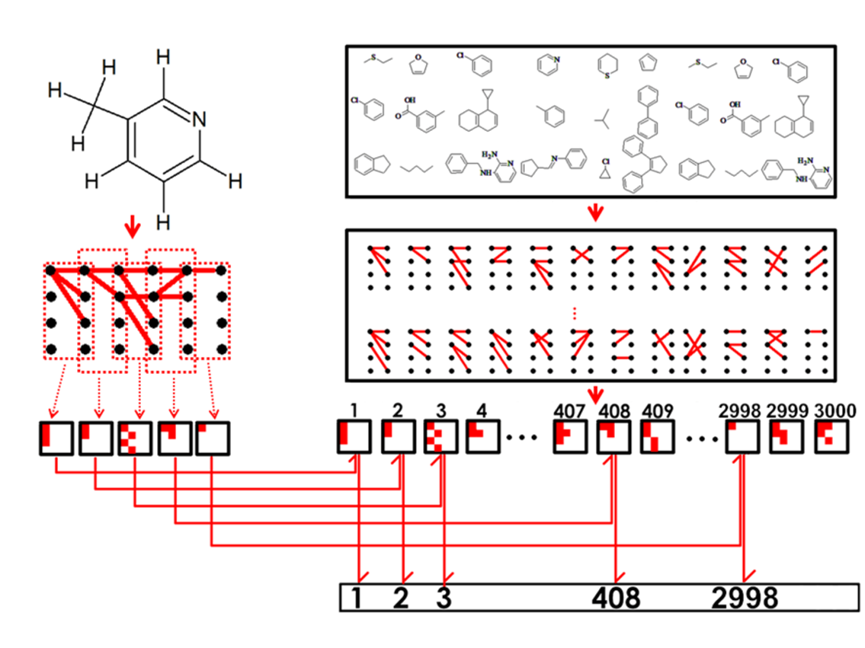


These tensors are regarded as the building blocks of molecules, and they are reused in the molecular library. We sort these building units according to the frequency of occurrence in the molecular library, and use the ordinal number as their index. When a molecule is broken down into tensors in order, the numerical value of the molecule can be encoded by finding their corresponding index.



**Conclusion**

该矩阵编码方法不仅可以用于图生成领域，分子性质的预测。

****