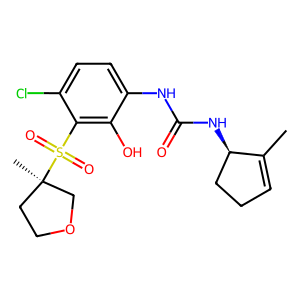
# 抗癫痫活性计算机虚拟筛选

1、筛选依据

脂溶性（BBB通透性）、成药性（成药性原则）、类药性（与临床常用癫痫治疗药物的比对，包括卡马西平、丙戊酸钠、司替戊醇、左乙拉西坦、唑尼沙胺、托吡酯）

2、导入化合物

使用 pandas 和 numpy 包合并 10 个化合物库，合并后的总化合物数量为 14,845。去重后，剩余 9,983 个化合物。利用 rdkit 包根据 SMILES 号将这 9,983 个化合物导入 Python。其中，有 424 个 SMILES 号是无效的最终成功导入 9,559 个化合物进行下一步的分析。



3、特征提取与特征工程

化合物具有局部特征（local features）和全局特征（global features）

4、局部特征的提取和处理

使用rdkit.Chem 模块，可以提取表一的特征，并构建原子矩阵和键矩阵。为了避免越大的值被认为越重要的歧义问题，对所有的原子和键的特征进行独热编码，除了原子质量，其他特征都除以100。

表一图示

描述已自动生成

|  |  |  |
| --- | --- | --- |
| Feature type | Attribute | Description |
| Atomic  features | Atom type | Known chemical elements (by atomic number) |
|  | Degree | Number of bonds the atom is involved in |
|  | Formal  charge | Electronic charge assigned to an atom |
|  | Chirality | Unspecifed, tetrahedral CW/CCW, or other types of chirality |
|  | Number of H | Number of bonded hydrogen atoms |
|  | Hybridization | sp, sp2, sp3, sp3 d, or sp3 d2 |
|  | Aromaticity | Whether the atom is aromatic |
|  | Atomic mass | Mass of the atom |
| Bond  features | Bond type | Single, double, triple, or aromatic |
|  | Conjugated | Whether the bond is conjugated |
|  | Ring | Whether the bond is in a ring |
|  | Stere | Stereochemistry of bonds (none, any, E/Z or cis/trans) |

为了最大程度地利用小分子的局部特征，需要将原子和键的特征聚合并嵌入到一个向量中，这有利于后续任务的使用。为此，对原子矩阵和键矩阵进行线性主成分分析（linear PCA）。分别选择每个矩阵中方差最大的第一个主成分来表征原子和键的局部特征。

通过这一步骤，每个分子中不同大小的局部特征都被聚合成一个具有相同维度的表示，有利于进一步的分析。为了进一步过滤掉变异较低的特征，将原子矩阵的主成分和键特征的主成分合并后，再次进行PCA。选择合并后的特征中的前 50 个主成分，以解释键和原子特征的大部分变异。

5、全局特征的提取和处理

利用 rdkit.Chem 中的 Descriptors 模块，获取了209个分子描述符，并将其作为全局特征进行进一步处理。为了防止分子描述符中绝对值较大的特征占据主导地位，我们对分子描述符进行了 z 分数归一化，以确保所有值都在相同的范围内。最后，将全局特征和局部特征合并后，过滤掉每一列方差为零的特征，以保证最终特征矩阵的多样性和信息含量。

6、聚类和评估

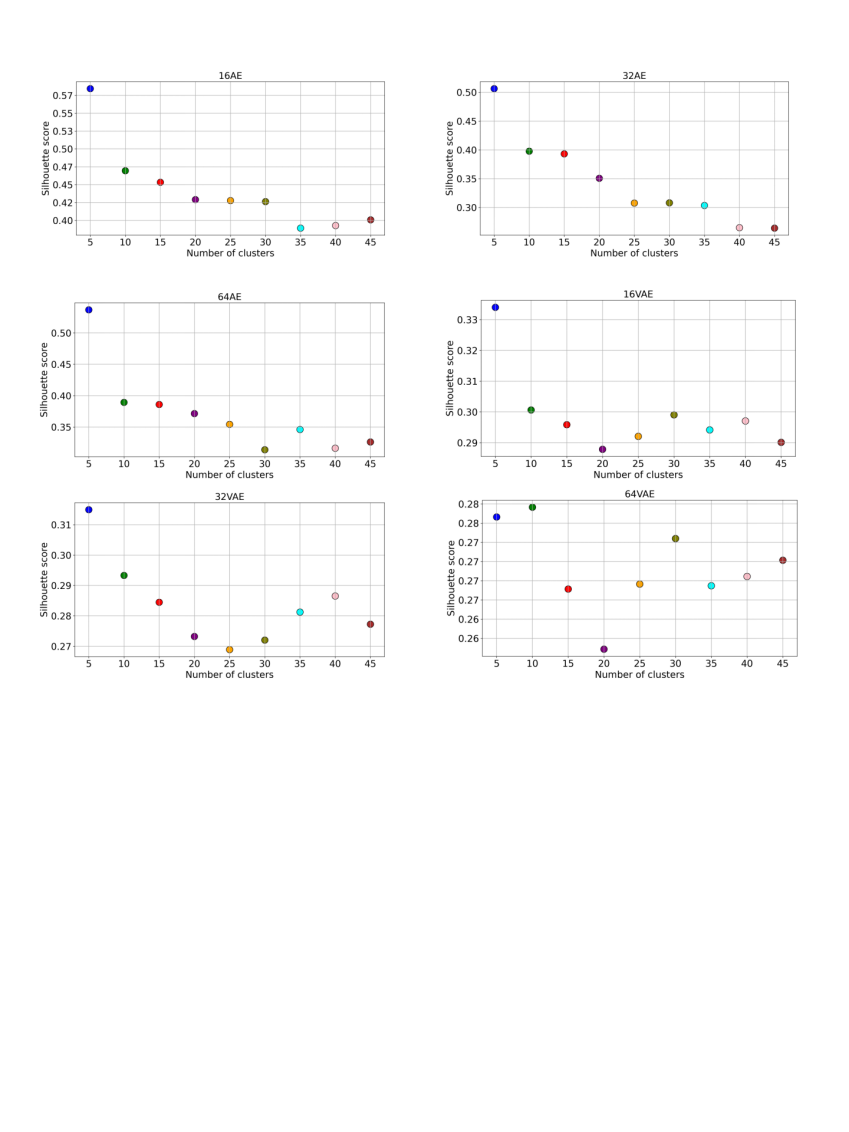
在建立构效关系之前，进行聚类分析具有以下优势：

1、有利于确保化合库的多样性；

2、能够了解小分子的异质性和相似性，从而深入理解药物的潜在作用机制；

3、有助于扩大化合物的选择范围。

由于合并局部特征和全局特征后依然存在大量特征，我们选择使用自编码器（AE）和变分自编码器（VAE）对数据进行压缩和降维。自编码器是一种无监督的神经网络，将输入分子映射到生成分子特异性特征，用于重建输入分子。AE在低维重构的同时不会丢失关键信息。而VAE是一种生成式神经网络，通过编码器和解码器实现输入的分子特异性特征以概率的方式嵌入到潜在空间中，然后通过解码器从潜在空间中生成输入数据。因此，VAE在处理大规模数据时更加实用和可行。因此，我们使用AE和VAE对特征进行降维和压缩（分别使用三种嵌入数量：16、32和64）。之后，采用K-Mean算法进行聚类，并通过Silhouette score指标对聚类结果进行评估。最终，选择最优的聚类结果进行进一步的活性筛选。



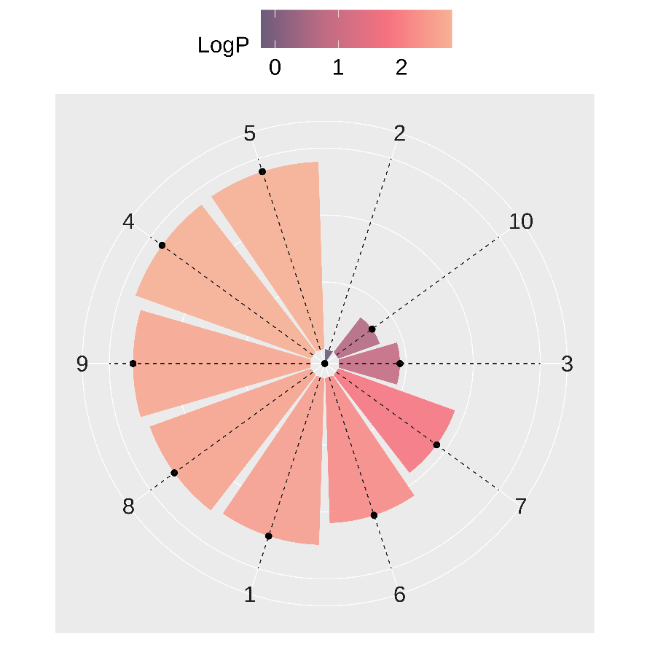
根据结果显示，当使用63个嵌入数时，VAE变分自编码器在分为10个类别时表现最佳。聚类结果如下图所示，第一张图展示了通过 t-SNE（t-distributed Stochastic Neighbor Embedding）将聚类结果可视化，而第二张图则表示了聚类的数量以及每个类别中小分子的数量。

图表

描述已自动生成

7、脂溶性原则筛选

使用rdkit中的Descriptors.MolLogP函数获取相应的LogP值，然后根据聚类结果计算每个类的LogP均值。结果如下图所示。在10个类中，1、4、5、8、9类别具有较好的脂溶性。因此，我们选择这几个类别进行下一步的筛选。



9、成药性原则筛选

利用rdKit 包中的 Descriptors 模块，分别获取相对分子质量（Molecular Weight）、脂溶性（CLogP）、氢键供体（Hydrogen Bond Donors）、氢键受体（Hydrogen Bond Acceptors）的值。结果如下图所示，将第1类排除（相对分子质量大于500），而其他的类别符合成药性原则。

表格

描述已自动生成

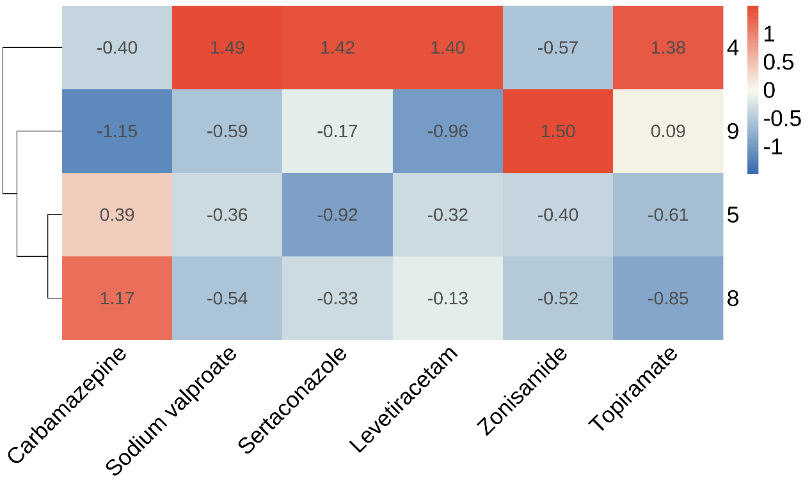
雷达图

描述已自动生成雷达图

描述已自动生成

9、类药性原则筛选 （临床治疗药物相似性）

为了进一步筛选化合物，我们将第4、5、8、9四个类别与临床治疗药物（卡马西平、丙戊酸钠、司替戊醇、左乙拉西坦、唑尼沙胺、托吡酯）进行结构相似性的比较。结果如下图所示，第4类和第9类与目前的临床治疗药物相比具有较好的结构相似性。



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 临床治疗药物 | 2D结构 | |  | 对比药物 | 2D结构 | 相似性 | 相似对比图 |
| 卡马西平 | ['Carbamazepine'] |  | | 吴茱萸碱 | ['Anatabine (dicitrate)'] | 0.13793 | Anatabine (dicitrate)-Carbamazepine |

10、结果桑基图（4、9）

我们将9559个化合物分成10个类，10个类所包含的化合物数量分别为1（1157）、2 （381）、3（498）、4 （355）、5 （1399）、6 （1020）、7 （44）、8 （3676）、9 （194）、10 （460）。其中4和9两个类通过筛选。4（355）、9（194）类一共具有个549化合物。我们选择这两个类进行下一步的实验验证。4类中具有355个化合物，9类中具有194个化合物。筛选过程如下图所示。

图表

描述已自动生成

