使用 MolAICal 计算药物化学过滤器 MCFs 及 MCE-18 描述符

作者：Qifeng Bai

邮箱：molaical@yeah.net

个人主页：https://molaical.github.io/baiqf.html

1. 引言

药物化学过滤器（MCFs）的作用是剔除含有所谓“结构警示”的化合物[1]。含有此类结构单元的分子要么具有不稳定或反应性基团，要么会发生生物转化，进而产生有毒代谢物或中间体。PAINS（泛测定干扰化合物）过滤器由一系列子结构过滤器组成，旨在减少筛选库中的假阳性结果、测定伪影以及非特异性生物活性分子。其设计依据是：结构中的某些片段可能会引发不良性质 —— 如反应性、螯合作用、胶体聚集形成以及染料相关效应等，这些都会干扰测定结果。由于 MCFs 和 PAINS 均涉及反应性，因此在 MolAICal 中，MCFs 的功能涵盖了 MCFs 和 PAINS 所涉及的结构。

“MCE-18”全称为“药物化学演进，2018”，是一种原创分子描述符[2]，用于有效评估具有药理学相关性分子的新颖性和先导化合物潜力。MCE-18 计算公式：

$$\text{MCE-18}=\left(\text{AR}+\text{NAR}+\text{CHIRAL}+\text{SPIRO}+\frac{\overbrace{\text{sp}^3+\text{Cyc-Acyc}}^{\text{NCSPTR}}}{1+\text{sp}^3}\right)\times Q^1$$

参数说明

1. AR：芳香环或杂芳香环的存在情况（0 = 不存在，1 = 存在）。
2. NAR：脂环或杂脂环的存在情况（0 = 不存在，1 = 存在）。
3. CHIRAL：手性中心的存在情况（0 = 不存在，1 = 存在）。
4. SPIRO：螺原子的存在情况（0 = 不存在，1 = 存在）。
5. sp^3：sp^3杂化碳原子的比例（范围为 0 至 1）。
6. Cyc：sp^3杂化的环碳原子比例（范围为 0 至 1）。
7. Acyc：sp^3杂化的非环碳原子比例（范围为 0 至 1）。
8. NCSPTR：复合项，计算公式为（sp^3+ Cyc - Acyc）。
9. Q[1]：归一化二次指数，用于表征结构骨架的支化程度。

如需了解更详细的 MolAICal 相关信息，请访问：https://molaical.github.io。

2. 材料

2.1. 软件要求

1. MolAICal：https://molaical.github.io 或 https://molaical.gitlab.io

注意：确保 MolAICal 安装正确！

2.2. 示例文件

1. 所有必要的教程文件可从以下网址下载：

https://gitee.com/molaical/tutorials/tree/master/024-mcf_mce18

第一部分：使用 MolAICal 计算化学过滤器（MCFs）

1. 为单个 SMILES 字符串计算 MCFs

```
#> molaical.exe -tool mcf -s "CC(=O)OC1=CC=CC=C1C(=O)O"
```

2. 为单个 SMILES 字符串计算 MCFs（不输出属性）

```
#> molaical.exe -tool mcf -s "CC(=O)OC1=CC=CC=C1C(=O)O" -p false
```

3. 计算 MCFs 并指定输出文件及格式（如 “text”“json” 或 “csv”）

```
#> molaical.exe -tool mcf -s "CC(=O)OC1=CC=CC=C1C(=O)O" -o result.csv -f csv
```

4. 计算 MCFs 并指定输出文件（不输出属性）

```
#> molaical.exe -tool mcf -s "CC(=O)OC1=CC=CC=C1C(=O)O" -o result.csv -f csv -p false
```

默认情况下，还会输出一些属性用于分子的进一步评估：

- ✧ **SMILES**：输入分子的 SMILES 序列。
- ✧ **Valid**：判断分子是否有效，值为 True（有效）或 False（无效）。
- ✧ **Passes filters**：判断分子是否通过 MCFs 评估，值为 True（通过）或 False（未通过）。若值为 True，表明该分子通过了 MCFs 和 PAINS 过滤器。
- ✧ **LogP**：辛醇 - 水分配系数，指物质在辛醇相中的浓度与在水相中的浓度之比。
- ✧ **SA Score**：即合成可及性分数，用于粗略评估特定分子的合成难度（评分 10 表示难，1 表示易）。该分数由分子片段的综合贡献得出 [3]。

- ✧ **NP Score：**即类天然产物分数，是一种贝叶斯指标，用于评估分子与天然产物所涵盖的结构空间的相似程度[4]。
- ✧ **QED：**即药物相似性定量估计，是一个 [0,1] 区间内的值，用于评估分子成为可行药物候选物的可能性。与 SA 分数类似，QED 中的描述符阈值在过去十年中有所变化，当前阈值可能未涵盖最新药物[5]。
- ✧ **Molecular weight（MW）：**化合物中原子量的总和。

从上述网址下载 "024-mcf_mce18/molecules.smi"；"molecules.smi" 可包含多个 SMILES 字符串，因此 MolAICal 可一次性为一个或多个分子计算 MCFs。请尝试以下示例：

1. 从文件中计算 MCFs（文件可包含多行 SMILES，每行代表一个分子的 SMILES）

```
#> molaical.exe -tool mcf -i molecules.smi -o result.csv -f csv
```

2. 从文件中计算 MCFs（不输出属性，文件可包含多行 SMILES，每行代表一个分子的 SMILES）

```
#> molaical.exe -tool mcf -i molecules.smi -o result.csv -f csv -p false
```

注意：用户可选择上述任一方式计算 MCFs。如需更详细的帮助，请参考 MolAICal 手册。

第二部分：使用 MolAICal 计算 MCE-18

1. 为单个 SMILES 计算 MCE-18

```
#> molaical.exe -tool mce18 -s "CC(=O)OC1=CC=CC=C1C(=O)O"
```

2. 仅输出 SMILES 和 MCE-18 值

```
#> molaical.exe -tool mce18 -s "CC(=O)OC1=CC=CC=C1C(=O)O" -sm true
```

3. 输入 SMILES 并输出 CSV 格式文件

```
#> molaical.exe -tool mce18 -s "CC(=O)OC1=CC=CC=C1C(=O)O" -o output.csv
```

4. 输入 SMILES 并输出 CSV 格式文件（不输出详细内容）

```
#> molaical.exe -tool mce18 -s "CC(=O)OC1=CC=CC=C1C(=O)O" -o output.csv -q true
```

5. 输入 SMILES 并输出 CSV 格式文件（仅包含 SMILES 和 MCE-18 值）

```
#> molaical.exe -tool mce18 -s "CC(=O)OC1=CC=CC=C1C(=O)O" -o output.csv -sm true
```

6. 输入包含 SMILES 字符串的文件（每行一个）并输出 CSV 格式文件

#> molaical.exe -tool mce18 -i molecules.smi -o results.csv

7. 输入包含 SMILES 字符串的文件（每行一个）并输出 CSV 格式文件（仅包含 SMILES 和 MCE-18 值）

#> molaical.exe -tool mce18 -i molecules.smi -o results.csv -sm true

对 MCE-18 的 值进行分析：

从实际应用场景来看，**45-78** 是一个相对理想的范围

对供应商化合物库（如 ChemDiv、Enamine）的分析显示：

MCE-18 <45：分子结构简单、新颖性低，多为 “旧骨架”，吸引力有限；

MCE-18: 45-78：新颖性足够，符合当前药物化学趋势（45-63），与制药公司近期专利化合物的结构相似性高（63−78）；

MCE-18 > 78：需要手动评估靶点适应性和药物相似性，因为过度复杂可能导致开发困难。

注意：用户可选择上述任一方式计算 MCE-18。如需更详细的帮助，请参考 MolAICal 手册。

如果只研究 MCFs 和 MCE-18 的计算方法，阅读至此已足够。

后续教程内容属于可选部分，在将分子从其它格式转换为 canonical SMILES 格式，或执行批量 canonical SMILES 转换时，可以参考以下教程。

第三部分：多种化学文件格式转换为 canonical SMILES 的方法

为了计算 MCE-18、SA 和 MCFs，常常需要将 mol2、sdf、mol 等化学文件格式转换为 canonical SMILES 表示法。在此背景下，MolAICal 提供了一种基于 RDKIT 的方法，可将 mol2、sdf、mol、pdb、inchi 和 SMILES 等格式的分子结构转换为其 canonical SMILES 表示形式。

1. Mol2 转 canonical SMILES

```
#> molaical.exe -tool tosmi -i lig.mol2 -t mol2
```

2. Mol2 转 canonical SMILES（不显示转换失败信息）

```
#> molaical.exe -tool tosmi -i lig.mol2 -t mol2 -q true
```

注：-q：取值为 true 和 false，“true” 表示抑制转换失败信息。

3. SMILES 转 canonical SMILES

```
#> molaical.exe -tool tosmi -i "CCO" -t smiles
```

4. InChI 转 canonical SMILES

```
#> molaical.exe -tool tosmi -i "InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3" -t inchi
```

5. PDB 转 canonical SMILES

```
#> molaical.exe -tool tosmi -i lig.pdb -t pdb
```

6. mol 转 canonical SMILES

```
#> molaical.exe -tool tosmi -i lig.mol -t mol
```

7. SDF 转 canonical SMILES

```
#> molaical.exe -tool tosmi -i lig.sdf -t sdf
```

注：“-q、-u 或 - r” 也可用于将.mol2、.sdf、.mol、.pdb、InChI、SMILES 等分子格式转换为 canonical SMILES。

第四部分：批量将多种分子文件格式转换为 SMILES 的脚本使用

有时，需要计算.mol2、.sdf、.mol、.pdb 和 InChI 等格式的分子文件的 MCFs 和 MCE-18 描述符。这些文件需批量转换为 SMILES 字符串，才能使用上述方法进行计算。

这里提供两个 Linux bash 脚本用于批量格式转换：

1） 打开 “024-mcf_mce18\batch”，查看 run_combine.sh：

```
--------------------------------------------------------------------------------
#!/bin/bash
cp /dev/null all.smi
for i in `ls -l | grep ".sdf" | awk '{print $9}'`
do
        # Above “.sdf” is a keyword. The below command is to remove the suffix.
        molaical.exe -tool tosmi -t sdf -q true -i $i > tmp.smi
        # Output last line.
       tail -n 1 tmp.smi >> all.smi
done
rm tmp.smi
--------------------------------------------------------------------------------
```

用户可对 “all.smi” 重命名，该文件包含从 SDF 格式示例文件中提取的所有 SMILES 字符串。

2） 打开 “024-mcf_mce18\batch”，查看 run_individual.sh：

```
--------------------------------------------------------------------------------
#!/bin/bash
for i in `ls -l | grep ".sdf" | awk '{print $9}'`
do
        # Above “.sdf” is a keyword. The below command is to remove the suffix.
        tmp=$(echo $i | awk -F'.' '{print $1}')
        # echo $tmp
        molaical.exe -tool tosmi -t sdf -q true -i $i > tmp.smi
        # Output last line.
        tail -n 1 tmp.smi > $tmp'.smi'
done
rm tmp.smi
--------------------------------------------------------------------------------
```

该脚本会将 SDF 文件逐个转换为包含 SMILES 字符串的文件。SDF 文件与其对应的包含 SMILES 字符串的转换文件具有相同的前缀名。

3） 运行 run_combine.sh 和 run_individual.sh

运行脚本前，需正确指定 “molaical.exe” 的路径。在 Linux 控制台中：

对于第一个脚本：

```
#> chmod +x run_combine.sh
#> ./run_combine.sh
```

对于第二个脚本：

```
#> chmod +x run_individual.sh
#> ./run_individual.sh
```

注： 上述示例基于 SDF 格式的分子，但该方法同样适用于 mol2、sdf、mol、pdb、inchi、SMILES 等其他格式的分子。请在对应的脚本中将文件扩展名（如".sdf"）修改为".mol2"等目标格式的扩展名。

参考文献

1. Polykovskiy, D.；Zhebrak, A.；Sanchez-Lengeling, B.；Golovanov, S.；Tatanov, O.；Belyaev, S.；Kurbanov, R.；Artamonov, A.；Aladinskiy, V.；Veselov, M.；Kadurin, A.；Johansson, S.；Chen, H.；Nikolenko, S.；Aspuru-Guzik, A.；Zhavoronkov, A.，Molecular Sets (MOSES)：A Benchmarking Platform for Molecular Generation Models. Frontiers in Pharmacology 2020, Volume 11 - 2020.

2. Ivanenkov, Y. A.；Zagribelnyy, B. A.；Aladinskiy, V. A.，Are We Opening the Door to a New Era of Medicinal Chemistry or Being Collapsed to a Chemical Singularity? Journal of Medicinal Chemistry 2019, 62 (22), 10026-10043.

3. Ertl, P.；Schuffenhauer, A.，Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Cheminform 2009, 1 (1), 8.

4. Ertl, P.；Roggo, S.；Schuffenhauer, A.，Natural product-likeness score and its application for prioritization of compound libraries. Journal of Chemical Information and Modeling 2008, 48 (1), 68-74.

5. Shultz, M. D.，Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs. J Med Chem 2019, 62 (4), 1701-1714.