# MolInsight 操作手册

> 适用版本：0.1.x ｜ 更新日期：2026-10-25 ｜ 界面语言：中文 / English
> 本手册面向实际使用者，按「导入数据 → 整理表格 → 绘图 → 分析 → 建模 → 导出」的顺序介绍每个模块。英文版单独提供（`manual_en.pdf`）。

## 目录


1. 软件简介
2. 安装与启动
3. 界面总览
4. 数据导入与数据集
5. 表格与列操作
6. 2D 结构面板
7. 绘图
8. 分析模块
9. QSAR 建模与预测
10. 导出与报告
11. 三种授权档位的异同
12. 主题、语言与高分辨率
13. 日志、故障排查与常见问题
14. 快捷键
15. 附录

# 1. 软件简介

MolInsight 是一款面向药物化学与 SAR（构效关系）分析的桌面工具，围绕「分子表格 + 2D 结构 + 统计绘图 + 化学信息学分析 + QSAR 建模」组织工作流。

核心能力：

- 导入 SDF / MOL / SMILES / CSV，统一为可筛选、可排序、可计算的分子表
- 35 项分子描述符、分子式/元素组成、标准化、药性规则
- 47 类统计图与化学图，支持与表格、结构面板联动
- 相似性、重复分子、SMARTS 计数、活性阈值/ROC、活性悬崖
- R 基团分解、骨架/侧链、指纹聚类、SOM、相似性网络
- MMP（匹配分子对）与转化规则、结构差异对比、分子对网络
- 3D 构象与形状相似度、虚拟库枚举
- QSAR 回归/分类建模：交叉验证、适用域、预测区间、特征选择、可解释性、模型保存与外部验证

技术栈：PySide6（界面）、RDKit（化学）、NumPy（计算）、Matplotlib（绘图）。除 RDKit 外，分析计算尽量使用纯 NumPy。

# 2. 安装与启动

> 本节适用于 Windows 打包版（`MolInsight.exe`）。**不需要安装 Python、RDKit、conda，也不需要运行任何脚本。**

## 2.1 系统要求

| 项目 | 要求 |
| --- | --- |
| 操作系统 | Windows 10 / 11（64 位） |
| 内存 | 建议 4 GB 以上（大型数据集建议 8 GB+） |
| 磁盘 | 约 500 MB 可用空间 |
| 显示 | 1280×720 以上；支持 125%–300% 缩放 |

## 2.2 获取与安装

1. 获取发布包（通常为 `MolInsight` 文件夹，内含 `MolInsight.exe`）。
2. 将**整个文件夹**复制到本地磁盘，例如 `D:\Apps\MolInsight`。
3. 双击 `MolInsight.exe` 启动。

- 文件夹版不要只复制 exe，需保留同目录的 `_internal` 等依赖文件。
- 单文件版只有一个 exe，直接双击即可；首次启动会稍慢。
- 程序不需要管理员权限，建议放在有写权限的目录。

### 2.3 首次启动

- Windows 可能弹出“已保护你的电脑”（SmartScreen）：点击“更多信息”→“仍要运行”。
- 首次启动会创建用户配置与日志目录，不会修改原始分子文件。
- 主窗口打开后即可导入数据；默认界面为中文。

### 2.4 升级与卸载

- **升级**：关闭程序，用新版本文件夹覆盖旧版本程序文件；个人设置、授权与日志不受影响。
- **卸载**：关闭程序后删除程序文件夹即可。
- 如需彻底清除个人数据，可删除：
  - `%LOCALAPPDATA%\MolInsight\`（日志、崩溃记录）
  - 用户目录下的 `.molinsight_license.txt`、`.molinsight_trial.txt` 等授权/试用文件
  - 注册表 `HKEY_CURRENT_USER\Software\MolInsight`（界面设置、最近文件）

### 2.5 数据与配置位置

| 内容 | 位置 |
| --- | --- |
| 运行日志 | `%LOCALAPPDATA%\MolInsight\molinsight.log` |
| 崩溃记录 | `%LOCALAPPDATA%\MolInsight\crash.log` |
| 授权 / 试用 | 用户目录下的 `.molinsight_license.txt` / `.molinsight_trial.txt` |
| 界面设置 / 最近文件 | 注册表 `HKEY_CURRENT_USER\Software\MolInsight` |

## 3. 界面总览

主窗口自上而下分为：

| 区域 | 说明 |
| --- | --- |
| 菜单栏 | 文件 / 表格 / 绘图 / 分析 / 视图 / 帮助 |
| 工具栏 | 文本筛选、结构大小、SMARTS 搜索、清除筛选、显示列 |
| 中央表格 | 分子列表：`#`、结构缩略图、名称、SMILES、导入属性、计算列 |
| 右侧停靠区 | 结构查看、绘图、MMP、骨架、各分析工具（可拖动/浮动/关闭） |
| 状态栏 | 左侧分子总数/可见数/来源；右侧当前行数值统计；最右品牌名 |

### 3.1 常用交互

- 点击表头排序；**Ctrl+点击**多列排序（新列为主键）。
- 双击任意单元格：打开结构面板并显示该行。
- 表格右键：查看大图、复制 SMILES、设置标签。
- 列宽与列序自动记忆；隐藏列状态不保存。
- 未保存修改时窗口标题带 `*`，关闭/打开/合并前会确认。

### 3.2 视图菜单

- **主题**：浅色 / 深色 / 跟随系统，立即生效并记住。
- **语言**：中文 / English，切换后**立即生效**；数据列名、SMILES 等内容不翻译。

## 4. 数据导入与数据集

### 4.1 支持格式

| 格式 | 扩展名 | 说明 |
| --- | --- | --- |
| SDF | `.sdf` `.sd` | 多记录，每条一个分子；保留 SDF 属性 |
| MOL | `.mol` `.mdl` | 单分子文件 |
| SMILES | `.smi` `.smiles` `.ism` | 每行 `SMILES [名称]`，空白分隔 |
| CSV / TSV / TXT | `.csv` `.tsv` `.txt` | 第一行必须是表头；逗号或制表符分隔 |

### 4.2 解析规则

#### SDF / MOL

- 解析失败的记录计入“失败数”并跳过，不中断其余记录。
- 名称取 `_Name`，没有则用 `Mol_序号`（MOL 用文件名）。
- SMILES 由 RDKit 重新生成规范化写法，可能与源文件原文不同。
- SDF 属性全部按字符串导入；`_Name` 不单独作为属性列。

#### SMILES 文件

- 空行、`#` 开头行跳过（但会计入失败数）。
- 第一行若为 `smiles` / `smi` / `smile` 表头则跳过。
- 第一个 token 为 SMILES，第二个 token 为名称，后续忽略。

#### CSV / TSV

- `.csv` 用逗号，`.tsv` / `.txt` 用制表符；**不自动探测分隔符。**
- 编码按 UTF-8（含 BOM）读取。
- 第一行始终当表头；没有表头的 CSV 会丢掉第一行数据。
- SMILES 列识别：
  1. 表头为 `smiles` / `smi` / `smile`（可写成 `SMILES|ID`）；
  2. 找不到时，若第一列表头不是 `name` / `id` / `mol_id`，用第一列；
  3. 若第一列是名称列且没有 SMILES 列，则所有行都会失败。
- 名称列：表头严格等于 `name` / `id` / `mol_id`。
- 其余列作为属性列导入；空单元格不建值。

> 建议 CSV 至少包含 `SMILES` 列；需要名称时再加 `Name` 列。

### 4.3 导入方式

- **文件 → 打开分子文件...**（Ctrl+O）：可多选，**替换**当前数据集。
- **文件 → 合并分子文件...**：追加到当前数据集，要求已有数据。
- 直接把文件拖入窗口：按“打开”处理（替换当前数据）。
- **文件 → 最近打开**：最多 8 条；菜单底部可「清除最近打开」。

导入完成后状态栏显示成功数；有失败行时弹窗提示成功/失败数量。

### 4.4 数据集与列

- `#`：原始行号（0 起）。
- `Name`：分子名称。
- `SMILES`：规范化 SMILES。
- `Structure`：表格中动态渲染的 2D 结构列。
- 导入属性：来自 SDF/CSV 的原始字段。
- 计算列：描述符、标签、分析写回列、建模写回列等。

**数值列识别**：对每列取前 50 行，至少一半能转成数值才视为数值列。全为空或前 50 行全是文本的列不会被当作数值列。

### 4.5 免费版分子数上限

免费版单次会话最多 **100 个分子**。导入或合并超过上限时，程序保留前 100 个并弹窗说明；被截断的行不会进入数据集。

## 5. 表格与列操作

### 5.1 排序、筛选与搜索

| 功能 | 入口 | 说明 |
| --- | --- | --- |
| 文本筛选 | 工具栏「筛选」 | 对名称、SMILES、导入属性、计算列做大小写不敏感子串匹配 |
| SMARTS 搜索 | 工具栏「SMARTS」+ 搜索 | 子结构匹配；非法 SMARTS 会提示 |
| 数值范围筛选 | 表格 → 表格显示 / 范围筛选 | 对某一数值列设最小–最大值 |
| 相似性搜索 | 分析 → 相似性搜索 | 按指纹 Tanimoto 相似度排序 |
| 重复检测 | 分析 → 检测重复分子 | 按 InChIKey 分组 |
| 清除筛选 | 工具栏 | 清空文本/SMARTS/骨架/数值范围等筛选 |

- 文本、SMARTS、范围、骨架、子结构等筛选之间是 **AND** 关系。
- 相似性搜索与重复检测可勾选「高亮而不筛选」，只着色不隐藏其他行。

- 高亮行不会因「清除筛选」而清除。

## 5.2 表格显示

- **结构大小**：小 150×110 / 中 200×140 / 大 280×190。
- **显示列**：工具栏右侧按钮，可全选/全不选/逐列勾选。
- **热力着色**：表格 → 表格显示 / 范围筛选，可选白→蓝、白→红、冷→暖、黄→绿；按当前可见行归一化。
- 表头悬停显示该列可见行的 n / min / max / mean / median。

## 5.3 计算性质（描述符）

**表格 → 计算性质…**（Ctrl+P）

- 35 项描述符，默认全部勾选，可全选/全不选。
- 后台计算，可取消；完成后写入计算列并刷新表格。
- **注意**：该操作会整体替换现有计算列（包括 Label、派生列、分析结果列），执行前请确认。

描述符分组：

| 类别 | 描述符 |
| --- | --- |
| 理化/拓扑 | MolecularWeight、ExactMolWt、MolLogP、MolMR、TPSA、NumHDonors、NumHAcceptors、NumRotatableBonds、NumRings、NumAromaticRings、NumSaturatedRings、NumAliphaticRings、HeavyAtomCount、FractionCSP3、FormalCharge、NumHeteroatoms、NumAmideBonds、NumAromaticHeterocycles、NumAromaticCarbocycles、NumSaturatedHeterocycles、NumAliphaticHeterocycles、NumSpiroAtoms、NumBridgeheadAtoms、Chi0n、Chi0v、Chi1n、Chi1v、Kappa1、Kappa2、Kappa3、EStateSum |
| 成药性 | QED、SA_score、CNS_MPO、LogP_SA_score |

## 5.4 计算列管理

- **表格 → 计算列管理…**（Ctrl+Alt+C）：重命名或删除计算列。
- **表格 → 清除计算性质**：清空全部计算列（含标签与分析写回列），无确认、无撤销。

## 5.5 数值列派生（pIC50 等）

**表格 → 数值列派生…**

- 选择源数值列与变换，可自定义新列名（留空自动命名）。
- 支持：pIC50（M / nM / μM）、由 pIC50 反算 IC50、log10、取反、平方、倒数、z-score、min-max 归一化。
- 结果中的 NaN/Inf 写为空值。

## 5.6 分子式与元素组成

**表格 → 分子式 / 元素组成列…**

- 可生成 `MolecularFormula` 与 `elem_C`、`elem_H`、`elem_N`、`elem_O`、`elem_S`、`elem_P`、`elem_F`、`elem_Cl`、`elem_Br`、`elem_I`。
- 默认勾选分子式与 C/H/N/O/S/P；至少选一项。
- 合并写入，不清空已有计算列。

## 5.7 公式计算列

**表格 → 公式计算列…**（Ctrl+Alt+F）

- 输入表达式，点「预览」查看前 200 行结果，再「写入计算列」。
- 运算符：`+ - * / % // **`；比较：`== != < <= > >=`；逻辑：`and or not`。
- 函数：`log10`、`log`、`ln`、`exp`、`sqrt`、`abs`、`min`、`max`、`round`、`floor`、`ceil`、`sin`、`cos`、`tan`、`isfinite`、`isnan`、`where(条件, 成立, 否则)`。
- 列名直接书写（区分大小写），例如 `6 - log10(IC50_uM)`、`where(Activity > 10, 1, 0)`。
- 只能引用导入属性与计算列，不能引用 `Name` / `SMILES` / `#`。
- 缺失值参与算术的结果为空。

## 5.8 整列批量操作

**表格 → 整列批量操作…**

| 操作 | 说明 |
| --- | --- |
| 字符串替换 | 查找值为空时只替换 `None`；非空时对字符串做区分大小写的子串替换 |
| 数值 clip | 把数值截断到下限–上限 |
| 数值替换 | 数值差 <1e-9 视为相等，替换为新值 |
| 截断小数位数 | 保留 N 位小数 |

- 作用于**全部原始行**，不受当前筛选影响。
- 可改写导入属性列，无撤销。

## 5.9 列拆分 / 合并

- **列拆分**：按分隔符把一列拆成多列，列名默认 `{原列名}_1`、`_2`…
- **列合并**：多选 2 列以上，按分隔符拼成一个新列；全空的行写空值。

## 5.10 分组聚合与数据透视

- **分组聚合**：按某列分组，对一个或多个数值列计算 n / min / max / mean / median / std；可导出 CSV，或把组均值写回数据行。
- **数据透视表**：分组列 × 数值列，聚合函数 mean / count / sum / min / max / std；结果在独立面板中展示，可导出 CSV。

- 分组值为空的行不进入任何组。

### 5.11 统计检验

**表格 → 统计检验…**

- **单样本 t 检验**：检验某数值列均值是否显著不为 0。
- **单因素方差分析（ANOVA）**：按分组列比较各组均值。
- 输出统计量、p 值与显著性（α=0.05）。

### 5.12 缺失值、列摘要与相关矩阵

- **缺失值统计**：逐列缺失数与占比（空串/纯空白也算缺失）。
- **列统计摘要**：数值列 n / mean / std / min / max / median。
- **列相关性矩阵**：所选数值列的 Pearson 相关，可切换显示 p 值。

结果都在右侧独立面板，可导出 CSV。

### 5.13 批量标注与标签

- **给选中行设置标签**（Ctrl+L）或表格右键「设置标签…」：写入 `Label` 列，留空即清空。
- **分析 → 批量标注**（停靠面板）：
  - SMARTS 匹配：把匹配行写为指定标签；
  - 按 Murcko 骨架分组：`{前缀}_{序号}`（需商业授权）；
  - 按聚类列标注：`{前缀}_{Cluster值}`（需已有 `Cluster` 列）。

## 6. 2D 结构面板

右侧「结构查看」面板随表格选中行同步：

- 显示 2D 结构、名称、SMILES 与部分属性。
- 复制 SMILES / InChI / MOL（属于导出，试用/免费版受限）。
- 保存结构图片 PNG（导出，需商业授权）。
- 上一行 / 下一行切换。
- ECFP 位点亮显：输入位点编号与半径，高亮支撑该位点的原子环境。

## 7. 绘图

### 7.1 基本流程

1. 在表格中筛选/排序到目标分子（绘图只使用**当前可见行**）。
2. 绘图菜单选择图表类型，或打开绘图面板后在「类型」下拉切换。
3. 在对话框中设置 X/Y/Z、分组、颜色、分箱等参数。
4. 确定后绘制到右侧绘图面板。

5. 可缩放/平移、点击联动表格与结构、保存图片、复制、导出绘图数据。

## 7.2 图表类型（47 类）

| 分组 | 图表 |
| --- | --- |
| 散点 / 直方 | 散点图、直方图 |
| 多维 / 响应面 | 3D 散点、并行坐标、3D 响应面、响应面等高线 |
| 分布诊断 | QQ 图、ECDF、蜜蜂群、山脊密度、分组 KDE、分位数曲线 |
| 散点增强 | 性质象限、2D 密度、分箱均值±误差带、六边形分箱、LOWESS、联合分布、置信椭圆、散点矩阵、相关性-目标条形 |
| 回归诊断 | 残差图、四联诊断、杠杆-残差、Bland-Altman |
| 轮廓 / 交互 | 雷达图、交互作用、分组柱状、瀑布图、哑铃图、环形图 |
| 构成 / 合规 | 帕累托、Lipinski 合规、骨架频率、骨架累积覆盖率、MMS 系列行进、活性悬崖气泡 |
| 分子对 / 分类评估 | MMP Δ、校准曲线、提升/增益、PR 曲线、决策曲线 DCA、阈值-指标扫描、预测概率分布 |
| 其他 | SAR 取代基矩阵（Ctrl+Alt+R，独立对话框） |

## 7.3 通用参数

- **列选择**：X / Y / Z、颜色映射、点大小映射、分组列、多选数值列。
- **分箱/分辨率**：柱数、分箱数、网格数、显示条数、阈值网格密度、百分位点数（0 通常表示自动/全部）。
- **统计与误差**：均值/中位数、SEM / SD / 95%CI、KDE 带宽、QQ 自由度。
- **外观**：透明度、cmap（plasma / inferno / cividis / RdBu_r / coolwarm）、标签、离群点、凹口、中位线。
- **参考线/阈值**：悬崖相似度与 Δ、覆盖率参考、残差界值、瀑布基线。

> 默认 X 轴常是 `#`（行号），绘图前请改成真正的数值列。

## 7.4 交互与导出

- 点击散点联动主表格与结构面板；悬停显示分子名与数值。
- 滚轮缩放、拖拽平移、复位。
- **保存图片**：PNG / SVG / PDF；DPI 默认随屏幕缩放（100%→120，200%→192），可在面板中调整。
- **复制**：Ctrl+C 复制当前绘图。
- **导出绘图数据**：CSV，列因图类型而异。
- 绘图配置按类型记忆，下次打开自动预填。

# 8. 分析模块

> 除特别说明外，分析工具都在右侧以**停靠面板**打开（非模态），可与其他面板并排、拖动或关闭。多数结果会写入**计算列**，写回后主表格自动刷新。

## 8.1 相似性搜索

**分析 → 相似性搜索…**

- 查询分子默认当前选中行。
- 指纹：Morgan (ECFP) / FeatMorgan / MACCS。
- 相似度阈值默认 0.30（0–1）。
- 结果：排名 / 分子 / 相似度；可「应用到表格（筛选相似分子）」。
- 可勾选「高亮而不筛选」，只着色。

## 8.2 检测重复分子

**分析 → 检测重复分子…**

- 按 InChIKey 分组，列出重复次数 ≥2 的组。
- 结果：InChIKey / 分子 / 组内重复次数。
- 可筛选到重复分子或仅高亮。

## 8.3 SMARTS 子结构计数

**分析 → SMARTS 计数…**

- 每行一条 SMARTS；支持 `# 注释`，行尾 `.` 表示忽略该模式。
- 对每个分子统计 `GetSubstructMatches` 的非重叠匹配数。
- 预览前 100 行；可写入数据集，每个 SMARTS 生成一列计数。

## 8.4 分子标准化

**分析 → 分子标准化…**

- 选项：去除盐/溶剂（保留最大有机片段）、中和电荷、归一化官能团、保留原始 SMILES。
- 先「预览前 200 个」，确认后「应用并写回数据集」。
- 写回会替换分子结构与 `SMILES` 列；勾选保留原始 SMILES 时新增 `OriginalSMILES`。
- 标准化流程：Cleanup → 片段选择 → 中和 → 归一化。

## 8.5 药性规则评估

**分析 → 药性规则评估…**

- 评估 RO5（Lipinski 五规则）、Veber、PAINS 等违例。
- 结果写入计算列（每个规则/违例一个布尔或计数列），可在表格中筛选。

## 8.6 活性阈值与 ROC/AUC

**分析 → 活性阈值与 ROC/AUC…**

| 参数 | 默认 | 说明 |
| --- | --- | --- |
| 活性数值列 | 第一个数值列 | 连续活性值 |
| 方向 | 值越大越活跃 | 或值越小越活跃 |
| 阈值模式 | 自动（中位数） | 自动中位数 / 自动 Youden / 自动平衡准确率 / 自定义 |
| 阈值 | 0 | 仅自定义可编辑 |

- 运行后状态栏显示：阈值、活跃数/有效数、活跃百分比、AUC、AP、EF(5%)、EF(20%)、BEDROC(α=20)。
- 可「写入数据集（Active 列）」（1/0，缺失为空）。
- 可显示 ROC 曲线或 PR 曲线。
- **注意**：自动 Youden 模式当前以中位数生成分类标签，ROC 图上标注 Youden 最优点；若需严格按 Youden 分类，请用自定义阈值。

## 8.7 活性悬崖检测

**分析 → 活性悬崖检测…**

- 参数：活性列、指纹（Morgan/FeatMorgan/MACCS）、相似度 ≥（默认 0.7）、活性差 ≥（默认 1.0）。
- 结果：左分子 / 左活性 / 右分子 / 右活性 / 相似度 / |Δ活性| / Δ方向。
- 可导出 CSV，或在绘图区显示「Δ活性 vs 相似度」散点。
- 只分析全部行，不受当前筛选影响；活性缺失的分子不参与配对。

## 8.8 SAR 取代基矩阵

**绘图 → SAR 取代基矩阵散点…**（Ctrl+Alt+R，模态对话框）

- 选择 X 类别列（如 R1）、Y 类别列（如 R2）、活性列（颜色）。
- 可勾选点抖动，避免相同类别点完全重合。
- 「绘制 SAR 矩阵」：类别轴散点，颜色表示活性。
- 「Δ 视图」：以出现频率最高的 R2 为参考，绘制各 R1 行相对参考的活性差热图。
- 前提：先用 R 基团分解把 R 位点写回数据列。

## 8.9 R 基团分解

**分析 → R 基团分解…**（需商业授权）

- 选择核心模板（参考分子）与活性列。
- 运行后：每个分子与参考分子做 MCS，公共核心至少 3 个原子；生成 R 位点列（如 `R@3`）。
- 结果表：分子 + 各 R 列（侧链 SMILES）。
- 可「写入数据集（追加 R 列）」、导出 CSV、绘制「R 位点 × 活性」热图。

- MCS 逐分子计算，分子多时较慢；可取消。

## 8.10 骨架 / 侧链分析

**分析 → 骨架 / 侧链分析**（Ctrl+S，需商业授权）

- 分析类型：骨架分析 / 侧链分析。
- 骨架类型：Murcko 骨架 / 通用骨架；最小分子数可过滤低频骨架。
- 结果表：骨架/侧链 SMILES、结构图、分子数/出现次数、占比；选活性列时追加平均/中位活性。
- 双击骨架行可筛选主表格；「清除骨架筛选」恢复。
- 骨架-活性箱线图：按 Murcko 骨架分组，组数 ≥2 时附 ANOVA p 值。
- 导出 CSV 或 SDF（骨架/侧链结构）。
- **注意**：活性下拉可能默认选中 `#`（行号），请手动改成真实活性列。

## 8.11 指纹聚类与投影

**分析 → 指纹聚类与投影…**（需商业授权）

| 参数 | 默认 | 说明 |
|---|---|---|
| 指纹 | Morgan (ECFP) | Morgan / FeatMorgan / MACCS |
| 半径 | 2 | 1–3 |
| 位数 | 2048 | 1024 / 2048 / 4096（MACCS 固定 167） |
| 投影 | PCA | 安装 scikit-learn 时可选 t-SNE |
| 算法 | 层次聚类（AGNES） | 层次 / KMeans / DBSCAN |
| 距离阈值 | 0.3 | 层次聚类，Jaccard 距离 |
| 簇数 K | 6 | KMeans |
| ε / MinPts | 0.3 / 2 | DBSCAN |
| 连接法 | average | average / complete / single |
| 最小簇大小 | 1 | 小于该值标为 0（未归类） |

- 运行后预览表：分子 / 簇号 / PC1 / PC2。
- 可「写入数据集（Cluster 列）」、「在绘图区显示投影」、「PCA 载荷/解释方差」、「高亮所选簇」。
- KMeans 固定随机种子 42；DBSCAN 使用 Jaccard 距离。

## 8.12 SOM 自组织映射

**分析 → SOM 自组织映射分析…**（需商业授权）

- 特征来源：分子描述符（数值列）或结构指纹。
- 描述符模式默认全选所有数值列；建议只保留需要的描述符。
- 参数：网格行/列（默认 10×10）、迭代轮数（300）、学习率（0.5）、邻域半径（0=自动）、种子（42）、U 分位数（0.5）、最小簇分子（2）、拓扑（矩形/六角）、标准化（z-score / median-IQR）、初始化（随机/PCA）。

- 「按 n 建议网格」：`ceil(sqrt(5n))`，限制 2–20。
- 训练后可查看：U-matrix、化学地图、命中计数、Component planes、QE 曲线、密度诊断、每格 QE、margin、性质场、分量平面相关、簇关系热图、簇统计、簇活性富集、分组格图、残差离群、拓扑距离。
- 写回列：`SOM_Cluster`、`SOM_Row`、`SOM_Col`、`SOM_Residual`。
- 可保存/加载 SOM 模型（`.npz`；保存需商业授权）。
- 双击地图格子可联动选中主表格中的分子。

## 8.13 分子相似性网络

**分析 → 相似性网络...**（需商业授权）

- 参数：指纹、相似度阈值（默认 0.7）。
- 分子数 >2000 时直接阻止。
- 结果：节点数、边数、连通分量数、孤立节点数。
- 可绘制网络拓扑图或相似度矩阵热图；热图可聚簇重排、按分组列着色。
- 可写入 `Network_Component` 列（连通分量编号，0 起）。

## 8.14 MMP 分析

**分析 → MMP 分析**（Ctrl+M，需商业授权）

| 参数 | 默认 | 说明 |
|---|---|---|
| 最大切点数 | 2 | 2–3 |
| 最大切断键 | 5 | 1–20 |
| Δ 性质 | （无） | 选数值列后计算 Δ = 目标 − 起始 |
| 范围 | 全部分子 | 或按 Murcko 骨架分组（同骨架才配对） |

- 只分析当前**可见/筛选后**的分子。
- 结果（转化规则）：SMIRKS、分子对数、平均 Δ、标准差、最小/最大 Δ、显著性 p（单样本 t 检验）。
- 选中规则后查看分子对；双击分子对可定位主表格；「结构差异对比」高亮变化侧链。
- 可导出规则 CSV / 分子对 CSV。
- 「分子对网络」可把匹配分子对画成网络：可选弹簧/环形布局、最少共享规则、每节点边数；点击节点查看结构，悬停边查看 SMIRKS 与 Δ 范围。
- p<0.05 整格高亮，`*` / `**` / `***` 表示 0.05/0.01/0.001 水平。

## 8.15 3D 构象与形状分析

**分析 → 3D 构象与形状分析...**（Ctrl+Shift+C，需商业授权）

- 选择分子、构象数（默认 10，最多 50，实际保留 ≤20）、随机种子。
- ETKDGv3 生成 + MMFF/UFF 优化；RMSD 去冗余。
- 结果表：构象 / RMSD→#0 / 回转半径 / 体积 / RMSD 簇。

- 「形状相似度 vs 全数据集」：以查询分子第 0 号构象为基准，计算 ShapeTanimoto 相似度，显示 Top5。
- 「写回相似度列」：列名 `ShapeSim_{索引}`。
- 构象聚类阈值固定 0.5 Å；分子多时较慢，无进度条。

### 8.16 虚拟库枚举

**分析 → 虚拟库枚举…**（Ctrl+Alt+V，需商业授权）

- 前提：已通过 R 基团分解写入 `R...` 列。
- 每位点保留出现频率最高的前 N 个取代基（默认 3）；组合数上限 5000。
- 结果：各 R 位点组合 + 综合得分（各位点平均活性的均值）+ Δ vs 参考组合。
- 表内只预览前 200 行，导出 CSV 包含全部组合。
- 参考组合：每个位点出现频率最高的取代基。

### 8.17 批量标注

见 5.13。

## 9. QSAR 建模与预测

**分析 → QSAR 建模…**（需商业授权；试用版可用但不能导出）

### 9.1 数据准备

1. 导入分子；特征直接使用表格中的数值性质列（可用「表格 → 计算性质」生成描述符，或在 QSAR 对话框点「计算指纹特征…」把 ECFP/FCFP/MACCS 指纹位写成列）。
2. 准备数值目标列：
   - 回归：连续活性值（如 pIC50）。
   - 分类：只能含 0/1（可用公式列 `where(Activity > 10, 1, 0)` 生成）。
3. 目标列必须是数值列；目标缺失的行会被剔除，特征缺失按「缺失填充」策略（mean/median/mode/zero/keep）处理。

### 9.2 基本流程

1. 选择任务（回归 / 分类）与目标列。
2. 选择方法并设置参数。
3. 选择分割方式与预处理。
4. 在特征列表中勾选表格性质列（至少选一列；可先用「计算指纹特征…」生成指纹列）；特征很多时可在筛选框输入 ECFP4/MACCS 等前缀，再点「全选」批量选择。
5. 点「训练与交叉验证」，在状态栏查看指标。
6. 按需做模型评估、适用域、特征选择、可解释性。
7. 保存模型 / 导出报告 / 预测新分子。

## 9.3 方法

### 回归（11 种）

| 方法 | 主要参数 |
| --- | --- |
| Ridge 岭回归 | α（默认 1.0） |
| PLS 偏最小二乘 | 分量数（默认 3） |
| KernelRidge 核岭回归 | α、核宽 γ（默认 0.1） |
| KNN 近邻回归 | k（默认 5） |
| GBM 梯度提升 | 树数（100）、学习率（0.05）、深度（1） |
| Lasso 套索回归 | α（默认 1.0，建议扫描） |
| ElasticNet 弹性网 | α、L1 占比 ρ（默认 0.5） |
| 随机森林回归 | 树数（100）、深度（8） |
| SVR 支持向量回归 | C（1.0）、ε（0.1）；线性核 |
| 贝叶斯岭回归 | 无界面参数 |
| MLP 神经网络回归 | 隐藏单元（64）、迭代轮数（300）、学习率（0.01）、L2（0.001） |

### 分类（7 种）

| 方法 | 主要参数 |
| --- | --- |
| Logistic 回归 | α（1.0）、类别权重（无 / balanced） |
| 高斯朴素贝叶斯 | 无 |
| 随机森林分类 | 树数（100）、深度（8） |
| GBM 分类 | 树数（100）、学习率（0.05）、深度（1） |
| KNN 分类 | k（5） |
| 线性 SVM | C（1.0） |
| MLP 神经网络分类 | 隐藏单元 / 迭代轮数 / 学习率 / L2 |

- 分类决策阈值：0.5 / Youden 最优 / F1 最优（在 OOF 概率上选择）。
- 「扫描超参数」与「学习曲线」仅回归可用。
- 随机种子只影响数据划分、置换、Bootstrap 等；模型内部使用固定种子。

## 9.4 分割方式

| 分割 | 说明 |
| --- | --- |
| 随机 70% / 30% | 实际比例由「测试占比」控制（默认 0.3） |
| 按 Murcko 骨架 70% / 30% | 同一骨架整组进入训练或测试，避免骨架泄漏 |
| K 折交叉验证 | 分类自动分层 |
| 重复 K 折 | 重复轮数默认 3 |
| 留一交叉验证（LOOCV） | 每折 1 个测试样本，n 大时很慢 |
| 嵌套交叉验证 | 外层评估 + 内层网格选参，避免调参乐观偏差 |

## 9.5 预处理

| 选项 | 默认 | 说明 |
| --- | --- | --- |
| 缺失填充 | mean | mean / median / mode / zero / keep |
| 低方差阈值 | 0（关） | 剔除方差过低的特征 |
| 去冗余阈值 | 0（关） | \|Pearson ρ\| ≥ 阈值时保留先出现的列 |

## 9.6 结果与指标

- **回归**：$R^2$、RMSE、MAE、Pearson r、Spearman ρ、RMSLE、MAPE；逐折与 OOF；过拟合差（训练 − OOF）。
- **分类**：ACC、平衡准确率、P/R/F1、MCC、AUC、PR-AUC、Brier、混淆矩阵；阈值可选 0.5 / Youden / F1。
- 可绘制预测 vs 实际（分类为 ROC）、残差诊断、校准曲线、学习曲线。

## 9.7 适用域与不确定度（回归）

- **适用域**：按到训练集的距离判断域内/域外，写回 `AD_{目标}_域内`、`AD_{目标}_距离`，并绘制距离分布。
- **Williams 图**：杠杆 h 与学生化残差；$h^* = 3(p+1)/n$，|残差| 警戒线 3。
- **预测区间**：保形（OOF 残差分位）或 Bootstrap；写回 `PI_{目标}_Lo`、`PI_{目标}_Hi`。

## 9.8 特征选择与可解释性

- **VIF**：方差膨胀因子，>10 提示严重共线。
- **RFE / 前向选择 / 重要性 Top-K**：按 OOF $R^2$（回归）或 OOF AUC（分类）选择特征，并勾选到特征列表。
- **PCA**：碎石图 + PC1–PC2 得分。
- **PDP/ICE**：部分依赖与个体条件期望。
- **近似 Shapley**：逐分子 × 特征贡献热图。
- **解读单分子**：均值扰动法，输出「特征 / 值 / 贡献 Δ」。

### 9.9 保存、加载与预测

- **保存模型**：JSON，含特征列表、预处理口径、训练填充值、模型参数。
- **加载模型**：恢复目标列与预处理；预测时沿用训练口径。
- **预测新分子**：粘贴 SMILES（每行一个），可选填真实值做外部验证；结果显示预测值/正类概率、域内外。
- 可导出预测 CSV 或把结果导入数据集：
  - 回归写回 `Prediction`、`AD_域内`；
  - 分类写回 `Pred_Prob`、`Pred_Class`。
- **导出模型报告**：Markdown，含配置、分割、逐折/汇总指标、过拟合差、适用域。

### 9.10 注意事项

- 不勾选特征 = 使用全部 35 个描述符，小数据集容易样本不足。
- 分类目标必须严格为 0/1。
- 训练会整行剔除目标或特征缺失的分子。
- 适用域/Williams/预测区间/特征重要性/解释都会按当前界面方法与参数重新拟合，不使用已加载模型。
- 保存模型、导出报告、导出预测 CSV 需要商业授权。

## 10. 导出与报告

| 导出 | 入口 | 说明 |
| --- | --- | --- |
| CSV | 文件 → 导出 CSV | 对话框可选全部/可见/选中行，并勾选要写入的列（#、Name、SMILES、属性） |
| SDF | 文件 → 导出 SDF | 对话框可选全部/可见/选中行与属性字段；始终写入分子名 _Name |
| 结构图网格 PNG | 文件 → 导出结构图网格 | 每行 4/8 个，可标注名称，标准或 2× 高清 |
| HTML 报告 | 文件 → 导出 HTML 报告（Ctrl+Shift+E） | 数据概览 + 每列统计 + 当前绘图（内嵌 PNG） |
| 分析结果 CSV | 各面板/对话框「导出」 | 需商业授权 |
| 绘图图片/数据 | 绘图面板 | PNG/SVG/PDF、CSV |
| 模型 JSON / 报告 / 预测 CSV | QSAR 面板 | 需商业授权 |

> 试用版与免费版**不能导出**，包括复制 SMILES/InChI/MOL。

# 11. 三种授权档位的异同

## 11.1 档位对比

| 项目 | 免费版（free） | 试用版（trial） | 商业版（commercial） |
|---|---|---|---|
| 授权性质 | 非商业用途 | 7 天全功能评估 | 商业授权（可永久或限期） |
| 获取方式 | 默认 | 帮助 → License 激活 → 开始 7 天试用 | 粘贴授权码 |
| 有效期 | 不限 | 7 天（每台机器一次） | 由授权码决定（可永久或指定到期日） |
| 分子数上限 | 100 | 不限 | 不限 |
| 导出（CSV/SDF/图片/报告/模型/剪贴板） | ✗ | ✗ | ✓ |
| MMP / MMP Δ | ✗ | ✓ | ✓ |
| 3D 构象与形状 | ✗ | ✓ | ✓ |
| 聚类 / SOM / 相似性网络 | ✗ | ✓ | ✓ |
| 骨架 / 侧链 / R 基团 / 骨架覆盖率 | ✗ | ✓ | ✓ |
| QSAR 建模与预测 | ✗ | ✓ | ✓ |
| 虚拟库枚举 | ✗ | ✓ | ✓ |
| 基础表格、描述符、规则、相似性、重复、活性、悬崖、SAR、大部分绘图 | ✓ | ✓ | ✓ |

## 11.2 相同点

- 三者共用同一套界面、数据模型、分析与绘图功能（只是入口被门控）。
- 都绑定本机机器码，授权码为 RSA 签名，无法伪造。
- 都支持主题切换、中文/英文、HiDPI。
- 都可以查看机器码、激活状态与到期信息。
- 都采用相同的本地授权校验机制。

## 11.3 免费版

- 默认档位：未安装授权、试用已耗尽时自动进入。
- 功能受限，最多 100 个分子；导入/合并超限时自动截断并提示。
- 适合个人学习、非商业评估。

## 11.4 试用版

- 7 天全功能评估，**唯独不能导出**（避免评估结果直接流出）。
- 每台机器只能开始一次；记录带签名并写多处副本。
- 试用记录含「最近一次看到」的高水位日期，回拨系统时间不会延长试用。

- 试用结束后自动回到免费版；安装正式授权会清除试用记录。

### 11.5 商业版

- 全部功能解锁，包括导出与剪贴板复制。
- 授权码可设置到期日；到期后自动降级。
- 授权到期校验同样使用高水位，回拨系统时间不会延长授权。
- 可在对话框中「移除本机授权」，删除后回到免费/试用状态。

### 11.6 激活、机器码与移除

1. **帮助 → License 激活…**
2. 查看/复制本机机器码（基于 Windows 系统标识生成，重装系统后会变化）。
3. 把机器码提供给授权方，获取授权码。
4. 粘贴授权码 →「激活授权码」。
5. 需要换机/续期时，可「移除本机授权」。

- 旧算法机器码签发的授权仍可验证（兼容）。
- 机器码变化（如重装系统）后，原授权会失效，需要重新签发。

### 11.7 功能门控的表现

- 受限菜单项置灰，悬停显示「需要商业授权」。
- 面板/对话框中的导出、复制按钮点击时弹统一授权提示。
- 通用绘图对话框中的 MMP Δ、骨架覆盖率也会被拦截。
- 免费版超过 100 个分子时截断并提示。

## 12. 主题、语言与高分辨率

### 12.1 主题

- 视图 → 主题 → 浅色 / 深色 / 跟随系统。
- 切换立即生效；2D 结构图、绘图背景、停靠面板同步换色。

### 12.2 语言

- 视图 → 语言 → 中文 / English。
- 切换后立即生效；数据列名、SMILES、CSV 内容不翻译。

### 12.3 HiDPI

- 界面随系统缩放（125%–300%）自动适配。
- 图标与结构缩略图按屏幕像素密度渲染。
- 窗口拖到不同缩放比例的显示器时自动重绘；超出屏幕的窗口自动收缩。
- 绘图导出 DPI 默认随屏幕；可用 `QT_SCALE_FACTOR=1.5` 强制固定缩放。

# 13. 日志、故障排查与常见问题

## 13.1 日志位置

- 运行日志：`%LOCALAPPDATA%\MolInsight\molinsight.log`
- 崩溃记录：`%LOCALAPPDATA%\MolInsight\crash.log`
- 日志自动轮转，最多 4 份。

## 13.2 常见问题

| 现象 | 处理 |
| --- | --- |
| 启动即崩溃 / 无法打开 | 确认是 64 位 Windows 10/11；重新复制完整程序文件夹（文件夹版需保留 `_internal`）；查看 `%LOCALAPPDATA%\MolInsight\crash.log` 并联系作者。 |
| CSV 导入后全部失败 | 确认第一行是表头、存在 `SMILES` 列、分隔符与扩展名匹配（`.csv` 逗号，`.tsv` / `.txt` 制表符）。 |
| 分子数被截断 | 免费版上限 100；升级或先筛选文件。 |
| 导出按钮不可用 | 免费版/试用版不能导出；检查「帮助 → License 激活」中的档位。 |
| 绘图默认 X 轴是 `#` | 手动改成真实数值列。 |
| 骨架活性图按行号计算 | 活性下拉默认可能是 `#`，请改选真实活性列。 |
| 分析结果为空 | 检查是否有足够有效数值、是否被当前筛选隐藏、是否缺少必要列（如 Cluster、R 列）。 |
| 语言切换没生效 | 确认在 视图 → 语言 中选择了目标语言；界面会立即刷新，无需重启。 |
| 授权无效 | 确认机器码与授权码匹配、未过期；可用「移除本机授权」后重新激活。 |

# 14. 快捷键

| 快捷键 | 功能 |
| --- | --- |
| Ctrl+O | 打开分子文件 |
| Ctrl+P | 计算性质 |
| Ctrl+M | MMP 分析 |
| Ctrl+S | 骨架 / 侧链分析 |
| Ctrl+L | 给选中行设置标签 |
| Ctrl+Alt+C | 计算列管理 |
| Ctrl+Alt+R | SAR 取代基矩阵 |
| Ctrl+Alt+V | 虚拟库枚举 |
| Ctrl+Alt+F | 公式计算列 |
| Ctrl+Shift+C | 3D 构象与形状分析 |
| Ctrl+Shift+E | 导出 HTML 报告 |
| Ctrl+C | 复制当前绘图 |

# 15. 附录

## 15.1 CSV 示例

```
SMILES,Name,IC50_uM,LogP
CCO,ethanol,1200,0.1
c1ccccc1,benzene,50,2.1
```

## 15.2 主要写回列

| 来源 | 列名 |
| --- | --- |
| 描述符 | 描述符原名（MolecularWeight、MolLogP…） |
| 分子式 | `MolecularFormula` 、 `elem_C` … |
| 标准化 | `OriginalSMILES` （可选） |
| 活性分类 | `Active` |
| 聚类 | `Cluster` |
| SOM | `SOM_Cluster` 、 `SOM_Row` 、 `SOM_Col` 、 `SOM_Residual` |
| 网络 | `Network_Component` |
| R 基团 | `R@...` |
| 构象 | `ShapeSim_{索引}` |
| 建模回归 | `OOF_...` 、 `AD_{目标}_域内` 、 `AD_{目标}_距离` 、 `PI_{目标}_Lo/Hi` |
| 建模分类 | `CLF_P_...` 、 `CLF_P_..._class` |
| 预测导入 | `Prediction` 、 `AD_域内` 、 `Pred_Prob` 、 `Pred_Class` |
| 划分 | `Split_30%` 、 `CV_5折` 、 `LOOCV` 、 `Nested_5折外` |

## 15.3 术语

| 术语 | 含义 |
| --- | --- |
| SAR | 构效关系 |
| QSAR | 定量构效关系 |
| MMP | 匹配分子对 |
| MCS | 最大公共子结构 |
| Murcko 骨架 | 去掉侧链后的环系+连接子 |
| AD | 适用域 |
| OOF | 交叉验证的折外预测 |
| EF / BEDROC | 富集因子 / 早期富集指标 |
| Tanimoto | 指纹相似度 |
| ECFP | 扩展连接指纹 |

## 15.4 第三方组件

见 `docs/third-party-licenses.txt` 。