项目说明文档：ChEMBL 数据预处理与入库工具

一、项目目标

本项目致力于从 ChEMBL 数据库自动下载并解析原始数据，精准提取药物（化合物）与靶点（蛋白）相关信息，将结构化后的数据存入 MongoDB 数据库。以此为药物靶点相互作用预测（DTI）任务提供高效、便捷的使用与查询数据基础。

二、项目结构与主要功能

项目依托三个核心 Python 脚本文件，各司其职完成数据处理全流程：

（一）download.py

功能：

1. 自动从 ChEMBL 官网下载`.dmp`格式的 MySQL 数据库备份文件；
2. 对`.tar.gz`格式的压缩文件进行解压处理，为后续数据加载做好准备。

主要思路：

借助 Python 的`requests`库实现数据下载，`tarfile`库完成解压操作，确保数据更新操作具备良好的可重复性，能及时获取最新数据。

（二）parser.py

功能：

1. 建立与本地 MySQL 数据库的连接，读取 ChEMBL 数据表；
2. 提取药物（化合物）信息，包含结构式、SMILES 表达式等关键数据；
3. 提取靶点信息，涵盖蛋白质名称、UniProt ID、作用机制等重要内容；
4. 整合药物与靶点数据，构建药物 - 靶点相互作用（DTI）数据集。

主要思路：

通过`pymysql`等数据库连接库连接本地 MySQL，运用 SQL 查询语句对关键数据表进行连接与筛选，将原始表数据转换为统一格式的 JSON 字典或 Python 数据结构，便于后续向 MongoDB 写入数据。

（三）ToMongo.py

功能：

将 parser.py 处理后的药物、靶点及相互作用数据，写入 MongoDB 数据库。

主要思路：

使用 pymongo 连接本地 MongoDB，分别创建药物集合（compounds）、靶点集合（targets）和相互作用集合（interactions），并将对应数据写入其中。同时注重数据结构一致性和索引优化，提升后续数据查询与建模效率。

三、总结

本项目达成以下目标：

1. 实现 ChEMBL 数据的自动化获取与更新；
2. 完成数据清洗与处理，提取适合 DTI 任务的数据结构；
3. 为机器学习模型构建结构清晰、查询便捷的 MongoDB 数据源，为药物靶点相互作用预测系统筑牢高质量数据根基。