**Цель:** Продемонстрировать навыки работы с химическими библиотеками, SAR-анализом, фильтрацией и структурной трансформацией соединений.

**Контекст:** вы готовите внутреннюю библиотеку соединений к виртуальному скринингу, оцениваете её качество и помогаете выявить перспективные scaffold’ы для дальнейшего синтеза.

### **Исходные данные:**

Файл input\_library.smi (формат SMILES + идентификатор), содержащий ~3000 молекул, часть из которых:

* содержит соли/гидраты,
* не соответствует drug-like правилам,
* имеет нежелательные подструктуры (PAINS, REOS и др.),
* включает повторяющиеся или близкородственные соединения.

### **Необходимо выполнить:**

### **1. Стандартизация и очистка библиотеки**

* Привести все структуры к canonical SMILES (RDKit),
* Удалить соли, гидраты, протомеры,
* Объединить дубликаты и определить ближайшие аналоги по Tanimoto-сходству (ECFP4),
* Отфильтровать соединения по:
  + правилам Lipinski, Veber, Ghose,
  + PAINS и REOS фильтрам (можно использовать готовые подструктурные SMARTS).

### **2. Вычисление дескрипторов и кластеризация**

* Рассчитать дескрипторы: MW, TPSA, logP, HBA, HBD, rotatable bonds, QED и др.,
* Выполнить кластеризацию библиотеки по структурному сходству (например, Butina или K-means на эмбеддингах),
* Определить и визуализировать основные кластеры, рассчитать их разнообразие (intra-cluster similarity).

### **3. SAR-анализ**

* Для заданного поднабора молекул с известной биологической активностью (будет выдан .csv с активностью):
  + построить matched molecular pairs (MMP),
  + выделить ключевые трансформации, улучшающие активность,
  + визуализировать примеры (например, “R-group analysis”).

### **4. Подготовка hit-листа**

* Из отфильтрованной библиотеки выбрать **25 наиболее перспективных молекул** по совокупности:
  + drug-likeness (QED, rule-based фильтры),
  + scaffold novelty (Bemis–Murcko),
  + структурное разнообразие,
* Сформировать отчет по каждой: SMILES, дескрипторы, принадлежность кластеру, краткий комментарий.

### **Формат сдачи:**

* Скрипт на Python (желательно Jupyter Notebook, можно с RDKit + Pandas/Plotly/Matplotlib),
* Итоговый отчёт (PDF или Markdown) с графиками, гистограммами, таблицами,
* Отдельный файл hitlist.csv с выбранными молекулами.

### **Дополнительно:**

* Реализация Workflow в KNIME или Pipeline Pilot,
* Визуализация ключевых трансформаций (до/после) с комментариями,
* Использование UMAP/t-SNE для проекции библиотеки на 2D-пространство.