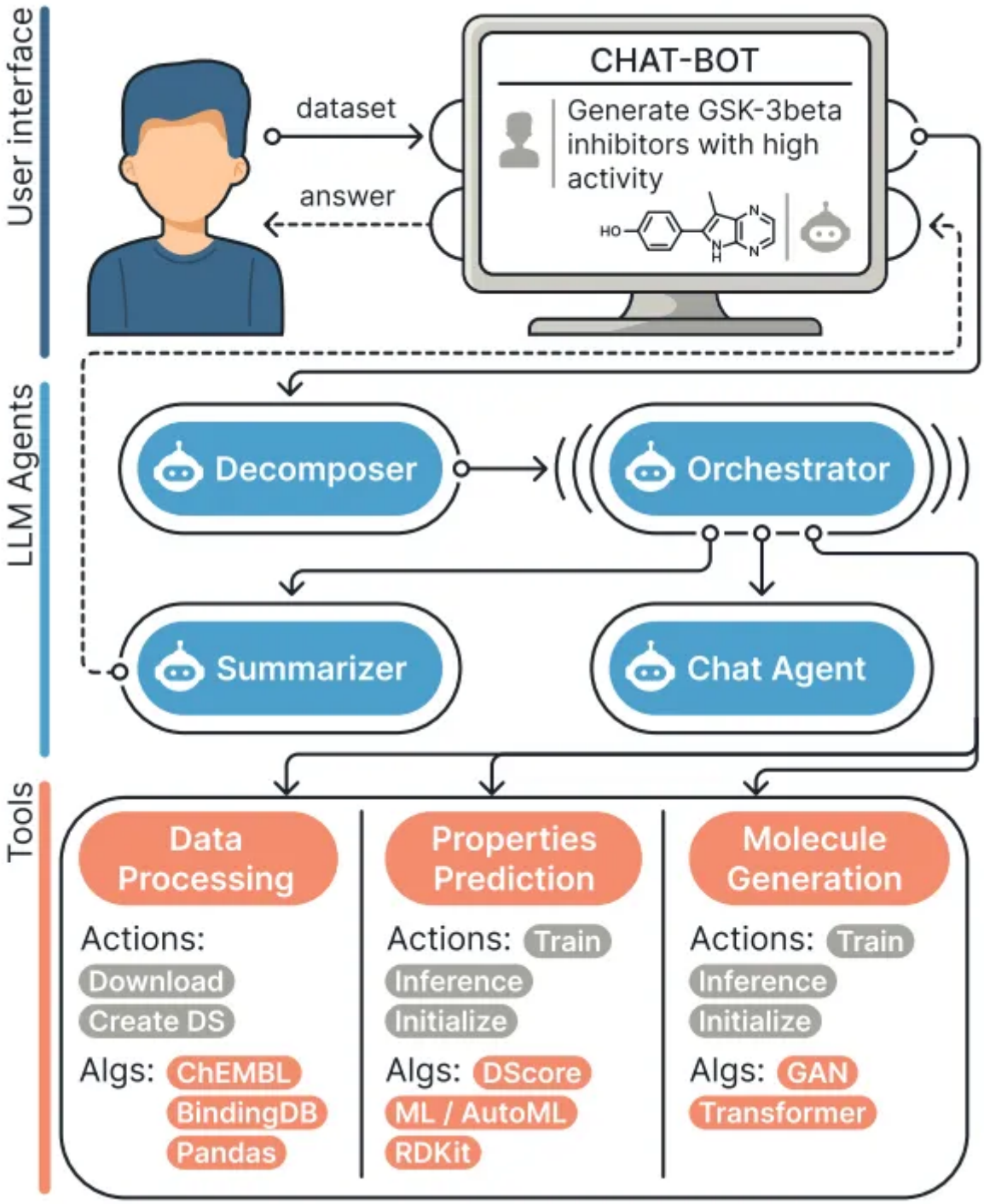
# AIDD-1




## 如何构建 LBP 候选药物打分模型（详细实操）

下面给出**候选评分的特征集、权重建议、模型结构与示例评分公式**，以及**验证指标与决策门槛**。

#### 4.4.1 候选评分的核心特征类别（输入特征）

分为 6 大类——每类下举例具体特征（用于可复现实现）：

A. 生物学相关性（Biological relevance）

- `differential_abundance_score`：在病例 vs 对照中显著富集的 effect size（log2 fold change）
- `prevalence_case`：在病例组中出现率（%）
- `prevalence_control`：在对照组出现率
- `co_occurrence_with_pathogen`：负相关系数（越负越可能抑制病原）
- `association_with_outcome`：与关键结局（exacerbation reduction）统计学关联 p-value -> 转为 score

B. 功能潜力（Functional potential）

- `SCFA_production_potential`（估算）
- `anti_inflammatory_pathway_score`（如IL-10 上调相关通路）
- `biofilm_degradation_enzymes_presence`（与噬菌体/lysins合作）

C. 安全性（必须硬性过筛项）

- `AMR_gene_count`
- `virulence_gene_presence`（boolean）
- `toxin_genes_presence`（boolean）
- `human_pathogenicity_score`（基于宿主致病基因集合）

> 任何 `virulence_genes_presence == True` 或 `AMR_gene_count > threshold` 的候选先行剔除或标为高风险。

D. 可培养性 / 工艺可行性（Manufacturability）

- `isolate_recovered_flag`（1/0）
- `growth_rate`（hours doubling）
- `aerobe_facultative_flag`（更易培养得分高）
- `freeze_dry_survival_rate`（实验或预测值）

E. 可递送性（Formulation / aerosol）

- `particle_aerodynamics_compatibility`（MMAD 模拟评分）
- `mucus_adherence`（实验或预测评分）

F. IP / Novelty / Freedom-to-operate

- `taxon_patented_flag`（是否被他人专利覆盖）
- `novelty_score`（是否未被广泛用于 LBP）

---

#### 4.4.2 评分/模型结构（示例）

我建议把评分分为两层：**硬性规则（rule gate）+ 可学习打分（ML model）**。

**硬性规则（Rule gate）**

- 如果 `virulence_genes_presence == True` → REJECT
- If `AMR_gene_count >= 2 (clinically relevant genes)` → REJECT or MANUAL REVIEW
- If `isolate_recovered_flag == 0` & `predicted_cultivability_score < 0.3` → LOW_PRIORITY

（规则旨在快速去除明显不合格项）

ML Ranker（示例 XGBoost）

输入：所有 numeric 特征（A–F）。输出：`score ∈ [0,1]`。

损失/目标可以是 **pairwise ranking**（学习把已知有效的候选排在前面）。如果没有大量监督 label，可以用 semi-supervised 或 weak supervision（比如把体外验证阳性定为正样本，negative 为体外无效）。

**打分公式（示例可解释混合公式）**

最终综合分 `S` = sigmoid( w_bio * BioRel + w_func * Func + w_safe * Safe + w_manuf * Manuf + w_deliver * Deliver + w_ip * IP )

其中：

- `BioRel = normalized(differential_abundance_score *`
- `Func = normalized(SCFA_production_potential + anti_inflammatory_pathway_score)`
- `Safe = 1 - normalized(AMR_gene_count_normalized + virulence_flag*10)`（higher is safer）
- `Manuf = normalized(growth_rate_score + freeze_dry_survival)`
- 权重建议（初始）：`w_bio=0.30, w_func=0.20, w_safe=0.25, w_manuf=0.15, w_deliver=0.08, w_ip=0.02`（可通过模型调优改变）

> 说明：上是 interpretable baseline。并行训练 XGBoost ranking model 来学习更复杂的非线性组合。

#### 4.4.3 训练策略与缺乏标签时的替代方案

- **监督学习**：如果你有历史候选的体外验证结果（positive/negative），用这些作为标注直接训练模型（优先）。
- **弱监督 / 半监督**：用因果/统计显著性（例如 p<0.01 的菌群）作为弱标签，或用 expert-provided rankings（专家打分的 Top/Bottom）训练。
- **启发式/多目标优化**：把问题当作多目标优化（maximize biological relevance & manufacturability & safety）并使用 Pareto-front identification 来选择候选组。
- **Active Learning Loop**：每轮选 Top N 进入 wet-lab，拿回 validation 结果后用这些真实 label 重新训练模型（可显著提高效率）。

---

#### 4.4.4 模型评价指标（实际度量）

- **Precision@k**（Top-10 中多少真阳性） — 最关键（实际研发成本昂贵，Top-k 准确率高价值大）
- **Recall**（对已知有效候选召回率）
- **ROC-AUC / PR-AUC**（若有二分类标签）
- **Enrichment Factor (EF)**：与随机抽样相比，你的 Top-10 比例提升倍数
- **Time-to-Proof（wet-lab）**：从候选导出到实验验证成功所需平均时间（业务 KPI）

目标示例：`Precision@10 ≥ 40%` 在早期即可显著节约成本。

---

#### 4.4.5 可解释性与决策支持（必须）

- 对每个候选输出：SHAP feature breakdown（哪三个特征最驱动评分） + Rule-gate flags（AMR/virulence/uncultivable） + 所属样本/院区的 prevalence statistics。
- 把解释结果做成决策卡（candidate card），便于科学家/CMC 做快速判断。

---

#### 4.4.6 模型部署与迭代（从 MVP 到生产）

- 把训练好的 model 存入 Model Registry（带 feature_version 与 training_data_version）。
- 在 Feature Store 做在线/离线特征服务（便于实时评分新样本）。
- 每轮 wet-lab validation 完成后触发自动 retrain pipeline（自动化：ToolUniverse CI/CD），并生成 new candidate list。

---

## 4.5 质量控制、统计注意事项与偏差修正

- **批次效应 (batch effects)**：必须在特征级进行批次校正（ComBat）或通过模型中加入 batch covariates（hospital_id, run_id）。
- **宿主读数高造成假阴性**：若 `host_read_fraction > threshold` 标记样本并尽可能用高深度测序或重新抽样。
- **因果混淆**：菌群-疾病关联可能受抗生素使用、住院时间等 confounders影响；模型训练时把这些临床变量作为 covariates 或用因果推断方法（DoWhy, propensity score matching）减少偏差。
- **外部验证**：务必保留独立医院/地域的数据作为外部验证集。