

面向抗糖尿病候选分子的 DPP4 效力与 DPP8 及 DPP9 选择性联合预测研究构想

2026 年 10 月 26 日 北京时间 研究构想与公开数据原型验证

一页要点

当前原型未通过外推验证，不用于候选优选或临床决策。 我建议将本方向定位为公开生物活性数据驱动的效力与选择性预测方法研究。已完成数据治理、训练前协议、固定外推拆分和小型模型比较，但未建立联合模型优于独立随机森林的可靠证据。下一步先补齐配对标签与优化收敛，再决定是否扩大模型或进入实验验证。

研究依据。 iPADD 体现了团队开展抗糖尿病候选分子计算预测的研究基础。[1] 本构想收窄到 DPP4、DPP8、DPP9 的体外 IC₅₀ 与选择性差值，不推定该 DPP 任务就是导师现有课题。已有三酶 3D-QSAR 及 DPP4 机器学习筛选研究，因此不以首次研究三靶点或使用机器学习作为创新。[2-3]

已有实质工作。 ChEMBL 37 初核 3883 条测量，经原文单位、数值关系和借用参考值审计，在拟合前排除 151 条，保留 3732 条。结构、缺失、删失、来源及分区均可追溯。用于探索性配对约束的 77 条标签已核对原文面板和数值，但方法信息未完整；严格完整方法匹配配对仍为 0。

原型告诉我们的事。 在预定 seed 下，骨架外活性宏平均 RMSE 为 RF 0.681，三个多任务变体为 0.802、0.874、0.873，未显示改善。神经网络均达到 300 次迭代上限而未收敛，限制了方法比较的解释。骨架外选择性主终点因缺 DPP9 精确配对不可计算；来源外只有 2 个分子的 4 条精确配对，只能描述，不能据此宣称验证成功。

后续要检验的问题。 在更充分且可比的标签与独立测试支持下，保留删失边界并利用配对约束，能否降低 ΔpIC_{50} 预测误差？主要终点仍为两脱靶任务的精确 ΔMAE 宏平均；排序、10 倍筛查和单靶点误差另行报告。当前输出只涉及体外测定效力与表观选择性，不推断临床疗效或安全。

我拟承担的工作。 结合临床药学背景，主导测量语义、来源审校、数据提取与治理、脚本、基线与消融、错误分析和复现材料。希望在导师指导下完善建模与验证，而不以未经证明的经历或现成实验资源作保证。

希望讨论的决定。 是否先以高质量配对证据和可复现外推评价为首期主题；如何设定新的配对覆盖与收敛闸门；达到这些条件后再决定是否增加模型复杂度。湿实验另列合作、预算与测定方案，不作为已具备的条件。

研究定位与预期增量

本研究关注候选分子对 DPP4 的体外抑制效力，以及相对于 DPP8、DPP9 的测定选择性。一个分子在 DPP4 上活性较强，并不自动说明对另外两个同族酶的作用较弱，因此三靶点结果应作为相互关联但不同的预测任务。若经后续独立验证，模型输出才可用于计算筛选与实验优先级讨论，不能直接推断降糖疗效、患者获益、临床毒性或用药安全。

iPADD 利用分子指纹和机器学习识别潜在抗糖尿病候选分子，为本构想提供了相关计算研究背景。[1] 我希望把候选分子分类进一步落实到有明确量纲和测定条件的生化终点，并研究什么样的数据和评价足以支持“选择性较好”这一结论。这是我提出的潜在合作方向，不表示团队已确定开展这一具体任务，也不替代对导师当前研究重点的进一步沟通。

三靶点选择性建模已有先例。2007 年的工作进行了 DPP4、DPP8 和 DPP9 抑制剂的对接辅助 3D-QSAR；2021 年的工作比较了 DPP4 活性预测模型，并开展虚拟筛选和选择性相关分析。[2-3] 因而，拟研究的增量不是首次联用三个靶点，而是检验在严格来源治理、删失边界保留和外推评价下，多任务与可靠配对监督是否仍然带来可重复的收益。该增量须由直接比较与失败分析证明，不能在训练前宣称创新已经成立。

本轮已经完成的原型与实际结果

数据基础及原文审计

本轮使用 ChEMBL 37 人源、直接单蛋白关系、confidence score 9、binding 单蛋白格式的 IC50 nM 记录。初核 DPP4、DPP8、DPP9 分别为 2601、721、561 行，合计 3883 行。三靶点同来源候选覆盖 331 个母体、32 个来源 ID，即 30 篇论文与 2 项专利；这不等于 32 项独立研究。标准化后得到 2533 个保留立体信息的结构和 1044 个用于防泄漏的骨架关联组；它们与初核三靶点候选的 125 个探索性骨架组口径不同。[4-5]

拟合前排除两项单位待解决专利来源的 137 行、原文关系或界限冲突的 7 行，以及借用参考值的 7 行，最终保留 3732 行。借用参考值中只有 1 条同时发现 1 nM 与 0.9 nM 不一致，其余 6 条是来源溯源问题，不能一概称为数值错误。原始记录不改写，分区不因剔除重抽。

最终有效配对门控为 77 条，其中 23 条精确差值、54 条区间差值，来自 Kim2007 和 Espadinha2024 两个原文面板。[6-7] 这些配对的原值、单位和关系已核验，但方法细节未完整获得，属于“原文面板核验但方法信息不完整”，不是严格同条件数据。严格完整方法匹配配对为 0；未知、不兼容、单位风险及数值冲突记录均未用于配对约束。

固定预算下的模型比较

已按训练前 v1 及 A1、A2 补充协议完成独立 RF 和三个小型多任务变体的比较。下表活性 RMSE 为三个靶点等权宏平均； Δ MAE 为 DPP4 对 DPP8 和 DPP9 的等权宏平均，数值越小越好。网络主表固定采用 seed 20261025，不是择优 seed 或集成结果。两个指标含义不同，不能横向混作同一终点。

模型	骨架外活性 RMSE	来源外活性 RMSE	来源外表现 Δ MAE
独立 RF	0.681	1.279	2.674
M1 普通多任务	0.802	1.562	3.162
M2 删失多任务	0.874	1.758	3.153
M3 面板配对约束	0.873	1.814	3.402

骨架外选择性主终点在所有模型中均不可计算：合格精确配对只有 DPP4 对 DPP8 的 2 条，DPP4 对 DPP9 为 0 条。来源外 Δ MAE 虽可计算，但仅依赖同一来源的 2 个分子、每分子 2 个靶点对，共 4 条，并非 4 个独立样本；它只展示管线如何计算，不足以支持跨来源泛化结论。所有配对均为上述部分方法核验层级。

来源留出面板另有 11 个分子的 22 条可判定精确或区间配对。在这个面板内，RF 预测达到 10 倍 DPP4 选择性的 20 条全部错误，其中 DPP8 为 9/9、DPP9 为 11/11；M3 为 22/22。分母是模型预测为选择性的可判定记录，不是所有化学空间的错误率。该来源以 DPP8/9 导向探针为主题，结果提示应检查跨来源化合物分布变化，但不能仅凭这一面板确定失效机制。当前原型不应作为候选分子优选工具。

骨架外活性误差没有支持联合模型改善。来源外及极小配对表中的数值也不能用来抵消标签覆盖不足或证明模型优势。三个网络均达到 300 次迭代上限而未满足收敛条件。独立复核显示主拆分检查点的目标与梯度有限、训练损失下降，达到迭代上限不同于数值崩溃；但收敛不足仍限制方法比较。本轮结果只说明预设实现与预算下的表现，不能据此否定多任务或删失方法本身。

三个预定 seed 的活性宏平均 RMSE 范围为：M1 骨架外 0.801 至 0.802、来源外 1.562 至 1.571；M2 骨架外 0.869 至 0.898、来源外 1.745 至 1.758；M3 骨架外 0.873 至 0.894、来源外 1.721 至 1.814。完整逐靶点结果、删失界限违反、逐行预测、失败案例和优化器状态均已保留，不挑选表现最好的运行。小样本配对的置信区间和排序指标不能提供可靠确认，部分指标因此不可计算。

固定比较完成后，另对 6 个预定 seed 检查点做了仅训练侧的数值诊断，各追加最多 700 次迭代。训练目标进一步下降约 3.0% 至 19.6%，但 6 项均再次达到上限而未满足收敛条件；未生成新的测试预测，也未替换本表的 300 步结果。这是后验优化诊断，不是新的性能验证，既不能证明延长训练可以挽救外推，也不能据此判定架构无效。

由结果支持的下一步

目前最明确的进展是形成了可审计数据和评价管线，并识别出单位风险、错误关系符、借用参考值和配对覆盖不足。下一阶段优先补原文及明确测定条件，改进训练区内的数值优化诊断，并预先设计新的外部验证集；本轮测试结果已被查看，不能继续把它当作未接触测试集调试到满意为止。暂不以扩大网络或筛选“优异候选分子”替代上述工作。

研究目标与可证伪假设

后续第一项目标是在已有可追溯数据集上补全关键方法与来源信息。每个标签保留原始值、单位、关系符、分子与母体 ID、assay ID、来源 ID 和原文定位，另存标准化结构、精确或区间标签、质量标记及分区。重测、重复发表、跨来源相同化合物和冲突测量应分别识别，不以平均数掩盖差异。

第二项目标是在充分配对覆盖和可接受优化收敛后，重新检验共享结构表示、删失信息和配对约束的增量。主假设为：相较独立靶点随机森林，保留缺失掩码及删失边界、并利用审核合格配对的多任务模型，在骨架外测试的选择性差值误差更低，并在来源外评价中保持相同改善方向。

本轮预测的是来源面板内表观 IC₅₀ 选择性，不是内在选择性或 K_i 选择性。主要终点为审核合格的来源面板内精确配对 Δ pIC₅₀ 的 MAE，对 DPP4 与 DPP8、DPP4 与 DPP9 等权宏平均。若其中任一靶点对缺少合格测试标签，则不能计算完整主终点，只能报告可用分量并注明主假设未能检验。排序相关、方向判断、10 倍选择性筛查和单靶点误差均为分开的辅助结果。

第三项目标是说明模型何时不适用。研究将报告低结构相似度、测定来源变化、标签覆盖差和任务负迁移下的失败，而不只呈现表现最好的模型或拆分。若联合模型没有稳定改善，仍可形成对公开数据选择性建模边界的可复现结论。

数据整理与标签设计

结构和测量分别治理

结构处理保留主要有机组分与立体信息，明确去盐、标准化和中和规则；不能将结构不一致的记录仅凭相同 parent ID 合并。原始行永久保留，只在同结构、同来源、同 assay 的精确重复上生成分区内中位数视图，并保留范围、测量数和全部记录 ID。不同实验体系不跨来源平均。

本轮只研究 IC50，不与 Ki 混池，也不把细胞或临床终点转写为同一种活性。来源和实验信息用于质量审计及分组，不进入结构模型特征。未测靶点用掩码处理，不补为不活跃；未知实验条件明确保留未知。

删失信息保留为界限

pIC50 定义为 9 减去 IC50 nM 的以 10 为底对数，数值越大表示体外效力越强。IC50 大于某值时，pIC50 是上界；IC50 小于某值时，pIC50 是下界。网络只在预测违反这些界限时受到相应惩罚，不把阈值当成测量真值。

选择性定义为 $\Delta 8 = \text{pIC50}(\text{DPP4}) - \text{pIC50}(\text{DPP8})$ ， $\Delta 9$ 同理。若 DPP4 IC50 为 10 nM、DPP8 IC50 大于 1000 nM，则 $\Delta 8$ 大于 2，而非精确等于 2。所有阈值和区间判断都需经过方向与开闭边界单元测试；区间不能区分的样本不强行产生排序。

配对实行证据分级

同结构同来源只形成候选配对。原文核查需关注蛋白来源与构建体、底物及相对 Km 浓度、缓冲/pH、温度、时长、预孵育和读出方法。描述相同但缺少原文证据的配对，不升级为严格可比。

独立原文审核区分严格可比、条件明确可用、原文面板核验但方法不完整、不兼容及未知。训练前补充协议允许前三类参与探索性表现 IC50 选择性分析，且训练与评价采用同一行级门控并分层报告。当前实际可用的只有第三类，不能称为完整方法核验。未知和不兼容均不进入配对约束及主要终点，但二者应分开记录。

模型路线与公平比较

本轮已采用统一 ECFP 分子指纹和小型模型，没有引入大规模预训练或复杂图网络。下一阶段仍应先完成强基线与收敛诊断。这样可以将问题集中到标签信息、任务共享和配对约束，而不是同时改变表示、网络规模及训练数据。更复杂表示只有在基线明确、覆盖足够后另立比较。

模型	核心信息	要回答的问题
独立 RF	各靶点精确标签	单任务结构基线能达到什么水平
M1 普通多任务	共享表示与缺失掩码	任务共享是否有效或造成负迁移
M2 删失多任务	M1 加上下界监督	保留测定边界是否有增量
M3 配对约束多任务	M2 加审核合格配对	配对差值信息能否改善选择性

RF 固定 200 棵树，只比较两种叶节点最小样本数；三个网络保持同一 32 单元隐藏层、三个输出和两档 L2 正则。使用固定验证指标和相同候选配置数，网络以三个预定 seed 复跑，保存全部运行与终止状态。验证指标是靶点宏平均精确标签 RMSE，作为模型选择的固定代理指标；它不是选择性研究的主要终点。

M1 与 RF 使用同样的精确训练信息，M2 增加删失监督，M3 增加通过审核的配对约束。比较解释必须保留这种信息差异，不能将所有改善都归因于网络结构。本轮 M3 仅使用训练前批准的原文面板部分方法核验配对，未将描述相似或同文献关系自动当作资格；任何新一轮仍应先固定门控和预算。

外推评价与结果解释

主评价以母体、连接结构和普通 Murcko 骨架关联形成的分组划分训练、验证、测试，目标比例 70 比 15 比 15。分组以固定哈希确定，不因标签分布或模型表现重抽。次评价按来源族拆分，并排除跨分区重复化合物；同一研究的不同来源记录不能被视为独立外部证据。

骨架外与来源外回答不同问题。骨架拆分仍可能共享实验来源，来源拆分也可能共享更广泛骨架；两种结果分别报告，并显示相互重叠，不使用“完全外部验证”概括两者。最近训练分子的指纹相似度用于分析模型适用范围，不以随机拆分高分替代新化学空间表现。

精确标签报告各靶点 MAE、RMSE 及共同测试分母。审核合格精确配对分层报告 Δ 误差；排序至少需三个不同真值才计算相关。删失记录单独报告界限一致性和可判定覆盖率，不能与精确 RMSE 混合。 $\Delta \geq 1$ 的 10 倍选择性只是预设筛查阈值，须报告错误选择性判断和不可判定数量，不把它解释为安全阈值。

区间估计以化合物聚类重采样，避免重复测量夸大精度，同时报告来源族和骨架组数。小规模配对或少数来源的结果只作描述性证据。若不同随机 seed、靶点对或外推场景方向不一致，应将其作为下一步是否值得投入的依据，不选择性报告阳性指标。

实施阶段与资源安排

公开数据与小型原型阶段已经完成到上述边界。以下后续安排以证据和资源是否充分为准，不是对导师平台或完成日期的承诺。

阶段	工作范围	进入下一阶段的依据
已完成的数据原型	逐行审计、来源族候选、原文配对门控与冻结拆分	仍需补齐方法信息及来源独立性证据
下一轮方法诊断	在训练和验证区内检查尺度、损失权重及优化收敛，先保留 RF 强基线	另立预算和未接触测试，不修改本轮结论
扩大回顾性验证	补全文献条件与新增可匹配对，另建外部数据	主终点有覆盖，增量稳定且不只来自单一来源
后续湿实验	在独立预算与合作条件下确认候选与阴性对照	测定方案、材料、盲法与重复设计另行审定

公开数据阶段以 CPU 计算、版本化文件、常规化学信息和统计软件为主，不以 GPU 或现成实验平台为启动前提。本轮记录了实际拟合时间与环境版本；尚未完整测量峰值内存和人工核查成本，不据局部耗时承诺整体资源需求。原文合法获取与实验细节缺失可能是比算力更重要的瓶颈。

若公开资料不足以支持审核合格选择性配对，可继续形成删失标签与数据治理方法的可行性报告，但应明确选择性主假设仍未检验。若单任务 RF 已足够、联合模型负迁移或来源外结果不稳定，则优先修正数据与任务定义，不通过无限扩展模型复杂度寻找阳性结果。

未来实验需同时测定 DPP4、DPP8、DPP9 并明确可比条件，设置代表性预测高选择性分子、边界样本和预测失效对照。候选数量、材料获得方式、浓度范围和重复次数，应根据实际预算和先导实验变异确定；计算评分不替代实验确认，也不直接进入人体研究。

我的工作重点与合作讨论

我希望把临床药学背景落实到明确的测量语义、选择性解释和验证边界，同时承担数据提取、程序实现与模型分析，而不仅是整理文献。首期可由我主导数据字典和审校规范、来源与 assay 核查、结构和标签处理、复现脚本、强基线及消融、逐例错误分析和研究记录。

方法层面，希望在导师指导下讨论共享表示是否适合这一样本规模、配对约束如何避免放大测定差异，以及如何设置真正具有独立性的外推验证。后续如有合适合作条件，再共同评估体外实验路径。本方案不预设导师已具备特定酶学平台、候选材料或患者数据，也不承诺论文数量或实验成功。

本轮已经形成可审计数据及字典、冻结协议和划分、可重复运行的代码与环境、全部模型结果和失败案例。现有证据支持先调整与完善，而不是宣布联合模型已验证成功。希望先就任务价值、配对可比标准和首期边界得到指导，再据证据决定后续投入。

技术验证附录

可重复性与有效样本

协议 v1 及拟合前 A1、A2 分别保留日期和 SHA256，质量排除均在拟合前确定；原始快照只读。最终固定拆分哈希以 6e17318e 开头，原型保留 3732 条测量。来源拆分另排除跨分区共享化合物的 195 条测量，因此其 train、validation、test 合计 3537 行；这不是另一套随机重抽数据。下表为实际模型输入口径，不是初核数量。

分区	全部测量行	精确行 DPP4 / DPP8 / DPP9
骨架外 训练	2657	1481 / 235 / 151
骨架外 验证	490	318 / 24 / 11
骨架外 测试	585	331 / 82 / 35
来源外 训练	2272	1351 / 190 / 80
来源外 验证	430	240 / 67 / 71
来源外 测试	835	417 / 64 / 27

DPP9 主验证区只有 11 条精确测量，主测试只有 35 条；单靶点测试行多于合格选择性配对数，二者不能互换。来源族仍是保守候选分组，完整来源元数据及独立性尚未覆盖全部记录。

固定实现与审计检查

结构输入为 2048 位 Morgan 指纹、半径 2、保留立体化学。缺失不活跃化、IC50/Ki 混池、把不等号阈值当精确值、来源 ID 或测定结果倒灌输入，均被禁止。单位、符号、区间运算、缺失掩码、损失梯度、结构及拆分泄漏和文件重跑一致性均纳入测试；测试输出随复现包保存。

RF 为 200 棵树，叶节点候选 1 或 3。网络为 32 单元 tanh 共享隐藏层，L2 候选 0.0001 或 0.001，L-BFGS-B 最多 300 次迭代，配对权重 0.25；候选数固定，不进行架构搜索。验证区宏平均活性 RMSE 只是调参代理指标。三个网络 seed 为 20261025、20261026、20261027，未选择最好 seed 替代主报告，也未将验证区合并回训练区。

独立复核已确认主拆分 RF 和 9 个选中网络的保存后预测可逐位复现；全部 20 个最终保存模型的往返验证通过。骨架分组留出中发现 2 个测试分子与不同环尺寸的训练分子 Morgan 指纹 Tanimoto 为 1，这

是有限半径二值指纹碰撞，不是相同结构泄漏。因此骨架分组不保证所有测试分子在所用表示中遥远，不将其宣传为全面化学空间外推。

模型环境记录 Python 3.12.14、NumPy 2.3.5、SciPy 1.17.0、scikit-learn 1.8.0。所有模型、配置、逐行预测、运行耗时、终止状态和输入哈希已保存。区间估计按化合物或相应拆分分组聚类重采样；仅表示固定测试样本内的不确定性，不涵盖模型选择或新来源泛化。选择性来源测试只有一个来源，不能凭重采样制造多个独立研究。

资料使用与边界

本项目数据派生自 EMBL-EBI ChEMBL 37，保留原始 ID 和版本。按 ChEMBL 官方许可说明，数据库及其衍生数据遵循 CC BY-SA 3.0，并保留来源署名及同许可要求；论文原文并不因此获得同样再分发授权。[8] 交付复现材料保留必要出处和数据处理记录，不再分发期刊论文 PDF。模型输出限于体外测定效力与表观选择性，不构成临床疗效或毒性证据。

参考依据

[1] Liu X W, Shi T Y, Gao D, et al. iPADD: A Computational Tool for Predicting Potential Antidiabetic Drugs Using Machine Learning Algorithms. *Journal of Chemical Information and Modeling*, 2023, 63: 4960–4969. DOI 10.1021/acs.jcim.3c00564。本方案仅据其任务和方法背景建立关联，不复述其性能为本研究结果。

[2] Kang N S, Ahn J H, Kim S S, et al. Docking-based 3D-QSAR study for selectivity of DPP4, DPP8, and DPP9 inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 2007, 17: 3716–3721. DOI 10.1016/j.bmcl.2007.04.031。用于确认三酶选择性建模已有研究基础。

[3] Hermansyah O, Bustamam A, Yanuar A. Virtual screening of dipeptidyl peptidase-4 inhibitors using quantitative structure–activity relationship-based artificial intelligence and molecular docking of hit compounds. *Computational Biology and Chemistry*, 2021, 95: 107597. DOI 10.1016/j.compbiolchem.2021.107597。用于确认 DPP4 机器学习筛选及选择性相关研究并非空白。

[4] Zdrazil B, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 2024, 52: D1180–D1192. DOI 10.1093/nar/gkad1004。本项目使用 ChEMBL 37 本地快照；DPP4 ChEMBL284、DPP8 ChEMBL4657、DPP9 ChEMBL4793。

[5] 本项目训练前研究协议 v1 及 A1、A2 补充，2026 年 10 月 25 日 UTC，即北京时间 26 日；冻结数据、assay 原文审计、各拆分评价结果和运行日志。实际原型数值均来自该冻结执行记录，不是文献性能。

[6] Kim D, et al. Triazolopiperazine-amides as dipeptidyl peptidase IV inhibitors: close analogs of JANUVIA (sitagliptin phosphate). *Bioorganic & Medicinal Chemistry Letters*, 2007. DOI 10.1016/j.bmcl.2007.03.098。原文表 1、表 2 用于数值与关系逐条核验；方法细节未完整核验。

[7] Espadinha M, et al. Active site-directed probes targeting dipeptidyl peptidases 8 and 9. *European Journal of Medicinal Chemistry*, 2024, 270: 116389. DOI 10.1016/j.ejmech.2024.116389。作者稿表 1 用于混合单位、关系符和面板核验，不将后续细胞成像作为生化 IC50。

[8] EMBL-EBI ChEMBL. General Questions 中的许可说明，北京时间 2026 年 10 月 26 日核对；CC BY-SA 3.0。