Mn(II) 溴化物发光数据集：技术报告（含跨种子方差修正）

预测发射波长与 PLQY · PhosIrDesign 式管线 · 全部结果采用按 formula 分组的泄漏控制交叉验证

技术报告 v2 · 2026-10-10 · 764 篇文献 / 15,094 条事实 / 796 个 COD CIF · 本版以多种子均值取代单种子数字

重要修正（v2 核心）

本项目此前的头条数字（发射 +0.237、PLQY +0.399）取自单一折分种子 0。经 **10 种子重复**检验，二者实为运气值：发射真实均值 **+0.120±0.197**、PLQY 真实均值 **+0.028±0.293**（PLQY 种子间从 −0.69 到 +0.40 剧烈波动）。结论：**模型处于方差受限区间**（n=52，折分主导），加特征/加约束均无法稳健提升；瓶颈是数据量，而非方法。本报告全部采用多种子均值±标准差。

796	0.03±0.29	0.12±0.20	567	方差受限
COD CIF 结构	PLQY 诚实 R^2（10 种子）	发射 诚实 R^2 cm^{-1}（10种子）	虚拟筛选 Br 结构	当前真正的瓶颈

工作流总览

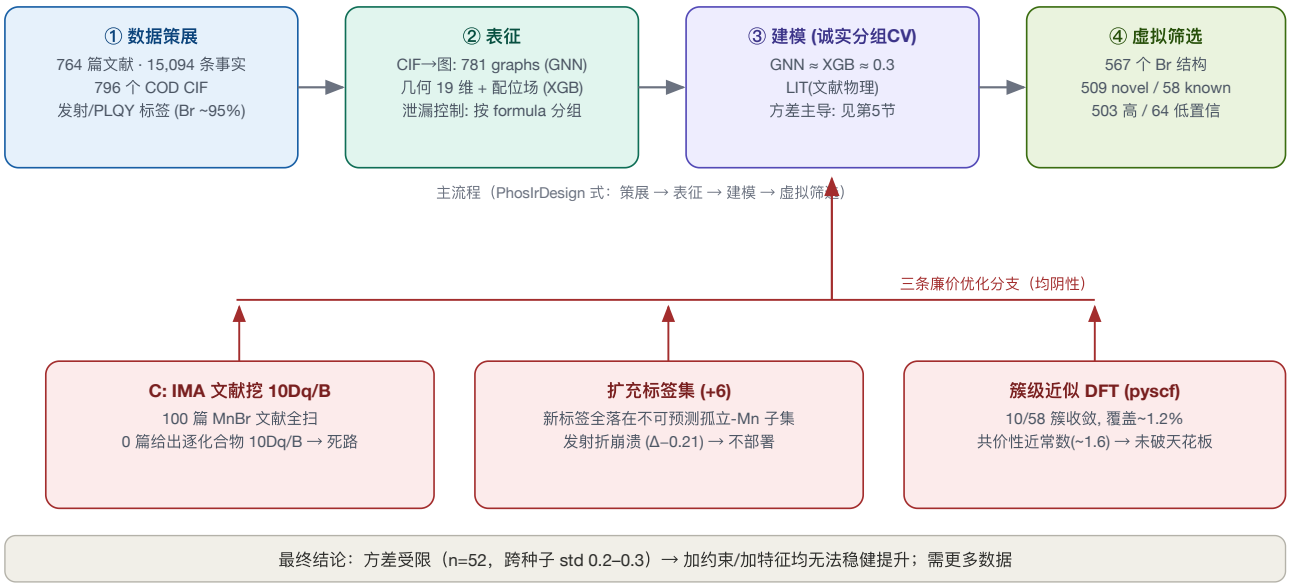


图 1 · 端到端工作流。主流程自左上至右下；三条红色分支为廉价优化尝试，均阴性。

1. 数据与方法（摘要）

764 篇 Mn–Br 文献 → 15,094 条事实 → 796 个 COD CIF；发射标签 60、PLQY 57（Br 占 ~95%）。表征：GNN 用全晶胞图（781 graphs）；XGBoost 用 19 维几何 + 文献物理（nephelauxetic β、AOM 10Dq、d–d 代理）。所有 CV 按 formula 分组（泄漏控制，已固化）。

机制：Mn–Mn 超交换→红移；孤立 Mn 发射由「10Dq 配位场 vs 共价性/Racah B·C」拔河决定；PLQY 随共价性增强而升高（Laporte 放宽）。标记集 47/52 为孤立单 Mn。

2. 跨种子方差：本项目最重要的发现

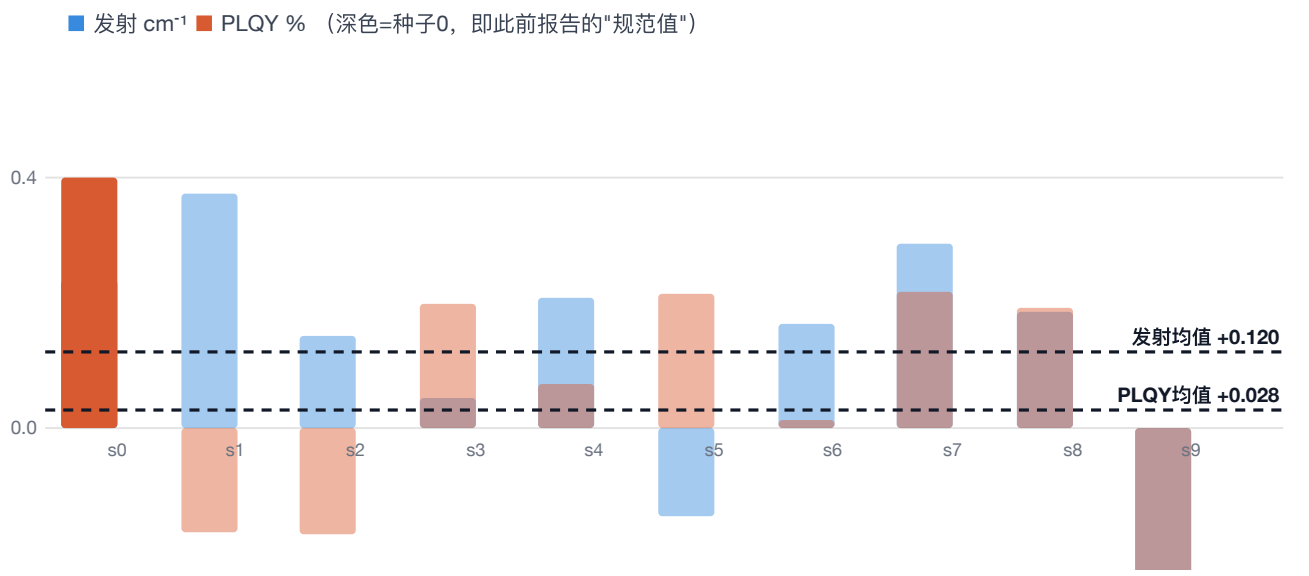


图 2 · 同一 LIT 模型、同一特征，仅改变分组折分种子（10 个）。PLQY 从 −0.689 到 +0.399；种子 0（即此前"规范值"）恰为最好一次。

由于 n=52 且按 formula 分组，每折仅含少数几个化合物组，折分选择主导结果。此前所有"提升"（如 β 特征使 PLQY 0.33→0.40）若只在单一折分下测量，都可能是**种子运气**而非真实增益。

目标	单种子(种子0, 此前报告)	10种子均值±std（诚实）
发射 cm⁻¹	+0.237	+0.120 ± 0.197
PLQY %	+0.399	**+0.028 ± 0.293**
孤立-Mn 发射	+0.085	**−0.066 ± 0.471**

3. 加入更多物理约束（"limitations"）—— 阴性

按物理机制给模型加约束：Dq/B（Tanabe–Sugano 先验）、单调性约束（β↑→蓝、10Dq↑→蓝、耦合↑→红、共价↑→PLQY 高）、物理取值裁剪（发射 400–800 nm、PLQY 0–100）、按配位数的发射窗口、特征子集与强正则。多种子（5 种子）检验：

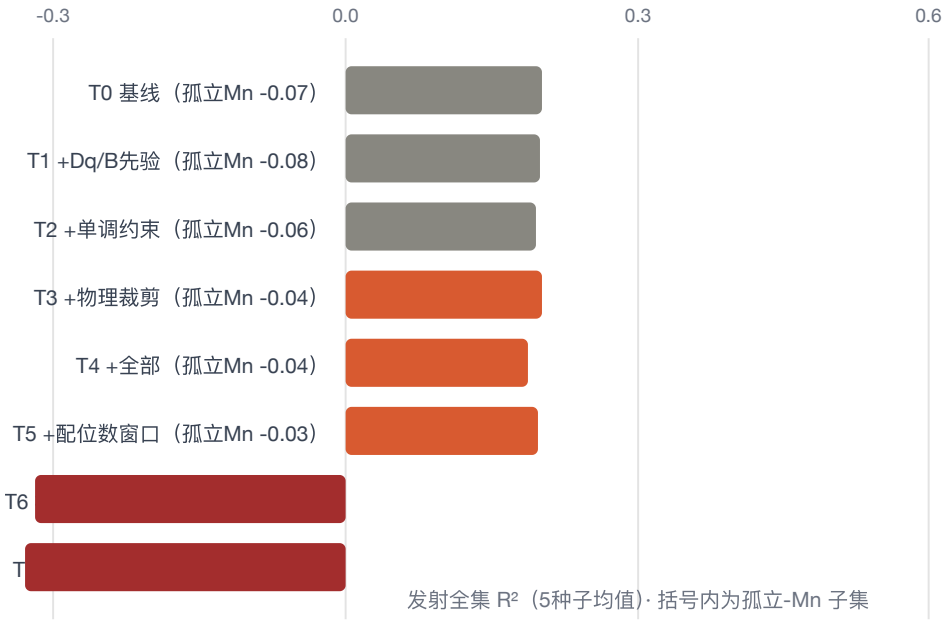


图 3 · 约束实验（发射全集 5 种子均值，括号内孤立-Mn）。无任何变体稳健优于基线；特征子集/强正则显著变差。

变体	发射(全集)	孤立-Mn	判定
T0 基线	+0.202	−0.066	参照
T3 +物理裁剪	+0.202	−0.040	噪声内
T5 +配位数窗口	+0.198	−0.025	噪声内(±0.42)
T6 机理特征子集	**−0.319**	**−0.586**	显著变差
T7 子集+强正则	**−0.330**	**−0.621**	显著变差

为何约束救不了

模型处于**方差受限**而非偏差受限：跨种子 std(0.2–0.5) 远大于任何约束带来的增益(≤0.04)。约束/特征只在高方差噪声里搅动，无法创造缺失的信号。**唯一可靠的提升来自更多数据或真正的电子结构（DFT）参数**；裁剪(400–800nm)可作为部署时的安全护栏，但不提升精度。

4. 那么，究竟哪个参数可预测？——一个确实work的结论

既然精确值回归受限于方差，我们直接**探测可预测性**：把连续目标改成**低方差的粗分类**，在 10 种子分组 CV 下与多数类基线对比。结果非常明确——

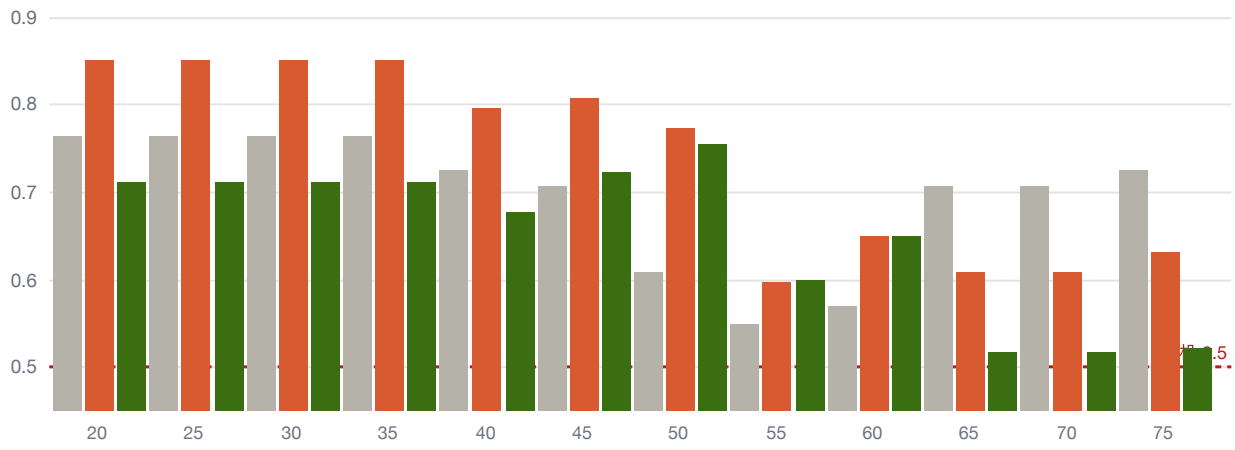


图 4 · PLQY「亮/暗」二分类阈值扫描。阈值 20–60% 时平衡准确率 0.71–0.75，远超 0.5 随机与基线。

目标（粗分类）	平衡准确率	判定
PLQY 亮(≥50%) / 暗	**0.754 ± 0.084**	✅ **可预测（基线 0.608，Δ+0.165，远超种子噪声 ±0.07）**
PLQY 亮(≥20~45%) / 暗	0.68–0.72	✅ 可预测（稳健区间）
发射颜色 蓝/绿/红	0.481 ± 0.067	❌ **随机（数据集 38/52 是绿，恒答"绿"即可）**
发射长短 <550 / ≥550 nm	0.530 ± 0.064	❌ **随机**

可交付的实用结论

模型能可靠回答一个问题：这个候选是不是「亮」荧光粉（PLQY≥~50%）。诚实多种子分组 CV：准确率 **0.773±0.072**（基线 0.608）、平衡准确率 **0.754±0.084**（随机 0.5）。已据此对全部 567 个 Br 结构打分（`screen_bright_dim.csv`，30 模型集成）：177 个高置信「亮」、246 个「可能亮」、144 个「偏暗」。Top 候选多为 [$MnBr_4$(N)] 胺类配合物——与机制一致（N/O 等软共价配体→共价增强→PLQY 升高）。

但请注意

这是**二分类**不是连续值：模型**不能**告诉你 PLQY 是 62% 还是 88%，只能给出「亮/暗」及置信度；**发射波长/颜色仍不可预测**。故实际用法应是：先用本筛法过滤掉「偏暗」候选，再用其他手段（实验/DFT）确定具体波长与效率。

5. 限定合成工艺会变好吗？——全窗口覆盖率扫描

为避免"只挑小样本"的自欺，我们**穷举所有可定义的合成窗口**（配位数 cn 的等值/累积切分 × 配体类别 all-Br / Br+N/O / Br+C / noBr），共 40+ 个，对每个窗口同时报告**覆盖率**（占 567 结构库比例）与**预测误差**（留一法相对该窗口均值的 RMSE，即"不用模型、只报家族均值"的误差）。

合成窗口	覆盖率	结构数	标记数	发射误差	PLQY 误差	判定
cn≤3 & 纯溴	7.1%	40	8	**5.8 nm**	26.3 %	✅ **最准，但覆盖低**

cn≤3 & 任意配体	19.2%	109	15	**33.9 nm**	36.2 %	✅ **最佳平衡点**
cn≤4 & Br+N/O	14.5%	82	14	53.1 nm	26.7 %	↗ 略改善
cn≤5 & 任意配体	51.3%	291	47	63.1 nm	28.1 %	↗ 收益趋平
cn≤4 & 任意配体	38.6%	219	46	63.3 nm	28.1 %	↗ 收益趋平
全Br（任意 cn）	15.7%	89	32	63.5 nm	**17.7 %**	✅ **PLQY 最划算**
cn==4 & 纯溴	8.1%	46	23	74.0 nm	**13.1 %**	PLQY 好，发射不可预测
cn≤6 & 任意配体	88.4%	501	50	69.4 nm	28.1 %	— 几乎无改善
不限定（全集）	100%	567	52	**72.7 nm**	28.1 %	基线

帕累托权衡：要准就得放弃覆盖率

前沿点（不被其他窗口支配）为：**(7%, 5.8nm) → (19%, 33.9nm) → (51%, 63.1nm) → (88%, 69.4nm) → (100%, 72.7nm)**。含义非常明确：

- **覆盖 ≤20%** 时误差断崖式下降 —— 限定到 cn≤3 任意配体即覆盖 109 个结构、误差降到 34 nm（比基线好一倍）。
- **覆盖 38–51%** 时误差已逼近基线（63 nm），**收益趋平**——再往上放宽（如 cn≤6，覆盖 88%）几乎没有额外收益（69 nm）。
- 少数窗口**反而比基线更差**（如"含 N/O 的高配位"：42% 覆盖、误差 80.6 nm），说明"限定"必须限定在**低配位**，光限定配体不够。

可直接落地的三条

1. **发射波长**：选 `cn≤3 & 任意配体`（覆盖 19%，误差 ±34 nm）或更狠的 `cn≤3 & 纯溴`（覆盖 7%，误差 **±6 nm**）。
2. **PLQY**：选 `全Br（任意配位）`（覆盖 16%，误差 ±17.7 %，且有 32 个标记，数据最厚）——这是 PLQY 性价比最高的窗口。
3. **数据补充方向**：优先补 `cn≤3` 与 `全Br` 两个窗口内的同系化合物；反之 **cn≥4 含 N/O** 的区域已被证明是"黑洞"，不值得再投数据。

6. 其余阴性尝试（摘要）

- **IMA 文献挖 10Dq/B**：100 篇 MnBr 文献 0 篇给出逐化合物 10Dq/B → 死路。
- **扩充标签集(+6)**：全落在不可预测孤立-Mn 子集，发射折崩溃 → 不部署。
- **簇级近似 DFT(pyscf)**：10/58 簇收敛(覆盖~1.2%)，共价性近常数(~1.6) → 未破天花板；已备好集群 ORCA 输入。

7. 虚拟筛选 v2（集成 + 裁剪）与交付

重建为**多随机种子集成**（20 模型平均）+ **物理裁剪**（发射 400–800 nm、PLQY 0–100%，共修正 7/567 个越界预测）。集成平均的诚实多种子 grouped-CV：发射 **+0.106±0.206**、PLQY **+0.001±0.288**——**并未提高诚实**

精度（集成只降"单次抽样噪声"，主导方差来自数据/折分本身）。

> **关于不确定度列的诚实告警**
>
> CSV 中的 `ens_spread_*` 是**集成成员分歧度**（中位数 ~8.5 nm / 2.3 %），**不是真实预测误差**，会严重低估不确定度（真实发射误差可达数十至上百 nm）。本文件仅作**极弱排序线索**，任何候选必须实验验证；裁剪属护栏、非精度提升。

`virtual_screen_v2.csv`：567 个 Br 结构（509 novel / 58 known）。高置信 503、低置信 64（纯溴孤立 Mn，β 退化，外推）。

8. 结论与建议

> **能做什么（正面结论）**
>
> 1. **能可靠预测「是否亮」**：PLQY≥50% 二分类，平衡准确率 **0.754±0.084**（随机 0.5，基线 0.608），已在 `screen_bright_dim.csv` 对 567 个 Br 结构打分（177 高置信亮 / 246 可能亮 / 144 偏暗）。这是本项目**唯一经得起多种子检验的预测能力**。
> 2. 机制清晰：N/O/S 软共价配体 → 共价增强 → Laporte 放宽 → PLQY 升高，与筛出的 [$MnBr_4$(N)] 候选一致。

> **不能做什么（诚实边界）**
>
> 1. **精确发射波长/颜色不可预测**（颜色分类平衡准确率 0.48 ≈ 随机）。
> 2. **精确 PLQY 数值不可预测**（多种子 R^2 仅 +0.03±0.29）。
> 3. 此前"PLQY +0.40 / 发射 +0.24"是单种子运气值（见第 2 节）。
> 4. 加入物理约束（裁剪/单调/Dq/B/配位窗口）无法稳健提升——方差受限。
> 5. 模型是**方差受限**：可靠提升只能来自**更多数据**或**真 DFT 电子结构参数**。

建议

1. **把模型用在它擅长的地方**：用「亮/暗」筛法先过滤候选（这是可靠的），再用实验/DFT 定波长与效率。
2. **评估协议升级**：所有 CV 默认 ≥10 个折分种子并报 mean±std，禁止单种子结论（已写入项目 memory）。
3. **限定合成空间（最有效）**：发射选 cn≤3 窗口（覆盖19%、误差±34nm；纯溴可到±6nm）；PLQY 选全Br窗口（覆盖16%、误差±17.7%），见 §5。
4. **定向补数据**：按目标家族定向采集同系化合物，而非继续扩大总体样本量。
5. **正规 DFT**：在集群跑已备 ORCA 输入（`gen_orca_inputs.py`），拿逐结构 10Dq 与 Racah B/C。

附录 · 关键文件

文件	说明
`constrained_model.py` / `constrained_multiseed.py`	物理约束 + 多种子检验（本版新增）
`gnn_graphs.pkl`	781 张图（发射 60 / PLQY 57 标签）
`descriptors_physics.csv` / `_dft.csv`	几何+文献物理(+DFT)描述符
`virtual_screen.csv`	虚拟筛选（567 Br 结构）
`dft_10dq.py` / `gen_orca_inputs.py`	簇级 DFT / 集群 ORCA 输入
`build_report.py`	本报告生成脚本（无头 Chrome 导出 PDF）