# 需要哪些数据才能让模型更准？——一份基于证据的实验设计

**数据科学团队 · 2026-10-13**

## 0. 先纠正一个前提

我们本地**已经持有全部 796 个 CIF 结构**。所以缺的不是结构，而是**实测性质**：

645 个溴化物结构里，只有 **58 个**带发射或 PLQY 标签（**9%**）。因此问题不是"要不要找更多结构"，

而是 **"在这已有的结构里，优先测哪几个、测什么"**。下面是用算法 + 实验证据得出的答案。

## 1. 三个反直觉的实验结论（先看这个，它决定了策略）

### ① 测量精度不是瓶颈

| 重复测量精度 s_rep | 发射 $R^2$ 理论上限 |
|---|---|
| ±3 nm | 0.998 |
| ±5 nm | 0.995 |
| ±10 nm | 0.980 |
| ±15 nm | 0.956 |

我们当前实测 $R^2$ 只有 **0.12**。即使测量误差高达 **±15 nm**，理论上限仍有 **0.94**。

→ **花大力气做重复性验证不会带来模型提升。**（但仍建议做 10–15 个化合物的三重复，

用来知道自己的地板在哪，以及确认没有系统误差。**优先级：低**。）

### ② 往"全集"里加标签，反而会变差

我们做了学习曲线（固定 30% 测试集，10 种子均值）：

| 训练分子式数 | 全集 Br 的发射 $R^2$ |
|---|---|
| 5 | −1.03 |
| 10 | −0.98 |
| 15 | −6.15 |
| 20 | −2.04 |
| 25 | −5.50 |

**$R^2$ 随标签增加而恶化。** 原因：全集把**性质相反的区域混在一起**——低配位窗口的规律

（发射锁死、可预测）和"黑洞区"（发射发散、不可预测）互相抵消。

→ **盲目扩大总数据量是有害的。**

### ③ 但在单个窗口内加标签，趋势是好的

| 训练分子式数 | cn4+cn5 窗口内 $R^2$（变化率） |
|---|---|
| 5 | −0.79 |
| 10 | −0.64 |
| 15 | −1.15 |

该窗口每多 5 个标签，$R^2$ 仅变化 **−0.036**（全集是 **−0.200**），即**退化慢 5 倍**。

→ **核心结论：数据必须"按窗口成建制地补"，不能"往全集里撒"。**

---

## 2. 因此的策略：三个战役（campaign），而不是一批 120 个

`MEASUREMENT_BATCH_DESIGN.csv` 已按此分好组，每个化合物都标了所属战役与推荐优先级：

### 🟢 C1 · 低配位窗口（最优先，先做这个）

**范围**：Mn 配位数 ≤3

**候选**：**35 个**（列表中 campaign = C1）

**为什么**：该窗口发射锁死，家族内误差仅 **±6~34 nm**。这是唯一能靠数据做出**可靠模型**的区域。

**测什么**：T1 发射 + PLQY（必测）、T2 寿命、T3 吸收光谱

**预期**：补到 ~35 个标签后，可支撑一个真正可用的窗口专属模型。

### 🟡 C2 · 四配位窗口（第二优先）

**范围**：cn = 4–5

**候选**：**29 个**

**为什么**：PLQY 在此收窄到 **±13~17.7%**，可预测性最好；但**发射不可预测**（误差 74 nm）。

**建议**：这一区主要用我们的**亮/暗分类器**（已验证 0.754 平衡准确率），发射交给 DFT/实验。

### 🔵 C3 · 八面体窗口（先探后投）

**范围**：cn ≥ 6

**候选**：**56 个**（库里最多的区域，但只有 3 个标签）

**重要前提**：**先做 20 个的吸收光谱**，验证 10Dq 能否解释其发射。

八面体 $d^5$ 的 Tanabe–Sugano 分析最成熟，10Dq 提取最可靠——**这是打开商用磷光体市场最关键的一块**，但**先验证再放量**，不要一上来就合成 56 个。

---

## 3. 每个化合物测什么（三个层级）

| 层级 | 内容 | 理由 | 成本 |
|---|---|---|---|
| **T1 必测** | 发射波长 + PLQY（给不确定度） | 没有标签就没有模型 | 低 |
| **T2 强烈建议** | 荧光寿命 τ | $k_{nr}=(1-\Phi)/\tau$，**这是决定"亮不亮"的物理量**，目前全库为 0 | 中 |

| 层级 | 内容 | 理由 | 成本 |
|---|---|---|---|
| **T3 建议（最便宜）** | **吸收光谱 250–800 nm** | → **10Dq**（第一吸收带）与 **Racah B/C**（多带拟合）。这是模型完全缺失的物理变量，且我们扫 100 篇文献 **0 篇报告过** | 低 |

> T3 性价比最高：**一次扫描同时解决 C1 和 C3 两个战役的问题**。

---

## 4. 选样算法（保证不是拍脑袋）

`design_measurement_batch.py` 的选样逻辑：

**化学合理性过滤**：剔除异金属配合物（Fe/Ni/Cr/Re/Ge/Co…）、Mn 核数 >4、分子式重原子数 >42、元素过少的非配合物 → 567 个 Br 结构剩 **455 个合格候选**（排除 112）。

**空缺配额**：优先补 `cn≤3` (30) / `cn6` (30) / `cn5` (15) / `noBr` (12) 四个区域。

**空间填充（max-min 距离）**：在标准化描述符空间里，每次选离"已标注+已选"最远的点。

→ 因为模型是**方差受限，把标签摊开比堆在一起更有价值**。

最终选出 **120 个结构**，每个都带 `why` 字段说明它为什么被选中（gap-fill 还是 space-filling）。

---

## 5. 给化学/实验团队的行动清单

| 优先级 | 动作 | 数量 | 说明 |
|---|---|---|---|
| **1** | **C1 低配位窗口全表征** | 35 个 | 发射+PLQY+寿命+吸收。做完我们能给出一个真正可用的窗口模型 |
| **2** | **吸收光谱批量扫描** | 现有样品 + 新样品 | 最便宜、解锁 10Dq/B；**顺便做 10–15 个三重复**确认噪声地板 |
| **3** | **C3 八面体先验证 20 个** | 20 个 | 只测发射+吸收，验证 10Dq 是否有效；**有效再放量到 56 个** |
| **4** | C2 四配位补 PLQY | 29 个 | 主要为了喂亮/暗分类器 |

---

## 6. 我们承诺

每收到一批数据，我们会：

重新画学习曲线，确认 $R^2$ 是否真的随标签上升（**如果没上升，立刻告诉你们停**）；

更新"哪些区域是黑洞"的结论，避免继续无效投入；

更新可交付的筛选清单（亮/暗预测 + 窗口专属预测区间）。

---

## 7. 一句话总结

> **不是"多测点数据"，而是"在正确的窗口里、成建制地测、并且把缺的物理变量（10Dq、寿命）一起测回来"。**

> 盲目往全集加标签会让模型变差；在单个窗口内加标签才是安全且有效的路径。