持续局部拉普拉斯（PLL）预测蛋白质–配体结合亲和力

Persistent Local Laplacian Prediction of Protein–Ligand Binding Affinities

刘健　重庆理工大学

第三届大湾区拓扑会议

内容提纲

1. 研究概览和背景
2. 方法和建模
3. 结果和分析
4. 总结与展望



背景：药物发现的困境

核心痛点

药物发现**耗时、昂贵、失败率高**。能否准确预测蛋白质–配体结合亲和力，直接决定小分子药物能否识别并调控其生物学靶点。

传统方法及其局限

- 分子对接（Docking）
- 自由能微扰（FEP）
- 经验打分函数（Scoring functions）
- **问题**：精度有限、计算昂贵、难以刻画复杂相互作用机制

ML / AI 的兴起

- AlphaFold 等蛋白质结构预测
- 图神经网络 / CNN / Transformer
- **关键瓶颈**：如何构造**好的分子表示（描述符）**

分子描述符的质量，直接决定模型性能的成败。

背景：从拓扑到谱的演进

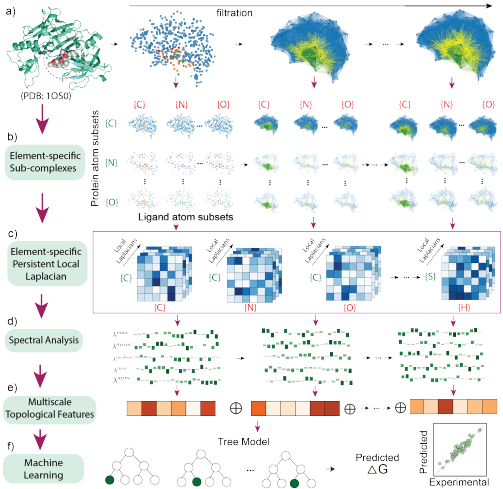
代数拓扑 / 谱方法为何适用于分子表示

生物大分子功能高度依赖其三维结构。拓扑与谱方法能从多尺度刻画几何与拓扑特征，特别适合描述氢键、静电、范德华力等物理化学相互作用。

- **持续同调（PH）**：多尺度拓扑抽象，曾在 D3R 国际药物设计挑战赛获顶尖成绩
- **持续拉普拉斯（PL）**：在 PH 基础上，利用拉普拉斯矩阵的**调和/非调和谱**编码更多几何信息
- **局限**：PH 与 PL 主要捕获**全局**拓扑，对**局部**结构变化不敏感
- **本工作 PLL**：显式引入**局部化**，同时刻画**局部拓扑 + 全局拓扑 + 几何信息**



PLL 流程总览



过程 (a)→(f): 原子 → 滤过 → 局部拉普拉斯 → 谱特征 → 多尺度特征 → GBDT 预测

从单纯复形到局部拉普拉斯

单纯复形（Simplicial complex）：由多个单纯形组成。p-单纯形是 $p+1$ 个仿射无关点的凸包。

- 0-单纯形 = 顶点，1-单纯形 = 边，2-单纯形 = 三角形，3-单纯形 = 四面体

单纯拉普拉斯（图拉普拉斯的高维推广）：

$$\Delta_k = \partial_k^* \partial_k + \partial_{k+1} \partial_{k+1}^*, \qquad k \geq 0,$$

其中 ∂_k 为边界算子，$(\cdot)^*$ 为伴随算子。

局部拉普拉斯（在顶点 v 处）

令 $\mathcal{L} = \mathcal{K} \setminus \{v\}$ 为不含 v 的最大子复形，则

$$\Delta_k^{\mathcal{K},v} = \overline{\partial}_{k+1} \overline{\partial}_{k+1}^* + \overline{\partial}_k^* \overline{\partial}_k, \qquad k \geq 0,$$

是相对链复形 $C_*(\mathcal{K}, \mathcal{K} \setminus \{v\})$ 上的自伴算子。

局部同调的切除定理（关键理论性质）

局部同调 $H_k^{\mathrm{loc}}(\mathcal{K}, v)$ 只依赖 v 的**任意小邻域**，与整体空间结构无关。

两个直接优势：

- 局部同调 / 拉普拉斯只需在**单点邻域**内计算 $\Rightarrow$ 大幅降低计算复杂度
- 不同点处的计算**天然可并行** $\Rightarrow$ 显著提升效率

第三个优势——可扩展性：

- 即使整个数据集规模增大，**已预计算的局部拓扑/几何特征不受影响、无需重算**
- 局部多尺度检测 **犹如一个可调放大倍率的显微镜**

持续局部拉普拉斯（PLL）

考虑单纯复形滤过

$$\mathcal{K}_0 \subseteq \mathcal{K}_1 \subseteq \cdots \subseteq \mathcal{K}_m,$$

对任一顶点 $v \in \mathcal{K}_0$，得到相对链复形的滤过。对任意 $s \le t$，定义自伴半正定算子：

$$\Delta_k^{s,t}(\mathcal{K}_\bullet; v) : C_k(\mathcal{K}_s, \mathcal{K}_s \setminus \{v\}) \to C_k(\mathcal{K}_s, \mathcal{K}_s \setminus \{v\}).$$

PLL 的双重内涵

- 其**调和分量**同构于**持续局部同调**（保留拓扑信息）
- 其**非调和谱**刻画底层结构的**局部几何特征**

关键结果：与链环拉普拉斯的酉等价

PLL 直接理解和计算较难，但可通过**持续链环拉普拉斯**高效求值。
链环复形：$\mathrm{Lk}_{\mathcal{K}}(v)$ 由所有满足 $v \notin \sigma$ 且 $\sigma \cup \{v\} \in \mathcal{K}$ 的单纯形 σ 组成。

定理 (酉等价)

设 $\mathcal{K}_\bullet$ 为单纯复形滤过，$v \in \mathcal{K}_0$。对任意 $s \le t$ 与 $k \ge 0$，

$$\Delta_k^{s,t}(\mathcal{K}_\bullet; v) \;\cong\; \Delta_{k-1}^{s,t}\big(\mathrm{Lk}_{\mathcal{K}_\bullet}(v)\big),$$

即持续局部拉普拉斯与持续链环拉普拉斯**酉等价**；其核心（持续调和空间）同构。

- 链环复形比相对商链复形**计算简单得多** $\Rightarrow$ 为大规模数据分析提供数学基础

元素特异 PLL 分子描述符

元素特异建模

将蛋白质–配体复合物中的原子按**元素类型分组**，捕捉氢键、静电、范德华力、疏水/亲水等相互作用。

- 蛋白质常见元素：{C, N, O, S}；配体常见元素：{C, N, O, S, P, Cl, Br, I, H}
- 共产生 **40 种**蛋白质–配体原子类型相互作用对
- 从配体向外截断距离取 **12 Å** 收集邻近蛋白原子
- 对每个点云，以 Vietoris–Rips 复形、滤过半径 **2–6 Å、步长 0.5 Å** 做 PLL 分析
- 每个原子用 9 类特征统计（和、均值、中位数、标准差、最大值、最小值、方差、平方正特征值之和、零特征值个数）



特征规模与组合

蛋白质–配体复合物特征

- 40 个原子类型组合
- 每个组合：蛋白 18 + 配体 18 = 36 维 @ 单半径
- 9 个滤过半径 ⇒ **324 维 / 组合**
- **合计 12,960 维**

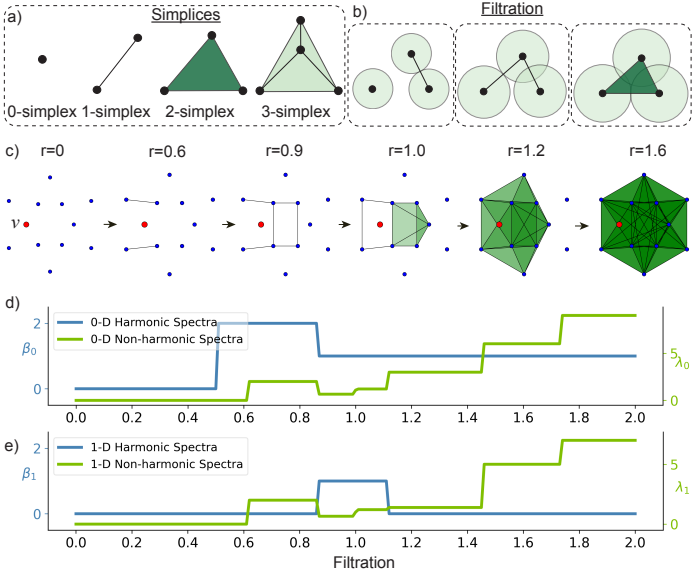
配体自身特征

- 13 个原子子集
- 每个子集 36 维 @ 单半径 × 9 半径
- **合计 2,106 维**
- 与复合物特征拼接 ⇒ 同时刻画**分子间 + 分子内**相互作用

PLL 在三个数据集上的滤过最优上限均为约 6 Å。



PLL 可视化示例：点云中的局部谱分析



基于序列的 Transformer 特征

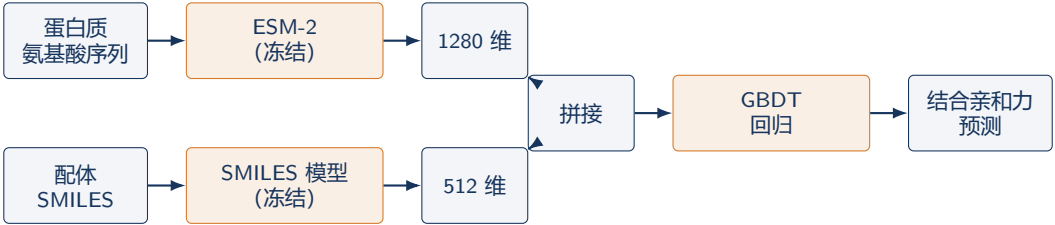
- **蛋白质**：ESM-2（esm2_t33_650M_UR50D，34 层、650M 参数），取末层 1280 维嵌入
- **配体**：SMILES 语言模型（model-CPZ），取首个 token 嵌入

机器学习模型：梯度提升决策树（GBDT）

- 独立训练：PLL 结构特征 **或** Transformer 序列特征
- 关键超参：30000 / 20000 棵树，最大深度 7，最小分裂 5，学习率 0.002，特征取平方根，子采样 0.8，重复 10 次
- **PLLML** = PLL 与 Transformer 输出**取平均**的一致性模型



Transformer 的使用（一）：作为“冻结”的序列特征提取器



- **不参与训练**：全程冻结预训练权重，只做一次前向推理，把**变长序列压成定长向量**；真正的回归学习由后面的 **GBDT** 完成。
- **两条特征线并行**：结构线（PLL，三维几何）与序列线（Transformer，一维序列）**各自独立训练。**
- **融合**：两模型预测**取平均** $\rightarrow$ **PLLML。**

Transformer 的使用（二）：蛋白质与配体的具体模型

蛋白质序列编码：ESM-2

- 模型：**ESM-2**（`esm2_t33_650M_UR50D`，约 650M 参数），在数亿条蛋白质序列上预训练。
- 输入：蛋白质的**氨基酸序列**（变长，一维）。
- 输出：取**末层** token 表示做**均值池化**，得到 **1280 维**向量。

配体序列编码：SMILES 语言模型

- 模型：在 **7 亿 +** 条 SMILES（ChEMBL/PubChem/ZINC）上预训练的 **model-CPZ**。
- 输入：配体的 **SMILES 字符串**。
- 输出：256×512 嵌入矩阵，取**首个 token**，得 **512 维**向量。

特点与定位

两编码器**全程冻结**（防止过拟合，便宜、稳定、可复现）；序列特征拼接后交给 **GBDT**，与 PLL 模型**并行互补**，**取平均**得 **PLLML**。

各策略性能对比

数据集	PLL	Transformer	PLL+Transformer
PDBbind-v2007	0.813 (1.971)	0.785 (2.006)	**0.827 (1.925)**
PDBbind-v2013	0.792 (1.976)	0.791 (1.974)	**0.813 (1.932)**
PDBbind-v2016	0.850 (1.672)	0.836 (1.713)	**0.861 (1.646)**
平均	0.818 (1.873)	0.804 (1.898)	**0.834 (1.834)**

括号内为 RMSE（kcal/mol），随机种子独立训练 10 次取平均。



数据集信息

数据集	总数	训练集	测试集
PDBbind-v2007	1300	1105	195
PDBbind-v2013	2959	2764	195
PDBbind-v2016	4057	3767	290

PDBbind 数据库

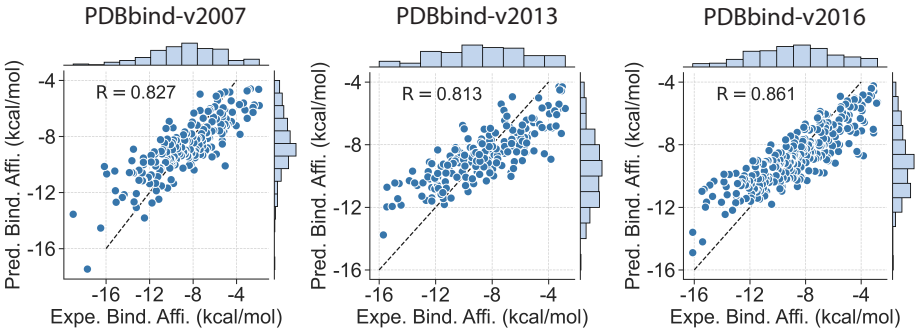
系统收集实验测定的蛋白质–配体复合物三维结构及其结合亲和力数据（K_d、K_i、ΔG），是评估该领域预测模型的**金标准**。

与先进数学机器学习模型的对比

模型	v2007	v2013	v2016	平均
PLLML	0.827	**0.813**	**0.861**	**0.834**
PerSpect-ML	**0.836**	0.793	0.840	0.823
TopBP-ML	0.818	0.804	0.848	0.823
PPS-ML	0.827	0.791	0.843	0.820

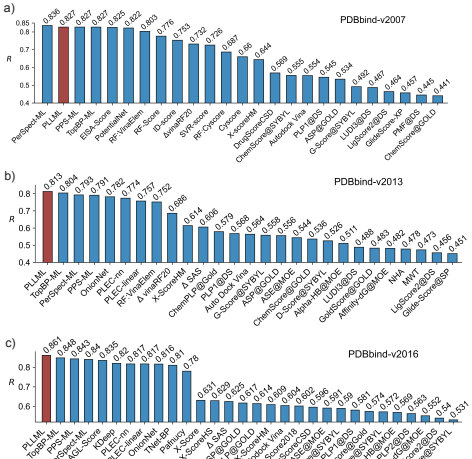
- PLLML 在 v2013 与 v2016 上**显著领先**
- 仅 v2007 上略逊于 PerSpect-ML（0.827 vs 0.836）
- 平均 R 值（0.834）明显高于全部对比模型

预测性能与实验值对比



PLLML 模型在三个 PDBbind 数据集上，预测结合亲和力与实验值的强一致性。

与已发表模型的 R 值对比



PLLML 与其它已发表模型在三个 PDBbind 数据集上 Pearson 相关系数（R）的对比。

滤过半径的影响（可解释性）

- 在 v2007 与 v2016 上，性能随滤过半径从 4 Å 增至 6 Å **逐渐提升**；超过 6 Å 略下降
- 在 v2013 上，最优为 **7 Å**（相对 6 Å 仅边际改善）
- **结论：6 Å 为 PLL 建模的最佳滤过上限**，兼顾性能与特征构造效率

为什么 6 Å 合理

约 6 Å 有效刻画了每个原子与其周围环境的**关键相互作用**——更短尺度捕捉直接接触（氢键、疏水作用、盐桥），4–6 Å 捕捉水介导相互作用。

计算复杂度优势（核心卖点）

拉普拉斯类型	构造复杂度	谱复杂度	总复杂度
组合（全局）拉普拉斯			
0 维 $\mathbf{L}_0$	$O(n_0^2 n_1)$	$O(n_0^3)$	$O(n_0^2 n_1 + n_0^3)$
1 维 $\mathbf{L}_1$	$O(n_1^2(n_0 + n_2))$	$O(n_1^3)$	$O(n_1^2(n_0 + n_2) + n_1^3)$
局部拉普拉斯（基于链环复形）			
链环复形	$O(n_0 q)$	–	$O(n_0 q)$
0/1/2 维链环拉普拉斯	$O(q)$	$O(q^3)$	$O(q + q^3)$
局部拉普拉斯总计	$O(n_0 q)$	$O(n_0 q^3)$	$O(n_0 q + n_0 q^3)$

当 $n_0, n_1, n_2 \gg q$（q 为含单个顶点的最大单纯形数）时，$O(n_0 q^3)$ **远小于全局的** $O(n_0^3)$ **或** $O(n_1^3)$，验证了局部拉普拉斯在大规模数据上的实用性。

研究总结

我们做了什么

提出**持续局部拉普拉斯（Persistent Local Laplacian, PLL）**这一新型数学框架，用它构建**分子描述符**，结合机器学习模型，实现对蛋白质–配体结合亲和力的**高精度预测。**

- **研究对象**：蛋白质–配体结合亲和力（药物发现的核心问题）
- **方法根源**：代数拓扑 / 谱理论——持续同调（PH）与持续拉普拉斯（PL）的深化
- **核心创新**：把“全局”拓扑谱分析，细化为“点/原子局部”的拓扑与几何谱分析
- **验证基准**：PDBbind 三大标准数据集（v2007 / v2013 / v2016）
- **最终模型**：PLL + 序列（NLP）Transformer 的集成模型 **PLLML**



主要结果一览

数据集	PLLML	对比方法
PDBbind-v2007	0.827	PerSpect-ML 0.836
PDBbind-v2013	**0.813**	PerSpect-ML 0.793
PDBbind-v2016	**0.861**	PerSpect-ML 0.840
平均	**0.834**	PerSpect-ML 0.823

- 平均 Pearson 相关系数 R 达 **0.834**，RMSE 低至 **1.834** kcal/mol
- 三项基准上**超越**现有最先进模型（TopBP-ML、PerSpect-ML、PPS-ML）
- 纯 PLL 模型（不借助 Transformer）在 v2016 上已优于大多数已发表模型

核心贡献

提出**持续局部拉普拉斯（PLL）**理论：通过局部拉普拉斯矩阵的谱分析，同时刻画**局部拓扑结构**与**几何信息**，是持续同调与持续拉普拉斯的重要深化。

- **理论**：局部同调由切除定理保证与全局无关；持续局部拉普拉斯与**链环拉普拉斯酉等价**，计算高效
- **方法**：元素特异 PLL 构造描述符 + NLP 序列特征 + GBDT，构成 **PLLML** 集成模型
- **结果**：三个 PDBbind 基准上平均 $R = 0.834$、RMSE=1.834 kcal/mol，**超越现有最先进模型**
- **效率**：局部拉普拉斯的谱复杂度 $O(n_0 q^3)$ **远低于**全局方法，支持大规模分析

价值与前景

为什么 PLL 有优势

- **局部敏感**：克服了 PH / PL 对局部变化不敏感的问题
- **高效**：单点邻域内计算，天然可并行
- **可扩展**：系统规模扩大无需重算局部特征
- **可解释**：谱特征与局部拓扑、几何有清晰对应

应用前景

除蛋白质–配体结合亲和力外，PLL 可推广至
药物发现、蛋白质工程及其它科学与工程中的大分子预测任务。



代码和数据

数据与代码

- 复现论文结果的代码与数据：https://github.com/hfeng2uncc/PLLML.git
- PDBbind 数据集：https://www.pdbbind-plus.org.cn/
- 论文原文：https://doi.org/10.1021/acs.jcim.6c02460





谢谢观看！

欢迎提问与讨论

持续局部拉普拉斯（PLL）预测蛋白质–配体结合亲和力