arXiv:2601.12219v1 [math.SP] 18 Jan 2026

突变引起的蛋白质稳定性与溶解度变化的持久层拉普拉斯分析

YimingRen[1],Junjie Wee[1],Xi Chen[2],Grace Qian[3],和 Guo-Wei Wei[1, 4, 5*]

[1]密歇根州立大学数学系，美国密歇根州东兰辛，MI 48824。[2]弗雷泽学校，美国佛罗里达州盖恩斯维尔，NW 89 大道 4700 号，FL 32606[3]拉西特高中，美国佐治亚州玛丽埃塔，GA 30066[4]密歇根州立大学生物化学与分子生物学系，美国密歇根州东兰辛，MI 48824。[5]密歇根州立大学电气与计算机工程系，美国密歇根州东兰辛，MI 48824。

2026年1月21日

摘要

基因突变常常会破坏蛋白质结构、稳定性和溶解度，是多种疾病的主要诱因。尽管这些分子改变至关重要，现有计算模型往往缺乏可解释性，也未能整合关键的物理化学相互作用。为克服这些局限，我们提出 SheafLapNet，这是一个基于拓扑深度学习（TDL）和持久层拉普拉斯算子（PSL）数学理论的统一预测框架。与持久同调等标准拓扑数据分析（TDA）工具不同，后者往往对异质信息不敏感，PSL 会将部分电荷等特定物理和化学信息直接编码到拓扑分析中。SheafLapNet 将这些层论不变量与先进的蛋白质 Transformer 特征及辅助物理描述符相结合，以多尺度、机制性的方式捕捉内在的分子相互作用。为验证该框架，我们针对回归和分类任务进行了严格的基准测试。在稳定性预测方面，我们采用了全面的 S2648 和 S350 数据集。在溶解度预测方面，我们使用 PON-Sol2 数据集，该数据集标注了溶解度增加、降低或中性的变化。通过整合这些多视角特征，SheafLapNet 在各类基准测试中均达到最先进的性能，表明层论建模能够显著提升对突变诱导的结构和功能变化进行预测时的可解释性与泛化能力。

关键词：蛋白质折叠稳定性、蛋白质溶解度、突变、持久拓扑拉普拉斯算子、层拉普拉斯网络

*通讯作者。电子邮箱：weig@msu.edu

1 引言

基因突变会改变氨基酸序列，从而严重破坏蛋白质结构精细的热力学平衡。这些原子层面的扰动常会损害关键理化属性，包括折叠稳定性、结合亲和力、反应性和溶解度，进而导致功能缺失表型，或引发与严重病理状况相关的毒性聚集。这类分子机制是阿尔茨海默病 [22] 和帕金森病 [18],等神经退行性疾病，以及各种癌症 [13]和代谢综合征 [59, 60]发病机制的核心。尽管稳定性和溶解度特征对临床至关重要，要准确预测突变引起的变化仍面临巨大挑战。蛋白质溶解度受多种因素共同影响，既包括固有的序列基序和翻译后修饰，也包括溶剂类型、离子强度和温度等外部环境条件。海量基因组数据进一步加剧了这一复杂性：尽管已编目超过四百万个错义变异，但其中大多数仍被归类为意义未明变异（VUS） [25], ，而现有数据集往往缺乏全面的环境注释，难以厘清这些细微差异。鉴于传统实验验证仍然耗费资源且通量较低 [30, 33], ，亟需先进的计算方法来解析这些复杂的生物物理相互作用，从而可靠地预测遗传变异的功能影响。

蛋白质稳定性和溶解度是蛋白质功能的基础，并直接参与人类疾病的发生；然而，预测突变引起的这些性质变化仍是一项复杂的计算挑战。在稳定性研究领域，深度学习架构充分利用了大规模数据集。例如，mutDDG-SSM [26] 将图注意力网络与梯度提升树相结合，而 TopologyNet [4] 则采用多任务框架，将稳定性数据与疾病关联联系起来。尽管取得了这些进展，过去十年开发的模型（如 STRUM [37], ）利用现代机器学习揭示了结构与稳定性之间的隐含关系，但往往缺乏足够的可解释性。与此同时，针对突变引起的溶解度变化的预测研究也取得了进展，推出了 CamSol[40], SODA [35], Solubis [45], PON-Sol [57],以及 PON-Sol2[58],等工具；后者利用扩展数据集进行梯度提升。然而，目前仍缺乏能够同时处理这两种相互关联性质的统一框架。此外，尽管许多现有模型具有较强的预测能力，但它们可能无法明确考虑支配分子行为[41, 43]的基本物理相互作用，包括氢键、范德华力、疏水性和静电作用。这一局限，加上溶解度任务中归一化正确预测率（CPR）等性能指标不够理想，凸显了对可解释人工智能（XAI）框架的迫切需求，以提供具有可解释性且以机制为依据的疾病因果关系见解。

为应对这些挑战，我们转向拓扑数据分析（TDA）。这是一门新兴的数学领域，运用代数拓扑分析复杂数据中的结构模式。其核心工具持久同调 [7, 17, 61], 结合了经典同调与滤过，可对数据进行多尺度分析。与机器学习相结合时，TDA便成为数据科学中的一种强大方法。尤其是，融合TDA与深度神经网络的拓扑深度学习（TDL）首次于 2017 [4]提出。多年来，TDL已成为理性学习领域的新前沿 [36]。各种深度神经网络也已相继提出 [2, 20]。TDL在解析生物分子方面取得了巨大成功 [3]。TDL持续展现出相较于其他竞争方法的明显优势，其中一些最有说服力的应用实例包括其在D3R大挑战中的出色表现

[34], 它在揭示SARS-CoV-2 [11], 进化机制方面发挥的作用，以及成功预测病毒变体的能力 [12]。

然而，基于持久同调的传统拓扑数据分析（TDA）存在固有局限：它对没有拓扑变化的同伦形状演化不敏感，无法区分几何异构体，也不能区分原子类型或编码有向关系。为克服这些局限，人们提出了持久谱理论 [47],（亦称持久拉普拉斯算子（PL） [31, 49], ）。通过利用非调和谱捕捉同伦形状演化，HERMES 软件 [49], 同伦延拓[52] 和 PETLS [23] 等计算算法推动了拓扑深度学习新范式的发展。这种方法在多种生物分子任务中展现出卓越效能，包括蛋白质-配体结合亲和力预测 [32], 蛋白质-蛋白质相互作用分析 [29, 51], ，以及提前预测主要 SARS-CoV-2 变异株，例如在世界卫生组织（WHO）公布奥密克戎 BA.4 和 BA.5 前两个月便对其作出预测 [10]。

尽管标准持久同调和 PL 均已证明具有实用价值，但二者都无法捕捉复杂数据集中的异质信息。为克服这一局限，最初提出了元素特异性持久同调，以明确区分不同原子类型 [3], 这一方法随后启发了一系列新颖的 TDA 方法 [1, 19,28, 29]。在此基础上，持久层拉普拉斯算子（PSL） [53] 近期作为一种数学上更优雅的理论崭露头角，它能够通过胞腔层理论将几何和部分电荷等异质信息直接嵌入拓扑分析 [14, 21]。与标准拉普拉斯算子不同，这种层论框架允许将代表化学性质的特定向量空间赋予蛋白质拓扑的开集 [53]。通过为原子（节点）关联特定权重，PSL 将局部拓扑和几何信息直接编码到其调和谱和非调和谱中，从而能够精确捕捉突变引起的内在物理和化学相互作用。关于持久拓扑拉普拉斯算子的综述见 [54]。 [42]综述了持久同调以外的 TDA 方法。 [36]回顾了 TDL 的最新进展。关于应用前景，读者还可参阅一篇全面综述，了解 TDA 和 TDL 在分子科学中的应用[51]。

在本研究中，我们提出了 SheafLapNet，这是一种基于 PSL 理论的新型深度学习框架，旨在预测突变对蛋白质稳定性和溶解度造成的变化。SheafLapNet 架构整合了三类不同特征，构建全面的分子表征：（1）由持久层拉普拉斯算子导出的多尺度拓扑和几何特征；（2）用于捕捉局部原子相互作用的辅助理化描述符；（3）从预训练 ESM-2 蛋白质 Transformer [27] 中提取的进化序列嵌入。这些多样化的特征集被整合到统一的神经网络框架中，用于预测稳定性和溶解度变化。我们采用 S2648 和 S350 数据集进行稳定性预测，并采用 PON-Sol2 数据集进行溶解性分类，对 SheafLapNet 与既有基准进行了严格评估。在这些全面的评估中，SheafLapNet 始终优于现有最先进模型。这些结果表明，模型的预测准确率有所提升，验证了将层论不变量用于突变影响预测的有效性。

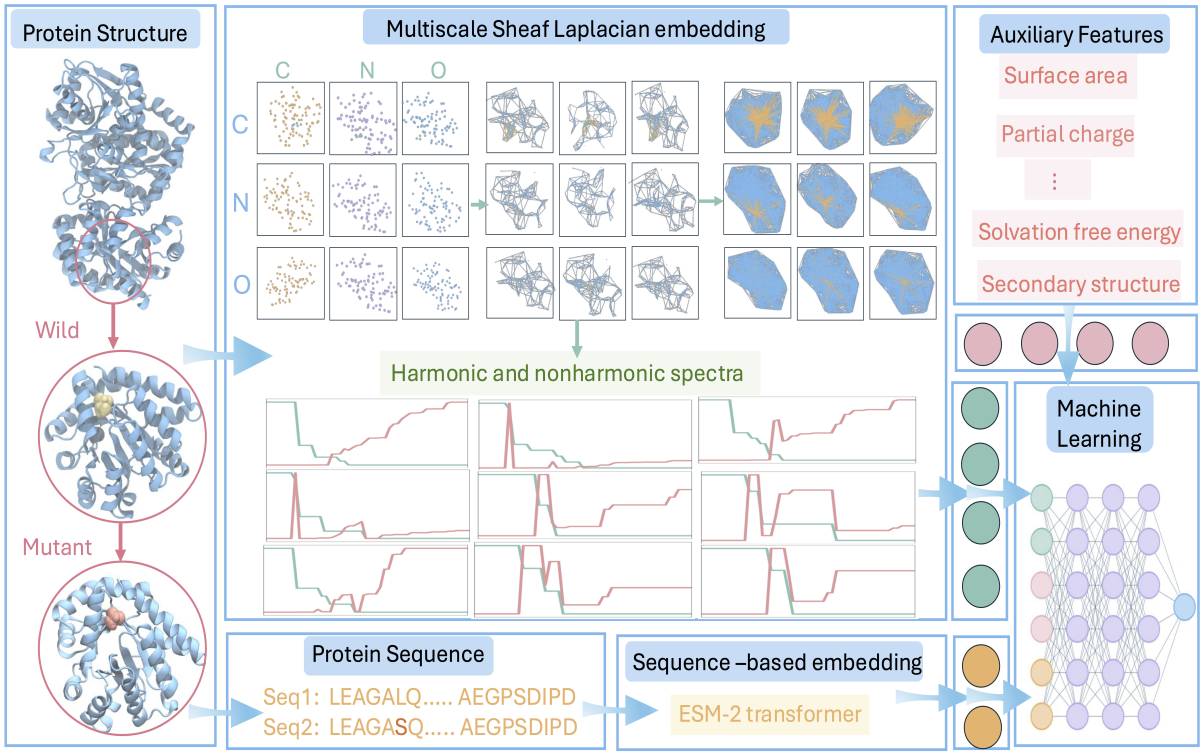


图 1： 持久层拉普拉斯算子（PSL）神经网络（SheafLapNet）工作流程示意图。该框架通过整合多尺度蛋白质表征，预测突变诱导的稳定性和溶解度变化。对于每个输入的蛋白质结构，特征生成流程提取三个不同的组成部分：(1) 从预训练蛋白质Transformer模型中获得的基于序列的嵌入；(2) 通过PSL框架计算的拓扑特征；以及 (3) 辅助物理化学特征。这三组特征经过拼接后，作为神经网络的输入，用于执行预测任务。

2 结果

2.1 SheafLapNet概述

图 1 概述了 SheafLapNet 的工作流程。作为一种标准的机器学习模型，SheafLap-Net 从蛋白质结构中提取特征，并利用神经网络预测突变导致的稳定性和溶解度变化。该流程从数据集中的三维蛋白质结构开始，并使用 Jackal 软件[56]生成相应的突变体结构。特征生成过程包括三个组成部分：从预训练蛋白质 Transformer 中提取的序列特征、来自持久层拉普拉斯算子的拓扑特征，以及辅助理化特征。对于拓扑特征，我们从野生型和突变体蛋白质中提取突变位点周围的原子子集，以构建元素特异性子复形。随后利用这些子复形，在结构滤过下计算层拉普拉斯算子的调和谱和非调和谱，生成层化拉普拉斯嵌入，以表征多个尺度上的原子间相互作用。对于序列特征，我们从复合物中提取野生型和突变体蛋白质的 FASTA序列，并输入预训练 Transformer模型。所得的潜在空间嵌入用作序列特征。对于理化特征，我们考虑原子级性质，例如部分电荷、电

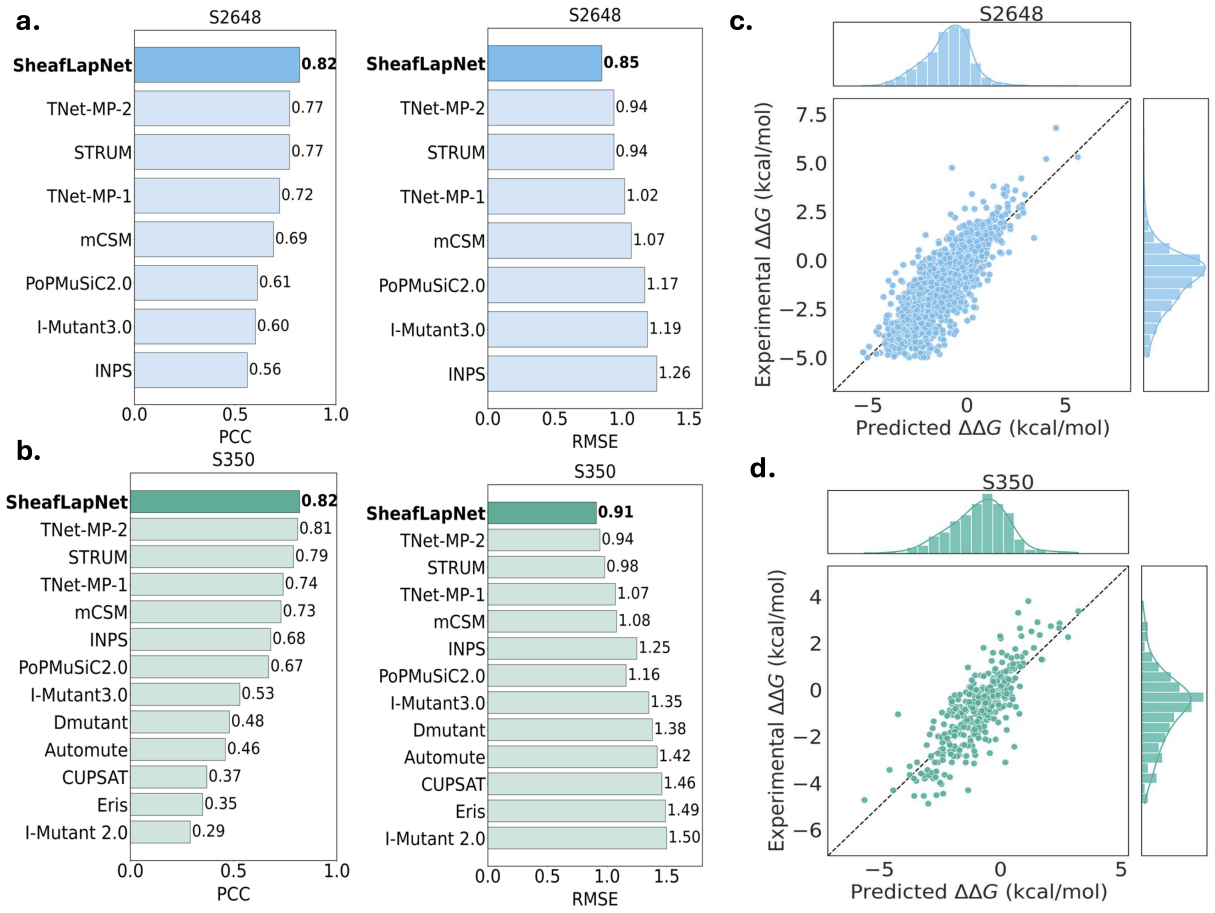


图 2： 突变后蛋白质稳定性变化的模型性能示意图。(a). SheafLapNet 在 S2648 数据集上的五折交叉验证性能与现有最先进模型的对比 [4, 37, 55]。(b). SheafLapNet 在 S350 数据集上的盲测性能与现有最先进模型的对比 [4, 6, 37, 55]。(c). S2648 数据集中实验测得的蛋白质稳定性变化与 SheafLapNet 预测结果的对比。(d). S350 数据集中实验测得的蛋白质稳定性变化与 SheafLapNet 预测结果的对比。

静电溶剂化自由能和库仑相互作用，以及残基级属性，例如突变位点邻域氨基酸组成、pKa 位移和其他理化性质。这三类特征嵌入拼接后形成输入特征向量，供机器学习算法（神经网络）预测突变引起的稳定性变化和溶解度变化。

2.2 突变引起的蛋白质稳定性变化预测

蛋白质稳定性中的突变诱导的扰动是许多遗传性疾病的一个基本机制。这些稳定性变化以 $\Delta\Delta G$ 量化，往往会损害蛋白质功能和结构完整性。为了严格评估我们的框架捕捉这些影响的能力，我们使用了 S2648 基准数据集。该数据集包含 131 种蛋白质结构中的 2,648 个单点突变，并标注了通过实验测定的稳定性变化。我们的评估方案分为两个不同阶段，旨在确保稳健

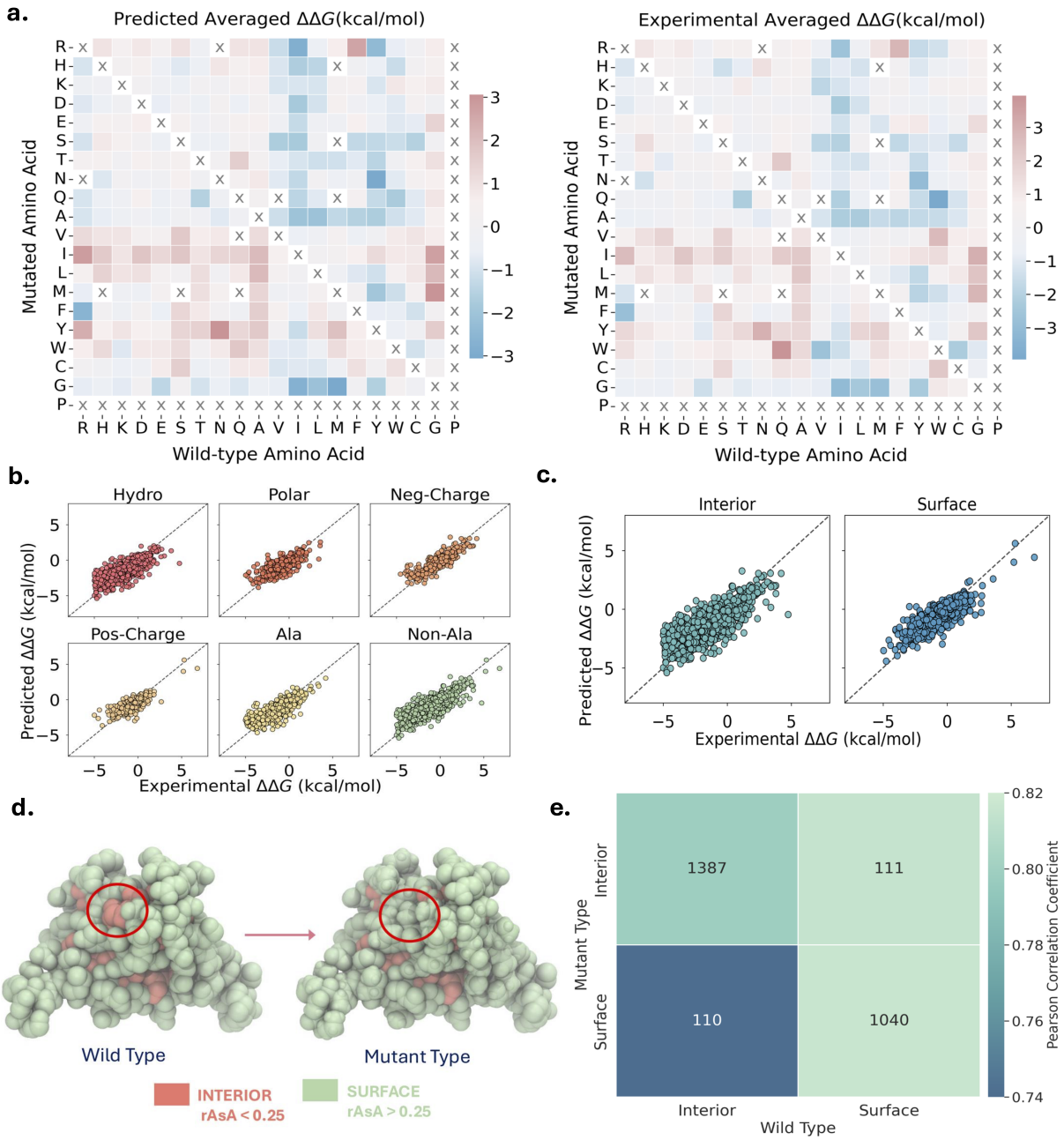


图 3： S2648 数据集中突变引起的蛋白质稳定性变化模型性能示意图。(a) 残基-残基矩阵，对比整个数据集中 2648 个突变的实验平均值与预测的突变诱导的稳定性变化 $\Delta\Delta G$，其中 X 表示无突变样本。(b) 基于残基理化性质的不同突变类型下的模型性能。(c) 按相对可及溶剂面积（rASA）定义的不同结构区域中的模型性能。(d) Q39G突变使蛋白质（PDB编号：1ARR）的结构从内部区域转移至表面区域。(e) 按四种突变区域类型分层展示的模型性能，单元格注释标示各类别的样本数量。

验证。首先，我们在完整的 S2648 数据集上进行了五折交叉验证，以确立泛化能力基准。随后，我们使用 S350 数据集进行了有针对性的盲测。该数据集是从 S2648 中筛选并整理出的高质量子集，包含来自 67 种蛋白质的 350 个突变。该子集是独立评估模型的标准基准。为便于与文献中已确立的方法直接比较， [4, 37, 55], 我们按照补充信息第 S1.1 节中的定义，使用皮尔逊相关系数（PCC）和均方根误差（RMSE）量化预测性能。

Sheaf Laplacian 框架的预测效能，得到了 SheafLapNet 在两种验证方案中优异表现的有力证明。在全面的 S2648 交叉验证分析中，我们将模型与现有最先进方法进行了基准比较，如图 2(a) 所示。SheafLapNet 的表现显著优于已确立的基线方法，包括 STRUM [37]以及领先的拓扑卷积神经网络 TNet-MP-2 [4]。这些方法的 PCC 约为 0.77，RMSE 为 0.94 千卡/摩尔；相比之下，我们的模型展现出更强的预测能力，PCC 达到 0.82，RMSE 为 0.85 千卡/摩尔。相较于 TNet-MP-2，皮尔逊相关系数提高了 6.49%，预测误差降低了 9.6%。如图 2(c) 所示，散点图显示，预测的稳定性变化与实验测量结果高度一致，验证了 PSL 嵌入在刻画突变诱导的结构变化方面的有效性。为进一步评估模型的稳健性，我们在 S350 基准上进行了盲测。在这项严格的评估中，SheafLapNet 保持了较高的预测准确率，平均 PCC 达到 0.82，RMSE 为 0.91 千卡/摩尔。如图 2(b) 的比较分析所示，我们的模型优于 TNet-MP-2；后者的 PCC 为 0.81，RMSE 为 0.94 千卡/摩尔。图 2(d) 展示了测试集实验值与预测 $\Delta\Delta G$ 值的直接比较。除预测准确率外，PSL 框架在数据覆盖范围方面也表现出更强的稳健性。例如，I-Mutant 3.0 未能评估完整数据集，在 S2648 基准中仅覆盖 2,648 个样本中的 2,636 个，在 S350 中仅覆盖 350 个样本中的 338 个。相比之下，SheafLapNet 成功为两个基准中的每个突变样本生成了预测。如此全面的覆盖凸显了层拉普拉斯嵌入的可扩展性，也体现了其计算可靠性。

利用 S2648 数据集上五折交叉验证所得的预测结果，我们评估了不同突变类型下实验稳定性变化平均值与预测稳定性变化平均值之间的一致性。如图 3(a) 所示，我们构建了一个残基-残基矩阵，其中横轴表示野生型残基，纵轴表示突变残基。特定突变体所在行与野生型残基所在列的交叉单元格，表示该替换对的平均 $\Delta\Delta G$ 值；每种替换对应的突变样本数见图 S1。为呈现完整的稳定性景观，我们明确纳入了反向突变，并为其赋予相反的稳定性变化值，从而形成反对称矩阵结构。预测矩阵与实验基准中观察到的模式高度相似，表明该模型成功捕捉了突变特异性稳定性趋势。尽管总体扰动景观得以保留，但与实验数据相比，预测值的方差似乎有所减小。这项分析揭示的一个显著特征是，相对于大多数野生型残基，突变为异亮氨酸始终会产生正的 $\Delta\Delta G$ 值。一种可能的解释是，异亮氨酸往往无法弥补有利亲水相互作用的损失，例如原本由极性或带电野生型残基提供的氢键和静电接触。因此，异亮氨酸与周围水相

溶剂缺乏化学相容性，进而导致值为正 $\Delta\Delta G$ 。

为进一步评估模型对特定理化性质的敏感性，我们将突变残基分为六类：疏水性、极性、带负电、带正电、丙氨酸和非丙氨酸。图3(b)展示了各类别中预测的稳定性变化与实验结果之间的对应关系。定量分析表明，模型在所有组别中均具有稳定的预测性能。具体而言，疏水性、极性、带负电、带正电、丙氨酸和非丙氨酸突变的皮尔逊相关系数（均方根误差）分别为0.802（0.890 千卡/摩尔）、0.750（0.816 千卡/摩尔）、0.853（0.768 千卡/摩尔）、0.798（0.754 千卡/摩尔）、0.815（0.859 千卡/摩尔）和 0.820（0.832 千卡/摩尔）。值得注意的是，模型在所有突变类型中均取得了稳健的结果，表明其能够有效泛化至各种氨基酸替换，且未表现出明显偏差。在这些不同化学类别中始终观察到较高相关性，这表明底层特征表示能够有效捕捉与各类特定残基相关的稳定作用力和去稳定作用力。

突变位点被划分为两种不同的结构环境：蛋白质内部和溶剂暴露表面。该分类依据相对可及溶剂面积（rAsA）确定，阈值为0.25。rAsA < 0.25 的残基归为内部区域，而rAsA ≥ 0.25 的残基则归为表面区域。从生物学角度看，发生在蛋白质内部的突变通常比表面突变造成更大程度的稳定性降低。在我们的数据集中，内部残基的实验平均 $\Delta\Delta G$ 为-1.392 千卡/摩尔，而表面残基为 -0.522 千卡/摩尔。这一趋势证实，埋藏区域的破坏性影响更大，这是因为突变会扰乱对蛋白质稳定性至关重要、排列紧密的疏水核心。图3(c)展示了模型在这些结构区域中的预测性能。分析显示，模型在两种环境中都具有较强的相关性：内部区域的皮尔逊相关系数为0.796，均方根误差为0.964千卡/摩尔；表面区域的表现略优，皮尔逊相关系数为0.819，均方根误差为0.614千卡/摩尔。表面区域的误差较低，可能反映了溶剂暴露位点的稳定性变化通常方差较小。为直观展示这些结构定义，图3(d)呈现了Q39G突变引起的蛋白质1ARR结构变化。在此例中，第39位残基从野生型内部区域的谷氨酰胺转变为突变体表面区域的甘氨酸。通过将野生型和突变体结构中的残基分别归类为内部区域或表面区域，我们可以考察突变后氨基酸暴露程度连续变化对稳定性变化的影响。图3(e)展示了四种特定突变轨迹的性能指标：内部到内部、内部到表面、表面到内部以及表面到表面。模型在大多数类别中均表现稳健，内部到内部、表面到内部和表面到表面类别的皮尔逊相关系数分别为0.80、0.81和0.82。内部到表面类别的相关性略低，皮尔逊相关系数为0.74，这可能是由于该突变类型的样本量显著较小。

2.3 突变诱导的蛋白质溶解度变化分类

由基因突变引起的蛋白质溶解度变化是决定蛋白质稳定性和功能的关键因素，通常会导致聚集相关疾病或生物活性改变。为模拟这些影响，我们采用了最初为 PON-Sol2 [58] 整理的数据集。该数据集包含来自 77 种不同蛋白质的 6,328 个突变样本。根据突变对溶解度的影响，将其分为三类：降低、增加和中性

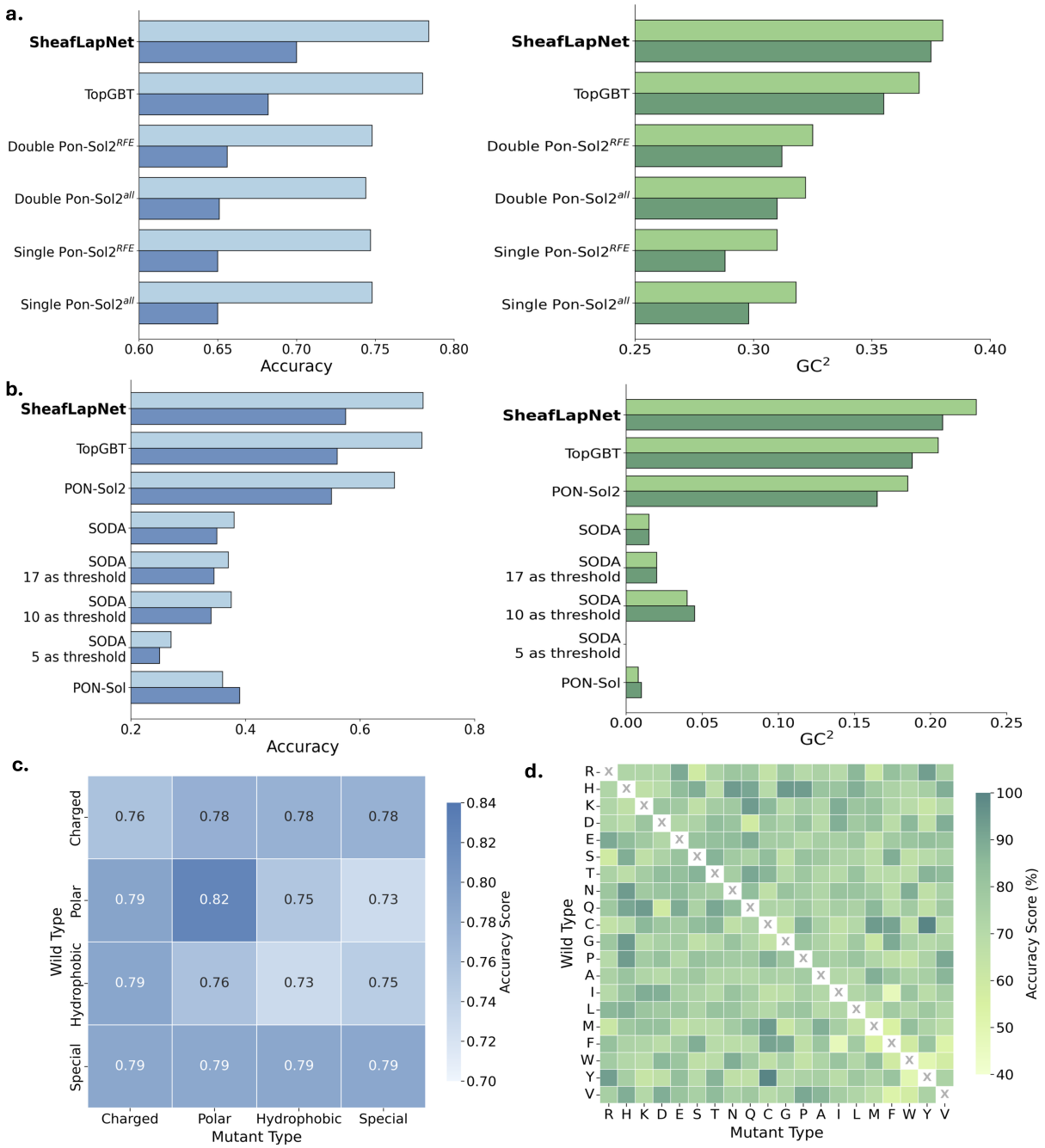


图 4：模型对突变引起的蛋白质溶解度变化进行分类时的性能示意图。浅蓝色表示准确率（CPR），深蓝色表示归一化准确率，浅绿色表示广义平方相关性（GC^2），深绿色表示归一化 GC^2。(a) 在 PON-Sol2 数据集上通过 10 折交叉验证得到的性能，对比 SheafLapNet 与现有模型 [50, 58]。对于现有的 PON-Sol2 模型 [58]，，RFE 指递归特征消除，all 指使用全部特征。(b) 在 PON-Sol2 数据集上的独立盲测结果，对比 SheafLapNet 与最先进模型 [50, 57, 58]。(c) 按理化性质突变组分层的 10 折交叉验证模型准确率。(d) 按特定氨基酸替换类型分析的 10 折交叉验证模型准确率，其中 X 表示未发生突变的样本。

（无变化）。这些突变样本的分布和分类见图 S2。具体而言，数据集包含 3,136 个溶解度降低的样本、1,026 个溶解度增加的样本，以及 2,166 个溶解度无变化的样本。这一分布显示类别存在明显不平衡，比例约为 1 : 0.69 : 0.33（降低 : 中性 : 增加），表明降低蛋白质溶解度的突变占比较高。

为严格评估我们框架的预测能力，我们采用了双重验证策略，包括随机 10 折交叉验证和独立盲测分类。在交叉验证阶段，我们将 SheafLapNet 与 TopGBT[50]，（一种基于持久同调的模型）以及现有的 PON-Sol2 预测模型[58]，进行基准比较；后者采用递归特征消除（RFE）等特征选择技术。为减轻固有类别不平衡的影响，我们采用归一化准确率分数作为全面评估性能的主要指标。鉴于这是一个具有三种不同溶解度类别的 K 类问题，我们采用补充信息 S1.2 节中定义的正确预测率（CPR）和广义平方相关系数（GC^2）进行全面评估。具体而言，CPR 衡量模型的整体准确率，而 GC^2 量化分类的相关系数，取值范围为 0 到 1。这些指标的值越大，表示性能越好。

所提出的模型 SheafLapNet 在 10 折交叉验证设置下，归一化 CPR 和 GC^2 得分分别达到 0.700 和 0.375。这些结果分别比现有 PON-Sol2 模型最高提升 6.71% 和 20.19%，比基于持久同调的 TopGBT 模型最高提升 2.64% 和 5.63%。鉴于该任务具有多类别特性，这些指标能够全面评估预测性能，如图 4(a) 所示。通过独立盲测进一步证实了层拉普拉斯框架的稳健性。在此次评估中，SheafLapNet 保持了稳定的预测能力，归一化 CPR 和归一化 GC^2 分别达到 0.570 和 0.201。这些得分分别比 PON-Sol2 模型最高高出 3.64% 和 21.82%，比 TopGBT 高出 1.79% 和 6.91%。基于 CPR 和 GC^2 得分的基准比较结果见图 4(b)。

转而关注突变类型后，我们的模型在 20 种不同氨基酸类型中对溶解度变化进行分类的能力同样值得探究。除此之外，我们还将氨基酸分为带电极性、疏水性和特殊情况三类。图 4(c) 展示了每种突变组对的准确率得分，而图 4(d) 显示了每种氨基酸对的得分。在此表示中，纵轴标示原始蛋白质的残基类型，横轴标示突变体的残基类型。对于反向突变，除非变化为零，否则标签采用反向溶解度变化。分析表明，不同突变类型之间的预测性能存在明显差异。值得注意的是，从极性残基到极性基团的突变具有最高的分类准确率。这可能是由于表面极性与溶解度之间存在直接关联，从而产生可预测的溶剂化能变化，并被模型的拓扑特征有效捕捉。相反，从极性残基到特殊残基的转变，或发生在疏水性组内的转变，其性能相对较低。准确率降低可能反映了对复杂结构约束进行建模的困难，例如特殊残基引入的骨架刚性，以及疏水核心所特有的细微堆积相互作用。

3 材料与方法

本节详细介绍了从蛋白质结构中提取的三种不同特征集，用于我们的分析：源自持久层拉普拉斯算子的拓扑特征、理化辅助描述符以及进化序列嵌入。本研究所使用软件包的完整汇总见补充信息的 S4 节，而辅助描述符的详细列表见 S3 节。

3.1 持久层拉普拉斯算子

在介绍持久层拉普拉斯算子之前，我们首先回顾标准的持久拉普拉斯算子。构造从考虑两个单纯复形 K 和 L 开始，其中 $K \subseteq L$。我们分别将 K 和 L 的实系数单纯链复形记为 C^K 和 C^L。由于单纯链复形中的链群 C_q 由单形形式生成，因此其自然构成有限维内积空间，从而确保边界映射 ∂_q 的伴随算子定义良好。

我们将 C^L_{q+1} 的一个子空间定义为 $C^{L,\ K}_{\ q+1}$，其为集合 $\{c \in C^L_{q+1} \mid \partial^L_{q+1}(c) \in C^K_q\}$。令 $\partial^{L,\ K}_{\ q+1}$ 表示边界映射 ∂^L_{q+1} 在该子空间 $C^{L,\ K}_{\ q+1}$ 上的限制。q 阶持久拉普拉斯算子 $\Delta^{L,\ K}_{\ q}$ 形式上定义为以下算子：

$$\Delta^{L,K}_q = \partial^{L,K}_{q+1}(\partial^{L,K}_{q+1})^* + (\partial^K_q)^*\partial^K_q. \tag{1}$$

为了将这一框架扩展到持久层拉普拉斯算子，我们必须首先定义胞腔层。胞腔层 $\mathscr{S}$ 由一个单纯复形 X （视为胞腔复形）构成，并为每个胞腔 σ 指定一个 X 有限维向量空间 $\mathscr{S}(\sigma)$，称为 $\mathscr{S}$ 在 σ上的茎。此外，对于每个面关系 $\sigma \preceq \tau$（其中 $\sigma \subset \bar{\tau}$），都存在一个向量空间的线性态射，记作 $\mathscr{S}_{\sigma\preceq\tau}$，称为限制映射。这些映射满足以下复合规则：

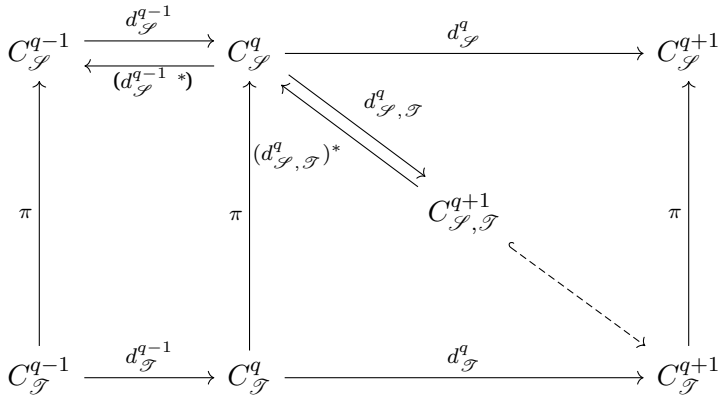
$$\mathscr{S}_{\rho\preceq\tau} = \mathscr{S}_{\sigma\preceq\tau}\mathscr{S}_{\rho\preceq\sigma} \tag{2}$$

对于任意 $\rho \preceq \sigma \preceq \tau$，其中 $\mathscr{S}_{\sigma\preceq\sigma}$ 是 $\mathscr{S}(\sigma)$的恒等映射。

与单纯复形类似，胞腔层会生成一个层上链复形。第q个层上链群 $C^q_{\mathscr{S}}$ 定义为 q维胞腔上的茎的直和。通过对单纯复形 X进行全局定向，从而为每个 $\sigma \preceq \tau$获得带符号关联关系 $[\sigma\colon \tau]$，余边界映射 $d^q\colon C^q_{\mathscr{S}} \to C^{q+1}_{\mathscr{S}}$定义如下：

$$d^q|_{\mathscr{S}(\sigma)} = \sum_{\sigma\preceq\tau}[\sigma : \tau]\mathscr{S}_{\sigma\preceq\tau}. \tag{3}$$

持久层拉普拉斯算子通过考虑两个胞腔层构造而成，$\mathscr{S}$ 在 K上，$\mathscr{T}$ 在 L上，且 $K \subseteq L$，并且 K 的茎和限制映射与L的相同。这些层的上链复形与持久谱结构之间的关系如下列交换图所示：



(4)

我们将子空间$C_{\mathscr{S}}^{q+1}{}_{\mathscr{T}}$定义为$C_{\mathscr{S}}^{q+1}{}_{,\mathscr{T}} = \{c \in C_{\mathscr{T}}^{q+1} \mid (d_{\mathscr{T}}^{q})^*(c) \in C_{\mathscr{S}}^{q}\}$。令$d_{\mathscr{S},\mathscr{T}}^{q}$表示限制算子$(d_{\mathscr{T}}^{q})^*|_{C_{\mathscr{S}}^{q+1}}$的伴随映射，$\mathscr{T}$。第 q个 PSL 记作 $\Delta^{\mathscr{S},}{}_{q}{}^{\mathscr{T}}$，其定义如下：

$$\Delta_q^{\mathscr{S},\mathscr{T}} = (d_{\mathscr{S},\mathscr{T}}^{q})^* d_{\mathscr{S},\mathscr{T}}^{q} + d_{\mathscr{S}}^{q-1}(d_{\mathscr{S}}^{q-1})^*. \tag{5}$$

PSL 的谱性质提供了关键的拓扑不变量。具体而言，该算子的零特征空间（核）的维数对应于第 q 个持久层贝蒂数，记作 $\beta^{\mathscr{S},}{}_{q}{}^{\mathscr{T}}$其中 $\beta^{\mathscr{S},}{}_{q}{}^{\mathscr{T}} = \dim(\ker \Delta^{\mathscr{S},}{}_{q}{}^{\mathscr{T}})$。

本研究采用在带标记单纯复形上构造的胞腔层，利用原子部分电荷对蛋白质结构进行建模，方法遵循现有文献 [53] 中确立的方案。首先定义一个通用框架，用于在带标记单纯复形 X 上构造层，其中每个顶点都与一个量 q 相关联。设 $F\colon X \to \mathbb{R}$ 是一个处处非零函数。该层通过为每个单形指定茎 $\mathbb{R}$ 来构造。对于一个面关系 $[v_0,\dots,v_n] \le [v_0,\dots,v_n,v_{n+1},\dots,v_m]$（方向无关紧要），线性态射 $\mathscr{S}([v_0,\dots,v_n] \le [v_0,\dots,v_n,v_{n+1},\dots,v_m])$ 定义为以下标量乘法：

$$\frac{F([v_0,\dots,v_n])q_{n+1}\cdots q_m}{F([v_0,\dots,v_n,v_{n+1},\dots,v_m])}. \tag{6}$$

为将这一框架用于蛋白质分析，我们引入非几何信息，采用从 PDB2PQR 软件包 [24]获取的原子部分电荷。我们构建图的 Rips 或 Alpha 过滤，其中顶点 v_i 表示原子，边e_{ij} 表示原子 v_i 之间的相互作用，v_j亦然。胞腔层的定义方式是：每个茎都是实数直线 $\mathbb{R}$。对于特定的面关系 $v_i \preceq e_{ij}$ （其中顶点 v_i 是边e_{ij}的面），限制态射被明确定义为乘以 $\frac{q_j}{r_{ij}}$，其中q_j 是相邻原子 v_j的部分电荷，而r_{ij} 表示边e_{ij}的欧几里得长度。由此得到的 PSL 的调和谱揭示了拓扑不变量，而非调和谱则代表数据的几何信息。因此，这些谱被用作 SheafLapNet 模型的输入特征。

为阐释持久层拉普拉斯算子框架，我们使用一个包含 12 个点的点云数据集作为示例，如图 5 所示。拓扑结构通过 Vietoris-Rips （ VR ）复形滤过进行建模，并在图 5(a) 中加以可视化。此处，异质信息被明确映射到定义域上，为节点赋予 1 和 0.01 两种不同的“电荷”值。随着过滤参数增加，点云的连通性不断演化，生成一系列嵌套单纯复形，为我们的层理论分析提供定义域

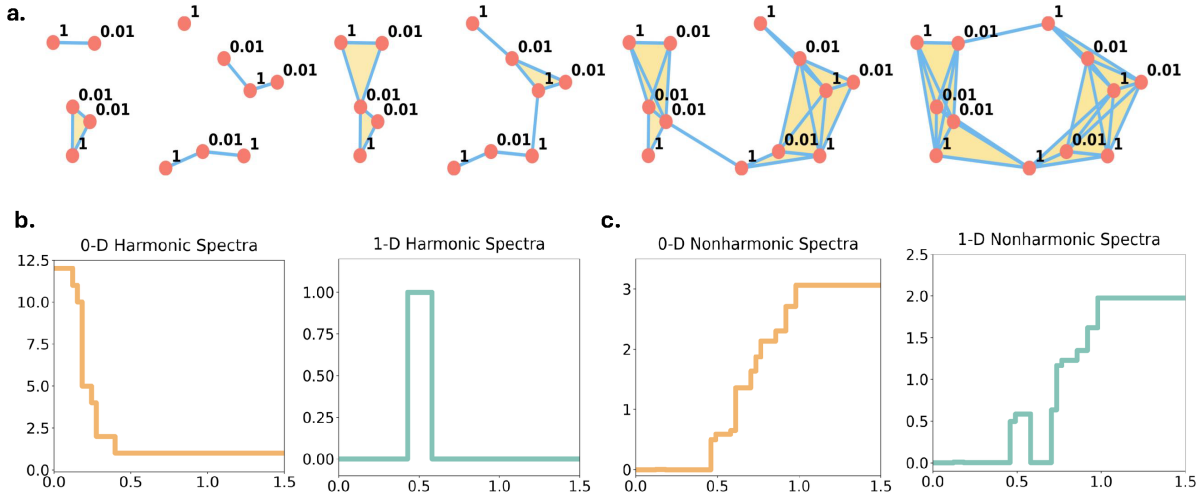


图 5：持久层拉普拉斯算子的示意图。(a)：从点云数据构建 Rips 复形的滤过过程。(b)：0 维和 1 维层拉普拉斯矩阵中零特征值的持久重数。(c)：零维和一维层拉普拉斯矩阵中非零特征值的最小值。

。图 5(b) 展示了由持久层拉普拉斯算子导出的相应调和谱分析，具体绘制了零特征值的持久重数。这些重数对应于层贝蒂数，通过纳入原子部分电荷等异质信息，对标准拓扑不变量进行了推广。作为对这一调和分析的补充，图 5(c) 展示了非调和谱信息，具体呈现了最小非零特征值的演化。这些数值量化了连通强度和局部几何刚性，为层中编码的内在物理和化学相互作用提供了连续度量。

3.2 蛋白质层拉普拉斯特征生成

在 SheafLapNet 架构中，蛋白质结构被建模为一组单纯复形，并据此计算持久层拉普拉斯算子以提取拓扑特征。为兼顾计算效率与捕获关键物理相互作用的能力，该模型采用元素特异性和位点特异性原子子集策略。三维蛋白质结构中的原子被划分为突变位点原子（记为 A_m）以及位于截断半径 r范围内的突变邻域原子（记为 A_{mn}（r））。这一划分同时应用于野生型和突变体结构。借助元素特异性表示框架 [5], ，模型聚焦于涉及碳（C）、氮（N）和氧（O）原子的相互作用。通过考察突变位点与其邻域中这些元素类型的交集，可构建九种不同的成对原子组合。这些组合表征特定的相互作用类型。例如，突变位点中的碳原子（$A_C \cap A_m$）与邻域中的碳原子（$A_C \cap A_{mn}$（r））配对，可捕捉疏水性 C-C 相互作用；而涉及碳和氧的组合（$A_C \cap A_m$ 和 $A_O \cap A_{mn}$（r））则表征极性 C-O 相互作用。随后，这些元素特异性和位点特异性集合构成多尺度层拉普拉斯嵌入的基础。

为了准确捕捉所定义原子集合内的拓扑相互作用，我们构建

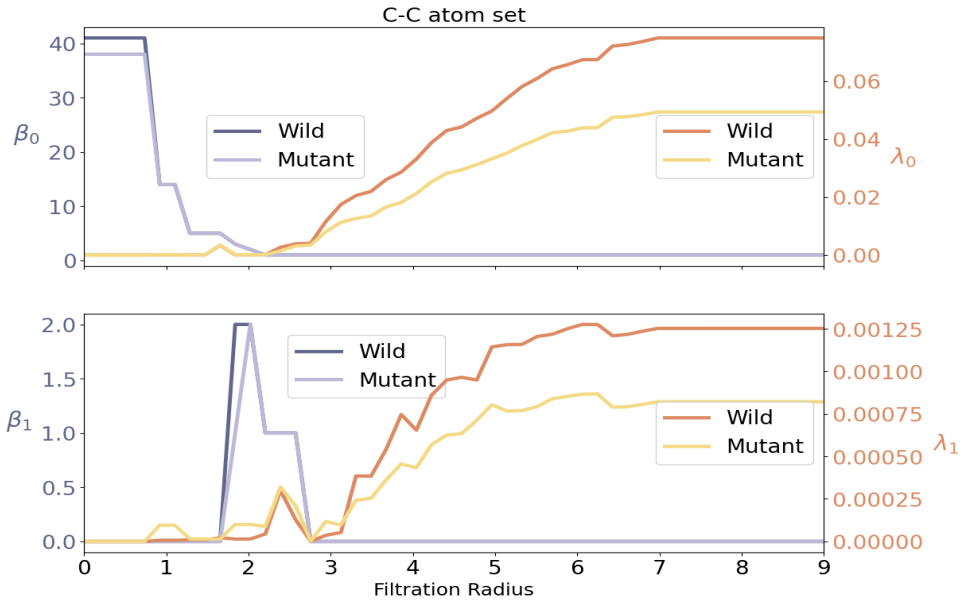


图6： 在 S2648 数据集的蛋白质 1A5E 中，第 37 位残基从 L 突变为 S 的突变位点处，使用修正距离过滤从 VR 复形中提取 C-C 相互作用的 PSL 描述符示意图。上图：零维 PSL 描述符示意图。下图：一维 PSL 描述符示意图。左侧坐标轴表示层贝蒂数 β_0 和 β_1。右侧坐标轴表示持久层拉普拉斯算子的最小非零特征值 λ_0 和 λ_1 。

VR 复形 [46] 和 Alpha 复形 [15]。构建 VR 复形时，我们采用修正距离函数 d，专门用于分离不同原子集合之间的界面相互作用（例如，突变位点 A_m 与邻域 $A_{mn}(r)$之间的相互作用）。修正后的度量定义如下：

$$d(a_i, a_j) = \begin{cases} E_d(a_i, a_j) & \text{if } (a_i \in A_m \wedge a_j \in A_{mn}(r)) \vee (a_i \in A_{mn}(r) \wedge a_j \in A_m), \\ \infty & \text{otherwise,} \end{cases} \tag{7}$$

其中， $E_d(a_i, a_j)$表示原子 a_i 和 a_j之间的欧几里得距离。通过将属于同一集合的原子对之间的距离设为无穷大，该度量有效地过滤掉集合内边，从而使拓扑分析仅关注突变位点与其周围环境之间的相互作用。此外，使用相同原子集合上的标准欧几里得距离构建 Alpha 复形 [15]，以捕捉局部几何约束。

从构建的复形中提取谱特征，作为局部 PSL 描述符。本研究采用距突变位点 16Å 的截断距离来识别突变邻域原子。对于 VR 复形，滤过范围为 3Å 至 9Å，步长为 1Å。在每个滤过步骤中，均根据持久层拉普拉斯算子推导谱性质。对于调和分量，记录零特征值的数量，从而为每个原子集合生成一个 7 维特征向量。对于非调和分量，则从非零特征值谱中提取八项统计性质：最大值、最小值、均值、总和、标准差、方差以及特征值数量。此过程为每个原子集合生成一个 72 维特征向量。同样，对于

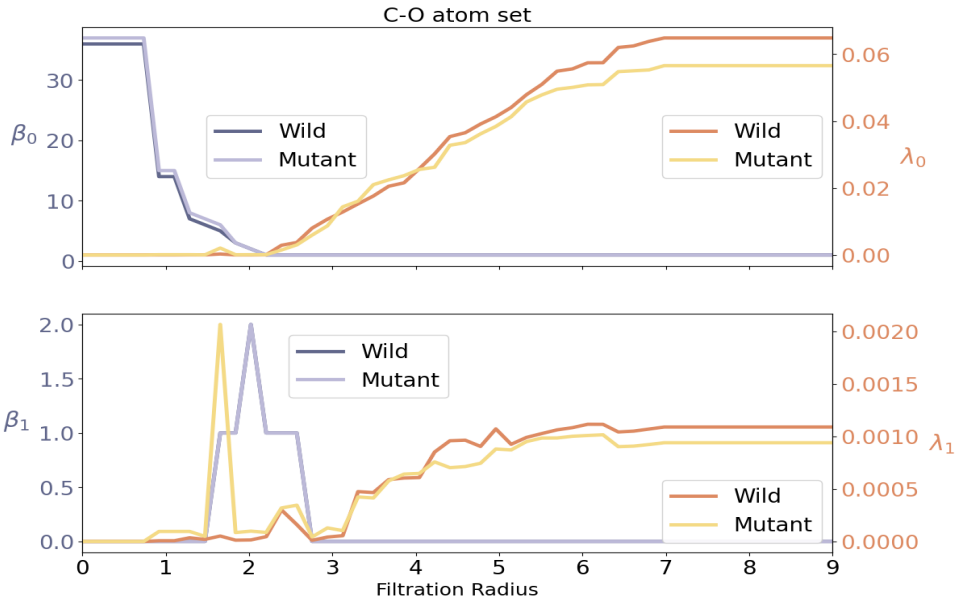


图 7： S2648 数据集中蛋白质 1A5E 的第 37 位残基突变位点由 L 突变为 S 时，使用修正距离过滤从 VR 复形中提取 C-O 相互作用的 PSL 描述符示意图。上图：零维 PSL 描述符示意图。下图：一维 PSL 描述符示意图。左侧纵轴表示层贝蒂数 β_0 和 β_1。右侧纵轴表示持久层拉普拉斯算子的最小非零特征值 λ_0 和 λ_1 。

Alpha 复形采用相同的统计提取方法分析持久层拉普拉斯算子的 一维调和分量和非调和分量。最终的拓扑特征表示由零维 VR 模型和一维 Alpha 模型提取的特征组合而成。分别为野生型结构、突变体结构及二者之差计算这些特征向量。将这些向量串联后，即得到单个蛋白质的综合特征表示，其维度为 3402。为展示这些特征的判别能力，图 6 和图 7 展示了 S2648 数据集中的蛋白质 1A5E 的野生型和突变体结构的一维和二维 PSL 描述符。图 6 使用原子集合 $A_C \cap A_m$ 和 $A_C \cap A_{mn}$ （r），通过经修正的基于距离 d 的滤过生成 VR 复形，从而有效揭示疏水性 C-C 相互作用。同样，图 7 使用原子集合 $A_C \cap A_m$ 和 $A_O \cap A_{mn}$（r），通过基于 d 的滤过生成 VR 复形，从而揭示极性 C-O 相互作用。

3.3 辅助特征

除了拓扑特征外，我们的模型还整合了一套全面的理化描述符，涵盖原子层面和残基层面，用于表征突变环境。在原子层面，我们计算溶剂排斥表面积、部分原子电荷、库仑相互作用能和范德华相互作用能，以及静电溶剂化自由能 [8, 24, 39]。其中，静电溶剂化自由能通过泊松–玻尔兹曼模型的 MIBPB 求解器计算 [8]。在残基层面，特征集包括局部邻域的氨基酸组成和物理性质、突变诱导的 pKa 位移、进化

保守性评分（由位置特异性评分矩阵（PSSM）得出），以及预测的二级结构参数。详细定义和计算流程见《补充信息》第 S3 节。

3.4 序列特征

蛋白质大语言模型（pLLM），例如 ESM（进化尺度建模） [38] 和 ProtTrans [16], ，已展现出卓越的能力，能够从数亿条未经注释的序列中捕捉深层进化模式。为将这一丰富的进化背景纳入我们的代数框架，我们采用了 ESM-2 Transformer 架构 [27]。该架构通过在大型蛋白质序列数据库上进行自监督掩码语言建模（MLM）训练，包含 33 个 Transformer 层和约 6.5 亿个参数。对于给定的输入序列，模型会生成维度为 1,280 的上下文化词元嵌入。我们从最后一层提取这些表征，并沿序列长度计算其均值，从而得到一个固定长度的 1,280 维向量。为明确捕捉遗传变异的影响，我们分别为野生型序列和突变序列生成这些嵌入，随后将它们拼接起来，形成一个最终的 2,560 维进化特征向量，用于我们的预测模型。

3.5 SheafLapNet 模型超参数

SheafLapNet 是一种深度神经网络，旨在通过处理由提取的 PSL 特征、Transformer 嵌入和辅助向量拼接而成的向量来处理更大的数据集。该网络架构由六个全连接隐藏层组成，每层包含 15,000 个神经元。为引入非线性并增强表征能力，每个隐藏层后均应用修正线性单元（ReLU）激活函数。网络中采用 Dropout 正则化，以防止过拟合并确保良好的泛化能力。最终输出层取决于任务类型：对于稳定性预测中的回归任务，输出连续标量值；对于溶解度预测中的分类任务，则输出离散的类别概率。模型训练进行了 200 个训练轮次，学习率固定为 0.001。针对每种任务类型分别优化批量大小：基于回归的蛋白质稳定性数据集采用 32 的批量大小，而基于分类的蛋白质溶解度数据集采用 50 的批量大小。

4 结论

准确预测基因突变对蛋白质稳定性和溶解度的影响，对于理解疾病的分子基础和推进精准医学至关重要。然而，对这些广泛存在的变异进行实验表征需要投入大量资源，而现有计算模型通常存在可解释性有限的问题，或将稳定性和溶解度视为彼此割裂的任务。因此，开发能够捕捉内在分子相互作用的统一、严谨的预测框架已成为一项重要且迫切的任务。为应对这一挑战，我们提出了 SheafLapNet——一种新型拓扑深度学习（TDL）框架，利用持久层拉普拉斯算子（PSL）预测突变诱导的蛋白质稳定性和溶解度变化。通过将 PSL 编码的异质几何信息与辅助理化描述符以及从预训练 ESM-2 Transformer 中提取的进化序列嵌入相结合，我们的模型实现了数学基础扎实且稳健的表征

分子改变。该整合方法捕捉了突变影响的多尺度和机制层面细微特征，使预测准确率在四个严格的基准数据集上均得到最先进水平的提升。

尽管本研究通过 PSL 证实了将异质原子信息嵌入单纯复形的有效性，但快速发展的拓扑数据分析领域仍为进一步改进提供了机遇。未来的研究可以探索更多独特的数学不变量，例如持久路径拉普拉斯算子 [48], 持久超有向图拉普拉斯算子 [9], 持久狄拉克算子 [1, 44] 以及层神经网络或余层神经网络 [2, 20]。这些先进工具或许能够更深入地揭示蛋白质结构中的有向相互作用和高阶相互作用，而这些内容可能无法仅通过单纯复形得到充分捕捉，从而有望实现更高水平的预测精度。

数据与代码可用性

本研究中的数据和代码可在 `https://github.com/yren24/SheafLapNet/tree/main`中获取

致谢

本研究部分得到了美国国立卫生研究院资助（项目编号 R35GM148196）、美国国家科学基金会资助（项目编号 DMS2052983）以及密歇根州立大学研究基金会的支持。

参考文献

[1] B. Ameneyro、V. Maroulas 和 G. Siopsis。量子持久同调。《应用与计算拓扑学杂志》，8(7):1961–1980，2024年。[2] F. Barbero、C. Bodnar、H. S. de Ocáriz Borde、M. Bronstein、P. Veličković 和 P. Liò。结合连接拉普拉斯算子的层神经网络。载于拓扑、代数与几何学习研讨会 *2022*，第 28–36 页。PMLR，2022年。[3] Z. Cang、L. Mu 和 G.-W. Wei。机器学习评分与虚拟筛选中生物分子代数拓扑的可表示性。《*PLoS* 计算生物学》，14(1):e1005929，2018年。[4] Z. Cang 和 G.-W. Wei。TopologyNet：用于生物分子性质预测的基于拓扑的深度卷积神经网络与多任务神经网络。《*PLoS* 计算生物学》，13(7):e1005690，2017年。[5] Z. Cang 和 G.-W. Wei。将元素特异性持久同调与机器学习相结合，用于蛋白质-配体结合亲和力预测。《生物医学工程数值方法国际期刊》，34(2):e2914，2018年。[6] E. Capriotti、P. Fariselli 和 R. Casadio。I-mutant2.0：根据蛋白质序列或结构预测突变引起的稳定性变化。《核酸研究》，33(suppl_2):W306–W310，2005年。[7] G. Carlsson。拓扑与数据。《美国数学学会通报》，46(2):255–308，2009年。[8] D. Chen、Z. Chen、C. Chen、W. Geng 和 G.-W. Wei。MIBPB：用于静电分析的软件包。《计算化学杂志》，32(4):756–770，2011年。[9] D. Chen、J. Liu、J. Wu 和 G.-W. Wei。持久超有向图同调与持久超有向图拉普拉斯算子。《数据科学基础》（密苏里州斯普林菲尔德），5(4):558，2023年。[10] J. Chen、Y. Qiu、R. Wang 和 G.-W. Wei。持久拉普拉斯算子预测奥密克戎 BA.4 和 BA.5 将成为新的优势变体。《生物学与医学中的计算机》，151:106262，2022年。[11] J. Chen、R. Wang、M. Wang 和 G.-W. Wei。突变增强了 SARS-CoV-2 的传染性。《分子生物学杂志》，432(19):5212–5226，2020年。[12] J. Chen 和 G.-W. Wei。奥密克戎 BA.2（B.1.1.529.2）：极有可能成为下一个优势变体。《物理化学快报》，13(17):3840–3849，2022年。[13] X. Chen、T. Zhang、W. Su、Z. Dou、D. Zhao、X. Jin、H. Lei、J. Wang、X. Xie、B. Cheng 等。癌症中的突变型 p53：从分子机制到治疗调控。《细胞死亡与疾病》，13(11):974，2022年。[14] J. M. Curry。层、余层及其应用。宾夕法尼亚大学，2014年。[15] H. Edelsbrunner。α 形状——综述。载于 R. van de Weygaert、G. Vegter、J. Ritzerveld 和 V. Icke 编，科学中的镶嵌：几何铺砌的优点、技术与应用，第 1–25 页。Springer，2011年。

[16] A. Elnaggar、M. Heinzinger、C. Dallago、G. Rehawi、Y. Wang、L. Jones、T. Gibbs、T. Feher、C. Angerer、M. Steinegger 等。ProtTrans：通过自监督学习理解生命语言。*IEEE*模式分析与机器智能汇刊，44(10):7112–7127，2021年。[17] C. Epstein、G. Carlsson 和 H. Edelsbrunner。拓扑数据分析。《逆问题》，27(12):120201，2011年。[18] M. Funayama、K. Nishioka、Y. Li 和 N. Hattori。帕金森病的分子遗传学：贡献与全球趋势。《人类遗传学杂志》，68(3):125–130，2023年。[19] J. Grbić、J. Wu、K. Xia 和 G.-W. Wei。数据科学中拓扑方法的若干方面。《数据科学基础》（密苏里州斯普林菲尔德），4(2):165，2022年。[20] M. Hajij、L. Bastian、S. Osentoski、H. Kabaria、J. L. Davenport、S. Dawood、B. Cherukuri、J. G. Kocheemoolayil、N. Shahmansouri、A. Lew 等。余层拓扑神经网络：一种广义深度学习框架。*arXiv*预印本 *arXiv:2505.21251*，2025年。[21] J. Hansen 和 R. Ghrist。走向胞腔层的谱理论。《应用与计算拓扑学杂志》，3(4):315–358，2019年。[22] E. M. Hurley、P. Mozolewski、R. Dobrowolski 和 J. Hsieh。家族性阿尔茨海默病相关的PSEN1突变通过增强Notch信号通路影响神经发育。《干细胞报告》，18(7):1516–1533，2023年。[23] B. Jones 和 G.-W. Wei。PETLS：持久拓扑拉普拉斯算子软件。*arXiv*预印本 *arXiv:2508.11560*，2025年。[24] E. Jurrus、D. Engel、K. Star、K. Monson、J. Brandi、L. E. Felberg、D. H. Brookes、L. Wilson、J. Chen、K. Liles 等。APBS生物分子溶剂化软件套件的改进。《蛋白质科学》，27(1):112–128，2018年。[25] K. J. Karczewski、L. C. Francioli、G. Tiao、B. B. Cummings、J. Alföldi、Q. Wang、R. L. Collins、K. M. Laricchia、A. Ganna、D. P. Birnbaum 等。从141,456名人类的变异中量化突变约束谱。《自然》，581(7809):434–443，2020年。[26] S. S. Li、Z. M. Liu、J. Li、Y. B. Ma、Z. Y. Dong、J. W. Hou、F. J. Shen、W. B. Wang、Q. M. Li 和 J. G. Su。基于自监督方法学习的几何表示预测突变引起的蛋白质稳定性变化。《*BMC*生物信息学》，25(1):282，2024年。[27] Z. Lin、H. Akin、R. Rao、B. Hie、Z. Zhu、W. Lu、N. Smetanin、R. Verkuil、O. Kabeli、Y. Shmueli 等。利用语言模型进行进化尺度的原子级蛋白质结构预测。《科学》，379(6637):1123–1130，2023年。[28] J. Liu、K.-L. Xia、J. Wu、S. S.-T. Yau 和 G.-W. Wei。生物分子拓扑：建模与分析。《数学学报（英文版）》，38(10):1901–1938，2022年。[29] X. Liu、H. Feng、J. Wu 和 K. Xia。基于Hom复形的机器学习（HCML）：预测突变引起的蛋白质-蛋白质结合亲和力变化。《化学信息与建模杂志》，62(17):3961–3969，2022年。

[30] A. J. Marian。遗传变异的临床解读与管理。*Basic toTranslational Science*，5(10):1029–1042，2020年。[31] F. Mémoli、Z. Wan 和 Y. Wang。持久拉普拉斯算子：性质、算法及其影响。*SIAM* 数据科学数学期刊，4(2):858–884，2022年。[32] Z. Meng 和 K. Xia。基于持久谱的机器学习（Perspect ML）：用于蛋白质-配体结合亲和力预测。《科学进展》，7(19):eabc5329，2021年。[33]I. Molotkov、E. R. Mardis 和 M. Artomov。理解错义突变：变异致病性预测的挑战与机遇。《疾病模型与机制》，17(12):dmm052218，2024年。[34] D. D. Nguyen、Z. Cang、K. Wu、M. Wang、Y. Cao 和 G.-W. Wei。用于D3R大挑战中姿态和结合亲和力预测与排序的数学深度学习。《计算机辅助分子设计杂志》，33(1):71–82，2019年。[35] L. Paladin、D. Piovesan 和 S. C. Tosatto。SODA：根据无序性和聚集倾向预测蛋白质溶解度。《核酸研究》，45(W1):W236–W240，2017年。[36] T. Papamarkou、T. Birdal、M. Bronstein、G. Carlsson、J. Curry、Y. Gao、M. Hajij、R. Kwitt、P. Lio、P. Di Lorenzo 等。观点：拓扑深度学习是关系学习的新前沿。《机器学习研究论文集》，235:39529，2024年。[37] L. Quan、Q. Lv 和 Y. Zhang。STRUM：基于结构预测单点突变引起的蛋白质稳定性变化。《生物信息学》，32(19):2936–2946，2016年。[38] A. Rives、J. Meier、T. Sercu、S. Goyal、Z. Lin、J. Liu、D. Guo、M. Ott、C. L. Zitnick、J. Ma 等。将无监督学习扩展到2.5亿条蛋白质序列后，生物结构与功能逐渐显现。《美国国家科学院院刊》，118(15):e2016239118，2021年。[39] W. Rocchia、E. Alexov 和 B. Honig。扩展非线性泊松-玻尔兹曼方程的适用性：多种介电常数与多价离子。《物理化学杂志 *B*》，105(28):6507–6514，2001年。[40] P. Sormanni、F. A. Aprile 和 M. Vendruscolo。CamSol 方法：通过理性设计提高蛋白质突变体的溶解度。《分子生物学杂志》，427(2):478–490，2015年。[41] S. Stefl、H. Nishi、M. Petukh、A. R. Panchenko 和 E. Alexov。致病性错义突变的分子机制。《分子生物学杂志》，425(21):3919–3936，2013年。[42] Z. Su、X. Liu、L. B. Hamdan、V. Maroulas、J. Wu、G. Carlsson 和 G.-W. Wei。超越持久同调的拓扑数据分析与拓扑深度学习：综述。《人工智能评论》，59:58，2026年。[43] S. Sun、P. Poudel、E. Alexov 和 L. Li。计算生物物理学中的静电作用及其对疾病影响的意义。《国际分子科学杂志》，23(18):10347，2022年。

[44] F. Suwayyid 和 G.-W. Wei. 有向图和超图上的路径持久狄拉克算子。《数据科学基础》（密苏里州斯普林菲尔德），6(2):124，2024年。[45] J. Van Durme、G. De Baets、R. Van der Kant、M. Ramakers、A. Ganesan、H. Wilkinson、R. Gallardo、F. Rousseau 和 J. Schymkowitz。Solubis：一种通过突变减少蛋白质聚集的网络服务器。蛋白质工程、设计与筛选，29(8):285–289，2016年。[46] L. Vietoris。关于紧空间的高阶连通性及一类保持连通性的映射。《数学年刊》，97(1):454–472，1927年。[47] R. Wang、D. D. Nguyen 和 G.-W. Wei。持久谱图。《生物医学工程数值方法国际期刊》，36(9):e3376，2020年。[48] R. Wang 和 G.-W. Wei。持久路径拉普拉斯算子。《数据科学基础》，5(1):26–55，2023年。[49] R. Wang、R. Zhao、E. Ribando-Gros、J. Chen、Y. Tong 和 G.-W. Wei。Hermes：持久谱图软件。《数据科学基础》（密苏里州斯普林菲尔德），3(1):67，2021年。[50] J. Wee、J. Chen、K. Xia 和 G.-W. Wei。整合持久拉普拉斯算子与预训练变换器，以预测突变引起的蛋白质溶解度变化。《生物学与医学中的计算机》，169:107918，2024年。[51] J. Wee 和 K. Xia。用于蛋白质-蛋白质结合亲和力预测的基于持久谱的集成学习（perspect-el）。《生物信息学简报》，23(2):bbac024，2022年。[52] X. Wei 和 G.-W. Wei。持久拉普拉斯算子谱的同伦延拓。《数据科学基础》（密苏里州斯普林菲尔德），3(4):677，2021年。[53] X. Wei 和 G.-W. Wei。持久层拉普拉斯算子。《数据科学基础》（密苏里州斯普林菲尔德），7(2):446，2025年。[54] X. Wei 和 G.-W. Wei。持久拓扑拉普拉斯算子综述。《数学》，13(2):208，2025年。[55] C. L. Worth、R. Preissner 和 T. L. Blundell。SDM：一种预测突变对蛋白质稳定性和功能异常影响的服务器。《核酸研究》，39(suppl_2):W215– W222，2011年。[56] J. Z. Xiang 和 B. Honig。Jackal：蛋白质结构建模软件包。哥伦比亚大学和霍华德·休斯医学研究所，纽约，2002年。[57] Y. Yang、A. Niroula、B. Shen 和 M. Vihinen。PON-Sol：预测氨基酸替换对蛋白质溶解度的影响。《生物信息学》，32(13):2032–2034，2016年。[58] Y. Yang、L. Zeng 和 M. Vihinen。PON-Sol2：预测变异对蛋白质溶解度的影响。《国际分子科学杂志》，22(15):8027，2021年。[59] Z. Zhang、J. Norris、V. Kalscheuer、T. Wood、L. Wang、C. Schwartz、E. Alexov 和 H. Van Esch。亚精胺合成酶中的Y328C错义突变会导致轻型斯奈德-罗宾逊综合征。《人类分子遗传学》，22(18):3789–3797，2013年。

[60] Z. Zhang、J. Norris、C. Schwartz 和 E. Alexov。对亚精胺合成酶中致病性错义突变位点可变性的计算机模拟和体外研究。*PLOS ONE*，6(5):e20373，2011年。[61] A. Zomorodian 和 G. Carlsson。计算持久同调。载于第二十届计算几何年度研讨会论文集，第347–356页，2004年。