专题: 生物分子模拟中的机器学习

蛋白质结构模型质量评估方法综述*

刘栋 崔新月 王浩东 张贵军†

(浙江工业大学信息工程学院, 杭州 310014)

(2023 年 6 月 30 日收到; 2023 年 8 月 1 日收到修改稿)

蛋白质模型质量评估方法是蛋白质结构预测的关键技术, 自 CASP7 以来一直是结构生物信息学领域的研究热点. 模型质量评估方法不仅可以指导蛋白质结构模型的精修, 还能够从多个候选构象中筛选出最佳模型, 具有重要的生物学研究和实际应用价值. 本文首先回顾了国际蛋白质结构预测关键评估竞赛 (CASP)、全球持续蛋白质结构预测竞赛 (CAMEO) 以及单体蛋白和复合物的模型评估指标, 主要梳理了近 5 年来包括共识方法 (多模型方法)、准单模型方法和单模型方法在内的模型质量评估方法的发展历程, 并介绍 CASP15 中的复合物模型评估方法; 鉴于深度学习在蛋白质预测领域所取得的巨大进展, 重点分析了深度学习在单模型方法数据集生成、蛋白质特征提取以及网络架构构建方面的深入应用, 并进一步介绍了本课题组近年来在模型质量评估方面开展的工作; 最后, 总结分析了目前蛋白质模型质量评估技术的局限性及所面临的挑战, 并对未来发展趋势进行了展望.

关键词: 蛋白质模型质量评估, 深度学习, 单模型方法, 复合物模型评估

PACS: 87.10.Vg, 87.14.E–, 87.16.A–, 87.55.de **DOI:** 10.7498/aps.72.20231071

1 引 言

蛋白质参与生命活动的各个过程, 是生命体的重要组成部分. 了解蛋白质结构可以进一步揭示生命过程中生物分子复杂的相互作用机制 [1–3]. 经过实验科学家近 60 年来巨大的努力, 已经解析出了二十余万种蛋白质结构. 然而, 由于生物实验过程耗时长且成本较高, 致使实验解析结构仅占已知两亿多蛋白质序列数量的 0.1%[4], 因此, 通过高效且准确的计算方法实现大规模蛋白质结构预测成为 50 多年来计算生物学家努力的方向 [5]. 广泛使用的 Rosetta[6], I-TASSER[7] 是蛋白质领域经典结构预测方法, 随着深度学习技术在该领域研究的广泛应用, 国内外学者陆续提出了 RaptorX[8], trRosetta[9], AlphaFold2[5], PAthreader[10], ESMFold[11] 等方法. 尤其是 DeepMind 和 Meta 研究团队基于 AlphaFold2 和 ESMFold 的方法, 分别构建了约两亿预测结构的数据库 AlphaFold Protein Structure Database[12] 和约七亿预测结构的数据库 ESM Metagenomic Atlas[11]. 针对同一序列, 上述方法预测出的结构存在显著差异. 为解决此类问题, 模型精度估计或者模型质量评估方法 (estimation of model accuracy, EMA)[13] 就成为蛋白质结构预测流程中一个关键的环节. EMA 方法主要目的是估计参考结构与预测模型在整体拓扑 (全局结构) 和残基级别 (局部结构) 相似的程度, 并能够进一步实现模型单残基、连续残基块的拓扑精修, 常用的指标包括 GDT-TS[14], TM-score[15], lDDT[16], CAD[17], SG[18] 等.

Moult 等 [19]1994 年创立的蛋白质结构预测的关键评估 (CASP) 被誉为蛋白质结构预测领域的

* 科技创新 2030—“新一代人工智能”重大项目 (批准号: 2022ZD0115103)、国家自然科学基金 (批准号: 62173304) 和浙江省自然科学基金重点项目 (批准号: LZF030002) 资助的课题.

† 通信作者. E-mail: zgj@zjut.edu.cn

© 2023 中国物理学会 Chinese Physical Society

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

奥林匹克竞赛. CASP 每两年举办一次, 目前开展了 15 届, 已经成为蛋白质结构预测技术发展的风向标[20,21]. 在 2006 年 CASP7 中引入了模型质量评估方法的评测, 这足以说明 EMA 方法对结构预测的重要性. 此外, 另一个重要的国际赛事 CAMEO[22] 自 CASP12 之后引入了每周在线的自动盲测评估服务器, 成为 CASP 两年间评测的重要补充平台. 值得一提的是, AlphaFold2 在 CASP14 中取得巨大的突破, 使得单体结构预测几乎到达了实验解析的精度[23]. 因此, 在 CASP15 中接触预测、优化和单体模型质量评估被取消, 而新增 RNA 结构、蛋白质与配体复合物、复合物结构及其界面的质量评估类别[24], 对于复合物评估, 除了全局结构与局部结构的精度估计之外, 还新增接触界面精度估计, 如 DockQ[25] 和 QS-score[26].

自 CASP7 至目前为止, 已经开发出许多蛋白质模型质量评估方法和在线服务器, 如图 1 所示. 本文梳理了最近 5 年主流的模型质量评估方法, 主要分为共识方法 (多模型方法)、准单模型方法、单模型方法[27]. 共识方法假设正确的结构包含在重复结构模式集合中, 通过聚类提取来自多个方法或不同模板生成的蛋白质结构模型的共识信息, 代表性方法有 Cheng 课题组开发的 MULTICOM 系列[28–30], Xu 和 Shang 课题组开发的 MUfoldQA 系列[31,32] 等. 在 CASP7—15 评测中, 共识方法在大多数情况下都比单模型方法表现得更好. 准单模型方法将单个模型输入的便利性与共识方法预测能力的优势相结合, 通过内部参考结构生成方法产生的一组蛋白质结构对预测模型进行评分, 代表性的方法有 McGuffin 课题组[33–35] 开发的 ModFOLD 系列等. 单模型方法基于单一蛋白质模型特征提取 (序列信息、几何结构、理化信息), 通过神经网络来评估残基或者拓扑的质量. 随着机器学习和深度学习技术在蛋白质结构预测领域广泛、深入地应用, 单模型方法在性能逐渐与多模型方法持平甚至超越, 成为 EMA 方法中一个热点研究方向, 代表性的方法主要有 Baker 课题组[27] 开发的 DeepAccNet 系列、Elofsson 课题组[36,37] 开发的 ProQ 系列, Venclovas 课题组[38–40] 开发的 Voro 系列, 杨建益课题组[41] 开发的 Yang_TBM, 张贵军课题组[42–44] 开发的 DeepUMQA 系列等.

本文将按顺序介绍 CASP 和 CAMEO, 其次详细讨论蛋白质模型质量评估的指标体系, 包括单体蛋白、复合物的评估指标以及综合性能分析指标. 然后, 对近 5 年来主流的共识方法、准单模型方法和单模型方法进行梳理, 并介绍 CASP15 的复合物模型质量评估方法. 考虑到深度学习对蛋白质领域的影响, 本文重点讨论单模型方法中的数据集、蛋白质特征和网络架构这三个方面, 并介绍了本课题组近年来在模型质量评估方面所开展的一些工作. 最后, 分析给出了蛋白质模型质量评估方法所面临的一些关键挑战, 并对未来可能的发展趋势进行了展望.

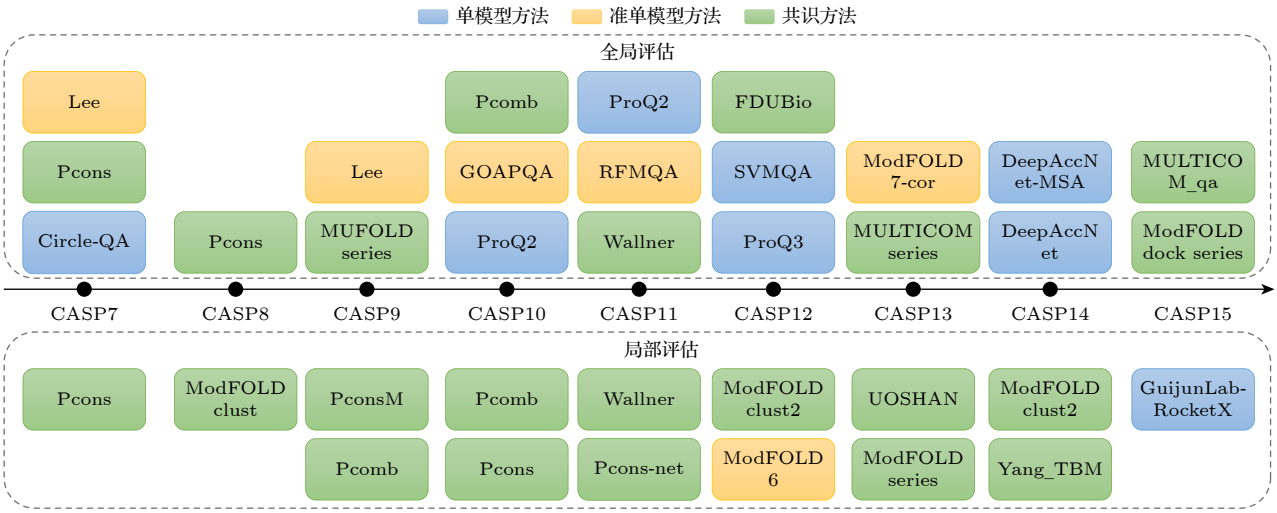

图 1 在 CASP 中主流的模型质量评估方法

Fig. 1. Mainstream model quality assessment methods in CASP.

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2 国际蛋白质结构预测的关键评估竞赛 (CASP) 和全球连续自动模型评估竞赛 (CAMEO)

CASP[19] 自 1994 年以来, 已成功举办了 15 届. CASP 为研究团队提供了一个客观测试蛋白质结构预测方法的平台, 并为研究团队和软件用户提供了对蛋白质结构建模最新技术水平的独立评估. 在 CASP7 中引入了蛋白质模型质量评估的评测, 其中蛋白质模型结构由三维结构预测组提交, 为评估模型质量方法提供了测试数据集. CASP 的评估过程分为两个阶段. 在第 1 阶段, 通过共识方法为每个蛋白质目标选择约 20 个蛋白质结构模型, 覆盖了整个模型质量范围进行评估; 在第 2 阶段, 选择前 150 个模型用于质量评估. 在这两个阶段中, EMA 方法需要评估每个模型的全局拓扑质量和残基级别的局部质量[45,46]. 第 1 阶段的结果仅用于与第 2 阶段的结果比较, 以确定 EMA 方法是否是单模型方法[47]. 在每届 CASP 比赛中, 表现最好的 EMA 方法通常代表了蛋白质质量评估领域的最新发展水平.

此外, 瑞士生物信息研究所和巴塞尔大学联合举办 CAMEO[48] 是一个全球持续进行的蛋白质结构预测平台, 被认为是蛋白质结构预测领域最重要的比赛之一. CAMEO 中每位参赛者每周对由世界范围内的结构生物学家最新破解出的 20 个蛋白质结构进行预测. 在 CAMEO-QE 中, 预测出的结构由模型质量评估参赛者进行评估并在线提交. 多年来, CASP 和 CAMEO 不断进步和相互促进, 为 EMA 研究带来了新的思路和方法, 并推动了这一领域的不断突破和发展.

3 蛋白质模型质量的评估指标

蛋白质结构的准确性和可靠性对于理解生命活动过程至关重要. 为了评估计算方法的性能, 必须使用有效的评估指标来衡量蛋白质模型的质量. 这些评估指标能够判断蛋白质模型与实验解析结构之间的相似程度, 并识别模型中可能存在的结构缺陷或误差, 从而进一步改进和优化模型. 此外, 蛋白质评估指标对于蛋白质设计和药物设计等领域也具有重要意义. 随着多年来蛋白质结构领域的发展, 衍生出了多种评估指标, 特别是在最近 CASP 或 CAMEO 比赛中采用的指标. 总体上来讲, 这些指标大致分为“单体结构质量评估指标”和“复合物结构质量评估指标”, 其中单体结构质量评估指标主要侧重于局部评估指标和全局评估指标, 下面将分别介绍一些常用的评估指标及其应用场景.

3.1 单体结构质量评估指标

对于 CASP 评估者而言, 其中一个主要挑战是定义合适的数值指标, 以量化预测与实验结构之间的准确度. 在 CASP 评估过程中, 研究者通过评估预测模型质量来反映结构预测技术的最新水平[16]. 均方根误差 (root mean square deviation, RMSD) 在 CASP 早期作为主要评估标准[49,50], 然而 RMSD 存在极易受到预测不准确区域的异常值影响、对模型中的缺失部分不敏感、对参考结构的叠加具有较高依赖性的问题[17]. 为了更为客观地评估蛋白质结构模型的质量, 研究者相应提出了多种评估指标来综合描述蛋白质结构的质量.

GDT-score (global distance test score)[14] 从 CASP4 引入以来一直被广泛使用. GDT-score 通过将预测与实验参考结构进行叠合后, 计算模型结构中某种原子 (如 C_α) 落在实验结构对应位置的某个阈值范围内所得到最大的原子数目. 通常 GDT-HA 使用的阈值为 0.5, 1, 2 和 4 Å, GDT-TS 使用的阈值为 1, 2, 4 和 8 Å, 计算公式[14] 如下:

$$\text{GDT-TS}_{(M_\text{p},\,M_\text{r})}=\frac{(P_1+P_2+P_4+P_8)}{4}, \quad (1)$$

其中 M_p 是预测模型; M_r 是参照模型; P_1, P_2, P_4 和 P_8 是 M_p 中的 C_α 原子与 M_r 的 C_α 原子距离小于 1, 2, 4 和 8 Å的概率. 此外, 根据所比较的原子类型, 分为使用侧链的原子 GDC_SC[51] 和全原子 GDC_ALL. 与 RMSD 相比, 局部低精度的原子不会对质量分数产生显著影响. 然而, GDT-score 对于蛋白质的大小具有依赖性. 当蛋白质序列的长度较短时, 它可能接近于随机选择结构模型. 这种显著依赖于序列长度的现象使得评分绝对值大小可能变得毫无意义[15]. 此外, GDT-score 评估中的缺失片段会导致较低的质量得分, 而类似于 GDT-score 这种基于全局叠加比对的度量方法, 其主要局限性在具有多个结构域的柔性蛋白质时更为突出. 全局刚体叠合会由最大的结构域主导, 因此较小的结构域无法正确匹配, 导致不合适的质量分数. 而且结构域相对位置轻微变化 (在生物学上可

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

能是可以忽略的) 可能会强烈影响 GDT-score. 这导致在 CASP 中需要将蛋白质模型分割成评估单元 (AU) 来减少结构域的影响, 并对其进行单独评估.

TM-score[15] 利用蛋白质长度相关的数值来消除之前评估指标中对于蛋白质长度的依赖性. 其次, 与设置特定距离阈值并仅计算低于阈值误差的部分不同, TM-score 会对齐预测模型与参考结构之间所有残基对进行评估, 计算公式[15] 如下:

$$\text{TM-score} = \max\left[\frac{1}{L_{\text{ref}}}\sum_{i}^{L_{\text{aligned}}}\frac{1}{1+\left(\dfrac{d_i}{d_0\left(L_{\text{ref}}\right)}\right)^2}\right], \tag{2a}$$

$$d_0\left(L_{\text{ref}}\right) = 1.24\sqrt[3]{L_{\text{ref}}-1.5}-1.8, \tag{2b}$$

其中 L_{aligned} 和 L_{ref} 分别是对齐的预测和参考结构的序列长度, d_i 是指预测蛋白中的残基与参考蛋白中相应残基之间的距离, $d_0\left(L_{\text{ref}}\right)$ 是用来归一化 d_i 的距离. 由于 TM-score 是基于两个结构之间单个叠加比对计算得出的分数, 当蛋白质长度依赖性对模型评估没有影响时, GDT-score 可以在多个阈值距离下进行评估, 综合考虑了更多的结构信息, 从而提供了更全面的相似性度量[17].

一般来讲, 单体蛋白全局结构模型质量的评估指标是从整体拓扑上比较预测结构与参考结构的相似度, 而局部结构质量评估指标能够细致地分析蛋白质中局部区域的结构特征和稳定性, 帮助研究者们识别和定位潜在的结构问题和缺陷.

为了更好地理解单体蛋白质主链中局部原子的相互作用, 验证其立体化学的合理性. lDDT (local distance difference test)[16] 通过比较参考结构中一定范围内较近的、不属于同一残基的原子对之间的距离进行计算. 如果模型中的距离与参考结构中的距离在一定的阈值范围内 (如 0.5, 1, 2 和 4 Å), 则被认为是符合要求的距离. 通过计算保留距离的比例, 可以得到预测模型的 lDDT. 其能够捕获结合位点中的局部几何结构, 并且对结构域的方位变化不敏感, 使得绝对值分数具有指导性的意义. 并且, 该指标可用于进一步指导结构模型的精细修正和拓扑微调.

由于蛋白质的空间结构是通过残基的相互作用形成, 而这种互作模式可以用空间结构上的接触表示. 因此, 通过量化蛋白质模型结构的接触预测相对于参考结构偏差, 并且不需要两个结构之间的对齐, 从而避免一些叠合对齐的问题. 基于接触面积差异的评估指标接触区域差异 CAD (contact area difference)[17], 它通过计算残基之间的接触面积差异来量化模型与参考结构之间的接触, 计算公式[17] 如下:

$$\text{CAD}_{(i,j)} = \left|T_{(i,j)} - M_{(i,j)}\right|, \tag{3a}$$

$$\text{CAD}^{\text{bounded}}_{(i,j)} = \min\left(\text{CAD}_{(i,j)}, T_{(i,j)}\right), \tag{3b}$$

$$\text{CAD-score} = 1 - \frac{\sum_{(i,j)\in G}\text{CAD}^{\text{bounded}}_{(i,j)}}{\sum_{(i,j)\in G}T_{(i,j)}}, \tag{3c}$$

其中 i 和 j 代表预测模型和参考结构中的残基, G 是参考结构中的接触残基对的集合, $T_{(i,j)}$ 和 $M_{(i,j)}$ 分别表示参考结构和预测模型中的接触面积. CAD-score 可以单独考虑残基主链和侧链, 具有处理模型中缺失残基的能力, 并且类似于 GDT-score, 能够对完整和不完整的模型进行排名. 此外, 另一个指标是 Sphere Grinder (SG)[18], 通过简单直观的方式识别预测模型中不正确的区域.

对于单体蛋白质模型的质量评估, 局部指标和全局指标相互弥补, 有效地揭示蛋白质模型的局部和整体结构质量, 并为蛋白质结构预测提供更可靠的指导.

3.2 复合物结构质量评估指标

随着人工智能技术在单体结构预测领域的突破, 之前的评估指标更适用于描述单体结构的质量, 而研究的重点逐步向复合物转移. 为了探究蛋白质与蛋白质之间的相互作用, 研究者们设计了专门用于复合物 (多聚体) 的评估指标, 这对于预测复合物的结构发展至关重要.

蛋白质相互作用的关键评估竞赛 (CAPRI) 旨在评估蛋白质对接方法和预测蛋白质与蛋白质相互作用关系[52]. CAPRI 引入 F_{nat}, LRMS 和 iRMS 指标用于评估模型[25]. F_{nat} 衡量了预测复合物界面中在实验参考结构中界面接触残基所占的比例, 界面接触被定义为两个相互作用的蛋白质 (受体和配体) 之间任意一对重原子之间的距离在 5 Å以内. LRMS 是在将预测和参考复合物的受体 (两个蛋白质中较大的一个) 进行叠合比对后, 计算配体 (较小的蛋白质) 预测和参考复合物的 RMSD. LRMS 是一个全局指标, 取决于配体的大小. 因此, 在接

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

触界面区域的匹配情况中, 它可能不是一个较好的评估指标. iRMS 仅针对接触界面残基的 RMSD, 其接触界面的残基距离范围重新定义为 10 Å以内, 即 $F_{\rm nat}$ 定义界面阈值的两倍. 虽然这些评估指标可以量化蛋白质对接模型质量的不同方面, 但在对模型排序、模型质量与评分函数的相关性分析以及在机器学习算法中作为目标函数时存在一定限制. 因此, 需要综合考虑多个指标, 以更准确地评估模型的质量. DockQ[25] 将 $F_{\rm nat}$, LRMS 和 iRMS 综合到一个介于 0 到 1 之间的单一评估指标中, 可以更加定量地评估蛋白质对接模型的质量, 计算公式[25]如下所示:

$$\mathrm{RMS}_{\rm scaled}\left(\mathrm{RMS}, d_i\right)=1/\left[1+\left(\mathrm{RMS}/d_i\right)^2\right], \tag{4a}$$

$$\begin{aligned}&\mathrm{DockQ}\\&=\frac{\left(F_{\rm nat}+\mathrm{RMS}_{\rm scaled}(\mathrm{LRMS}, d_1)+\mathrm{RMS}_{\rm scaled}(\mathrm{iRMS}, d_2)\right)}{3};\end{aligned} \tag{4b}$$

其中 $\mathrm{RMS}_{\rm scaled}$ 表示与 LRMS 或 iRMS (RMS) 中的任何一项相对应的缩放后的 RMS 偏差, d_i 是一个缩放因子, d_1 用于 LRMS, d_2 用于 iRMS. $F_{\rm nat}$ 被定义为预测的复合物界面中保留的原生界面接触的比例. 在评估 CAPRI 中的蛋白模型时, DockQ 几乎可以重现原始的 CAPRI 分类, 这意味着不需要使用阈值对预测模型进行分类, 并且可以使用 Z-score 来评估模型质量, 类似于 CASP 中使用的方法.

在蛋白质与蛋白质对接模型评估指标的发展历程中, 主要集中在二聚体的相互作用. 然而, 对于多聚体 (链数大于两条) 需要将其分解为二聚体可能需要大量的比较工作, 并且可能会缺失一些整体结构的接触界面残基. 因此, 研究者设计了 QS-score[26], 用于量化界面之间的相似性, 该相似性取决于共同的界面接触. 其能够区分不同的多聚体结构和结合模式, 计算公式[26]如下所示:

$$\begin{aligned}Q=&\sum_{s(i,j)}\mathrm{w}\left(\min\left(d_i, d_j\right)\right)+\sum_{n-s(i)} w\left(d_i\right)\\&+\sum_{n-s(j)} w\left(d_j\right),\end{aligned} \tag{5a}$$

$$\begin{aligned}&\mathrm{QS\text{-}score}=(1/Q)\times\\&\sum_{\mathrm{shared}(i,j)} w(\min(d_i, d_j))\left(1-\frac{|d_i-d_j|}{12}\right),\end{aligned} \tag{5b}$$

$$w\left(d\right)=\begin{cases}1, & d\leqslant 5,\\ \mathrm{e}^{-2\left(\frac{d-5}{4.28}\right)^2}, & d>5,\end{cases} \tag{5c}$$

其中 d 代表残基之间的欧式空间 C_β 距离, $|d_i-d_j|$ 代表相对误差 (将 12 Å作为最大误差), w 是加权函数. 当涉及的所有残基都被“映射”时, 形成的接触被定义为 s. 而那些接触但未被“映射”的残基对, 或者只在其中一个寡聚体中形成接触被定义为 $n-s$. 这里所提及的“映射”是指一个复合物中的蛋白质链与另一个复合物中蛋白质链之间的对应关系. QS-score 能够评估组装界面的质量, 适用于比较链的相对方位. 在最近的 CASP15 中, 评估者还使用界面接触分数 (ICS) 和接触区域分数 (IPS) 来评估模型. ICS 以 F1-score[53] 的形式计算, 用于衡量预测的链间接触的精准率和召回率之间的关系. IPS 则通过计算模型预测的接触残基与参考结构接触残基之间的部分, 得出 Jaccard[54] 系数.

伴随着结构预测领域的发展, 复合物结构的评估逐渐变得尤为关键. 复合物的评估指标可以从多个独立计算却相关的指标综合成一个评估指标, 并且可以从二聚体拓展到多聚体的评估指标.

3.3 评估结构精度估计的指标

模型质量评估 (EMA) 是 CASP 重要的组成部分, 理想情况下, EMA 方法可以提供与计算的评估指标分数相关的模型质量估计. 在 CASP14 之前的比赛中约有 70 多种参赛方法[55], 这凸显了模型质量评估对蛋白质结构预测的重要性, 并且研究人员通常将模型质量估计整合到建模流程. 蛋白质模型的精度估计包括了每个模型的全局精度评估和每个残基的局部精度估计. 此外, CASP 对参赛组进行分别排名, 这些排名通常使用多个评估指标综合计算得出.

评估全局结构精度估计包含 Top1 loss[47], AUC (area under the curve)[56], 相关性和绝对误差分析. Top1 loss 用于对比蛋白质结构预测模型的精度估计, 并选择排名第一的模型作为最佳模型. 在不同指标下, 计算选定的最佳模型与实际最佳模型质量的绝对误差. 相关性分析使用 Pearson 和 Spearman[57] 来评估预测全局模型与真实模型质量之间的相关性. 通过绝对误差分析 (MAE 或 MSE), 分析不同指标下模型质量预测值与真实值之间的差

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

异. AUC[56] 用于判断预测模型质量是否可以接受, 它通过计算 ROC 曲线下的面积衡量模型的性能, 而 ROC 曲线则反映了在不同质量阈值下, 准确和不准确模型的真阳性率和假阳性率之间的关系.

局部结构精度评估是在评估单元 (EUs)[47] 级别进行. ASE(average S-score error)[47] 是通过计算每个残基的 S-score 误差的平均值来评估:

$$\mathrm{ASE}=\left(1-\frac{1}{N}\sum_{i=1}^{N}\left|S\left(e_i\right)-S\left(d_i\right)\right|\right)\times 100, \quad (6)$$

其中第 i 个残基的 S-score 误差是对预测模型中评估单元 (EU) 的第 i 个 C_α 原子的预测距离误差 (e_i) 和实际距离误差 (d_i) 之间的差值. 通过 LGA[14] 在评估单元的叠合后, 使用 S-function 函数来计算, N 是评估单元中的残基数目. ULR (unreliable local region)[47] 是由预测模型中 3 个或更多连续残基组成的区域, 其在最佳叠合下与相应参考结构的残基之间的距离偏差超过 3.8 Å. 相隔一个残基的两个 ULR 将合并为一个 ULR. 确定 ULR 后, 计算它们的准确度和覆盖率, 并在实际 ULR 边界上以及在两个残基以内的预测被认为是准确预测. 对于每个 CASP 评估组, 通过调整阈值计算以最大化平均 F1-score[53]. 在 CASP 中, 组的排名往往是根据蛋白质目标的评估指标对应平均 Z-score 统计, 其中每个组的 Z-score是对每个目标的结果计算的均值和标准差, 将 Z-score 设置为–2—2.

随着 AlphaFold2 在单体结构预测方面的巨大进展, 几乎解决了单体结构预测问题, 促使 CASP15 将重点转向复合物的预测和模型质量评估. 其中, 整体模型拓扑质量评估采用 GTD-Score 和 TM-Score 指标; 链间相互作用质量评估采用 DockQ 和 QS-Score 进行衡量; 界面接触残基质量评估采用 CAD-Score, lDDT, PatchQS 和 PatchDockQ[24] 指标衡量. CASP 参赛组的性能往往是通过这些指标对应的 Pearson, Spearman, AUC 和 Loss 进行综合加权给出最终排名.

在蛋白质结构预测领域, 质量评估对于建模过程具有重要意义. 质量评估指标提供了一种客观、量化的方法来评估模型的准确性和质量, 同时为改进和优化建模过程提供了指导和依据.

4 蛋白质模型质量方法

在最近的 CASP 中, 研究者已经开发了许多方法, 包括共识、准单模型和单模型的质量评估方法, 主要步骤如图 2 所示. 此外, 鉴于复合物模型评估的重要性, 我们回顾了 CASP15 中的复合物质量评估方法. 最后, 介绍了本课题组近年来在模型质量评估方面开展的工作.

4.1 数据集

训练数据集在神经网络中起着至关重要的作用, 它是神经网络学习和理解模式的基础[58]. 通过训练数据, 神经网络可以从中学习到输入与输出之

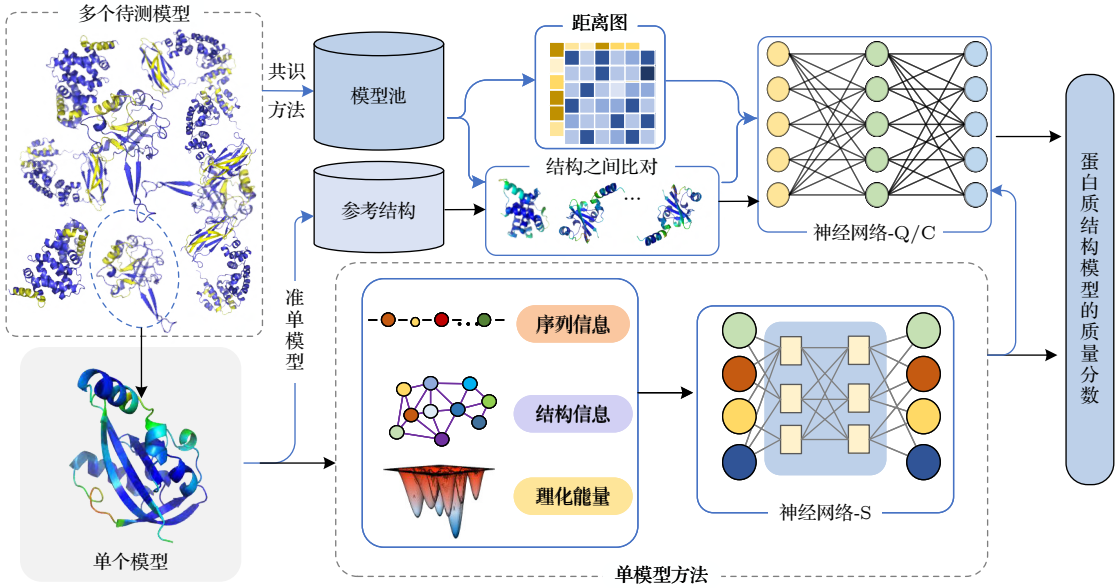

图 2 模型质量评估三类方法示意图

Fig. 2. Schematic diagram of three methods of model quality assessment.

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

间的关联性, 使其能够对新数据进行准确的预测和推断. 丰富、多样且代表性的训练数据可以帮助神经网络克服过拟合和欠拟合等问题, 提高模型的泛化能力和稳定性. 因此, 对基于神经网络的蛋白质模型质量评估而言, 高质量数据集需要包含不同精度的结构并且达到一定程度的数量, 这可以使网络学习到蛋白质的结构与质量的潜在映射关系.

CASP1-CASP15 数据集由每届参加 CASP 结构预测组提交的模型构成. 每个蛋白质目标至少包含 150 个预测结构, 这些结构的精度各不相同, 往往被用于训练和测试模型. 截止至 2023 年 6 月 28 日, CAMEO-QE 数据已经持续评估了 74704 个蛋白质预测模型, 针对每个蛋白质目标的模型数大约为 10 个, 相比于 CASP, 模型的相似度较高且预测难度较低. AlphaFoldDB 和 ESM Metagenomic Atlas 分别是 AlphaFold2 与 ESMfold 预测的高精度蛋白质模型数据库. 虽然大部分结构还未通过实验解析出来, 但是这两个数据集对于蛋白质结构领域的研究具有重要的意义. Zhanglab 服务器中非冗余的蛋白质目标所生成的诱饵结构包含 3DRobot 数据集、I-TASSER 数据集、QUARK 数据集等. 而 DeepAccNet, GNNRefine, DeepUMQA, DeepUMQA3, GraphCPLMQA 和 GraphGPSM 这些方法都采用大致相同的数据集制作思路: 从 PDB 库中筛选出一批非冗余的蛋白质目标, 通过不同的方法生成预测模型结构 (Decoys) 用于训练神经网络. 在开发基于深度学习模型质量评估的方法, 往往可以组合这些数据进行训练, 如表 1 所列.

4.2 共识方法

共识方法在 CASP 蛋白质模型精度评估上具有显著优势. Cheng 课题组[28–30] 开发的 MULTICOM 系列结合了各种质量评估技术, 包括半聚类方法、单模型机器学习方法以及组合方法. 其中, MULTICOM-cluster 和 MULTICOM-construct[29] 在 CASP 质量评估测试中表现优异. MULTICOM 系列评估方法通过结合来自 12 种不同 EMA 方法 (9 种单模型方法和 3 种多模型方法) 以及 1 种蛋白质接触预测方法 (DNCON2[47]) 的预测结果, 生成 10 个质量分数作为预训练深度神经网络的输入特征. 对于 MULTICOM-construct, 这 10 个质量分数取平均值. 而 MULTICOM-cluster 则将 13 个初步预测结果和 10 个 DNNs 预测结果的组合输入另一个 DNN, 进一步预测最终的质量分数. 该研究方法表明, 使用残基与残基接触特征可以显著提高该方法的性能. 在 MULTICOM-AI[16] 中, 基于深度学习技术和共进化分析, 新增了残基间距离特征, 其计算一组结构模型中的残基距离与 DeepDist[30] 预测的距离之间的相关性. 此外, MULTICOM-AI 还使用了基于 DNCON4 生成残基间接触特征.

Xu 和 Shang 课题组开发的 MUfoldQA[31,32] 系列方法, 在 CASP13 中涵盖了 MUfoldQA_M 和 MUfoldQA_T 两种方法, 其核心思想是利用一组参考模型对每个候选模型进行评分. 它们之间的区别在于选择参考模型和计算给定一组参考模型的候选模型评分方式. MUfoldQA 结合了准单模型的质量评估方法, 首先通过在 PDB 数据库中搜索蛋白质序列来获得一组模板. 然后, 从候选模型中选

表 1 模型质量评估的蛋白质结构数据集 (诱饵)

Table 1. Protein structure dataset (Decoys) for model quality assessment.

Data sets	URLs
CASP	https://predictioncenter.org/download_area/
CAMEO	https://www.cameo3d.org/
Zhanglab	https://zhanglab.ccmb.med.umich.edu/decoys/
AlphaFoldDB	https://alphafold.ebi.ac.uk/
ESM Metagenomic Atlas	https://esmatlas.com/resources?action=search_structure
DeepAccNet	https://github.com/hiranumn/DeepAccNet
GNNRefine	http://raptorx.uchicago.edu/download/
DeepUMQA	https://academic.oup.com/bioinformatics/article/38/7/1895/6520805?login=true
DeepUMQA3	https://www.biorxiv.org/content/10.1101/2023.04.24.538194v1.full.pdf+html
GraphCPLMQA	https://www.biorxiv.org/content/10.1101/2023.05.16.540981v1.full.pdf+html
GraphGPSM	https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbad219/7197734?searchresult=1#supplementary-data

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

择一个子集作为参考模型, 并根据与模板的相似性对每个参考模型进行评分. 最后, 每个候选模型根据其与参考模型的相似性进行评分, 并考虑到参考模型的评分进行加权. 此外, MUfoldQA_G[59] 结合了蛋白质模板和参考模型的信息, 以优化最大化皮尔逊相关系数的 QA 指标. MUfoldQA_Gr 通过重采样训练数据并训练模型, 学习到更好的共识模式, 同时最小化了平均 GDT-TS 误差. MUfoldQA_G 将 MUfoldQA_Gr 和 MUfoldQA_Gp 的结果相结合, 使最终的预测结果接近 MUfoldQA_Gr 的低平均 GDT-TS 误差, 并保持与 MUfoldQA_Gp 结果相同皮尔逊相关系数.

McGuffin 开发的 ModFOLDclust2[60] 是一种基于自动聚类的领先方法, 用于对局部和全局模型的质量评估. ModFOLDclust2 服务器在 CASP9-CASP14 中测试的方法基本相同. ModFOLDclust2 最初的开发目标是减少计算代价, 并提供比 ModFOLDclust[61] 更高的预测精度. ModFOLDclust2 的全局质量分数为 ModFOLDclustQ 和 ModFOLDclust 全局质量评估分数的平均值. 为了进行全面的比较模型, 使用了一种修改后的无结构比对的 Q-measure[62]. ModFOLDclust2 的残基的质量评估分数是直接从 ModFOLDclust 中获取.

杨建益课题组[41] 开发 QDistance(Yang_TBM) 是基于 trRosetta 预测的残基间距离估计全局和局部质量. QDistance 使用 trRosetta 预测查询蛋白的残基间距离和结构模型. 为了预测每个模型的全局质量评估分数, 设计了三组特征, 包括基于 2D 距离矩阵比对、势能分数和其他单一 QA 方法以及 1D 结构特征比较的特征. 这些特征被输入到线性回归模型中, 以预测 GDT_TS. 为了进行局部 QA 预测, 首先选择排名靠前的模型 (根据预测的 GDT_TS 分数), 然后使用共识分析来推断每个模型的局部质量分数.

clustQ 是 Bhattacharya 课题组[63] 基于加权距离比较的无超聚 (superposition-free) 方法评估质量. clustQ 对在序列中相隔较远的残基, 分配了较高的权重. 这类残基之间相互作用相对于局部短程相互作用提供了更多的信息, 并且使用基于 Q-score[62] 扩展的 WQ-score 对模型之间进行了配对比较, 以估计预测模型质量精度.

此外, UOSHAN[64] 是基于聚类 SARTclust_G 和 SARTclust_L 的评估方法. 在全局和局部评分中, 根据 SART_G 分数对预测模型进行排名, 形成一个包含前 N 个模型的参考集合. 然后, 将待评估模型与参考集合中的所有模型进行 TM-score 比对. 对于全局评分, 计算 N 个比较得到的 GDT_TS 分数, 并使用 SARTclust_G 对这些分数进行加权平均. 对于局部评分, 计算相应残基之间的 N 个距离值, 然后使用 SARTclust_G 对这些 S-score 进行加权平均. MESHI_consensus[65] 是基于 LightGBM[66] 随机森林回归器, 利用结构、序列和共识特征来估计蛋白质模型的质量.

4.3 准单模型方法

共识方法在 CASP 测试中表现出色, 因为它们能够利用多个模型之间的信息来生成更准确的预测. 然而, 共识方法的性能很大程度上受候选模型池质量和全面性的影响. 如果候选模型池质量较低或缺乏全面性, 那么共识方法的性能可能会受到影响. 鉴于共识方法的局限性, 准单模型方法通过参考其内部方法生成的一组蛋白质结构来评估预测模型, 从而避免了依赖于候选模型池的问题.

McGuffin[35] 开发 ModFOLD 系列方法作为准单模型方法在 CASP 测试中表现出色, 其中 ModFOLD6[67], ModFOLD7[68] 和 ModFOLD8[33] 在 CASP 评测中表现突出. 它们具有类似的工作流程, 通过使用不同的单模型和准单模型方法对蛋白质模型进行独立评估, 并生成局部质量评分. 这些局部质量评分被视为特征, 并输入到神经网络中, 以推导出最终的预测的全局评分. ModFOLD6 采用了多个评估方法, 如 ProQ2[36]、接触距离一致性 (CDA)、二级结构一致性 (SSA)、无序 B-factor 一致性 (DBA)、ModFOLD5(MF5s) 和 ModFOLDclustQ (MFcQs). 在 ModFOLD6[69] 中, 为了提高局部质量预测的准确性和单模型排名的一致性, 它采用了与之前类似的十种单模型和准单模型方法. ModFOLD7 还提供了两个版本, 分别是在排序 Top 1 模型方面表现最好的 ModFOLD7-rank 和在反映估计绝对误差方面表现良好的 ModFOLD7-cor. ModFOLD8[35] 结合了来自 13 种评估方法 (包括 9 个单模型和 4 个准单模型) 进一步发挥多个单模型和准单模型方法的各自优势提高预测准确性.

此外, QMEANDisco[70] 利用与同源模型结构的距离分布, 使用训练神经网络将多模板 DisCo 分数和单模型 QMEAN[71] 分数加权组合, 得到 QMEANDisCo 复合分数.

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.4 单模型方法

随着机器学习和深度学习的发展, 在蛋白质领域单模型评估方法得到越来越多关注与研究. 这些方法只需要一个模型作为输入, 并能够表现出与共识方法相似或更好的性能. 单模型方法可以分为基于传统机器学习和基于深度学习的评估方法, 并鉴于深度学习对蛋白质领域的影响, 将对基于深度学习模型评估方法从特征、网络以及架构展开描述.

基于传统机器学习的单模型质量评估方法通常使用多种特征作为输入, 包括基于能量的特征、基本的物理化学特征和统计特征. 例如 SVMQA[72] 方法则将基于势能的特征和基于一致性的特征作为输入, 使用随机森林算法预测全局质量. 此外, 还通过改变特征组合改善质量得分. MESHI-enrich-server, MESHI-corr-server 和 MESHI-server 使用机器学习训练的 3 种不同损失函数分析对该方法性能的影响.

对基于深度学习的单模型质量评估而言, 蛋白质模型特征和网络架构对于方法的性能有关键影响. 特征可以显性刻画蛋白质的属性, 其中包括蛋白质的结构特征和非结构特征. 对于结构的特征, 3DCNN[73] 仅利用 3D 结构的原始原子密度作为特征, 没有进行任何特征调整. Ornate[74] 表示基于体素化特征的蛋白质拓扑结构, 这些体素化特征根据骨架中原子的方向构建立方图, 描绘了残基及其邻域. Atom-ProteinQA 设计了两个提取几何和拓扑原子级关系模块. 几何感知模块捕捉输入蛋白质的几何特征, 生成细粒度的原子级预测, 基于化学键构建原子级图通过拓扑感知模块的消息传递并行输出残基级别的预测. 这些方法通过低维空间关系来表示蛋白质几何模型结构.

对于非结构特征, ProQ3D[75] 采用了基于 Rosetta 能量项的两个特征, 即全原子 Rosetta 能量项和粗粒化中心点 Rosetta 能量项. Venclovas 课题组 [38] 开发的 VoroMQA, 将统计势的概念与原子球的 Voronoi[76] 分割相结合评估模型质量. 其将蛋白质结构表示为一组原子球, 每个球具有对应于原子类型的范德瓦耳斯半径分配的空间区域, 并使用 Voronoi 面和球面的三角表示, 接触面积被计算为对应三角的面积. 其中, VoroMQA-A 通过使用 SCWRL4[77] 重构其侧链对输入模型进行预处理, 而 VoroMQA-B 在评估之前不会修改输入模型. 此外, 特别是, 序列信息中在包含潜在的蛋白质进化关系, 可以提高模型评估的准确性. ProQ4[78] 使用多序列比对的统计信息熵提升原有评估的精度. Bhattacharya-QDeepU(QDeep[79] 的变体方法) 使用从全基因组序列数据库与宏基因组数据库合并生成的多序列比对信息 (MSA) 进行训练. VoroCNN-GEMME 使用 GEMME[80] 计算了每个残基的共进化描述符, 其预测了在该序列位置发生突变对其他每个氨基酸的影响程度, GEMME 的输入也是 MSA 信息. DeepAccNet-MSA[27] 通过 trRosetta[9] 网络将 MSA 信息转换为几何约束特征输入神经网络预测质量分数.

深度学习网络可以捕获蛋白质内部的潜在联系. Venclovas 课题组 [81] 开发 VoroMQA-dark 是基于部分 VoroMQA, 通过神经网络 (NN) 来预测局部 (每残基)CAD-score 值. 其针对每个氨基酸残基输出包括 3 个 CAD-score: CAD-score-level0 是基于涉及中心残基的所有氨基酸残基间接触; CAD-score-level1 是基于涉及至少一个来自中心残基的第一层邻居 (直接邻居) 的所有氨基酸残基间接触; CAD-score-level2 是基于中心残基的直接邻居和直接邻居的邻居与所有氨基酸残基之间的间接接触来计算的. 输入向量已经进行了预卷积操作, 最终只使用了一个全连接隐藏层. VoroCNN[40] 是一种基于深度卷积神经网络的模型质量评估方法, 它处理无向加权图表示的蛋白质模型. 为了处理这些图, VoroCNN 由一个基于消息传递图卷积层和一个池化层组成. 此外, VoroCNN-GDT 网络输出层之前增加了一个 1D 卷积层, 以实现在蛋白质序列上有更好的局部质量预测的平滑性. Bhattacharya 课题组 [79] 提出的 QDeep (Bhattacharya-QDeep) 采用堆叠式深度 ResNet 估计模型在四个不同距离阈值 1, 2, 4 和 8 Å下每残基的误差. 其中, 4 个 ResNet 网络独立训练. DeepQA[82] 使用多个特征 (包括能量、物理化学性质和结构信息) 输入到深度置信网络中预测质量, 该网络由受限玻尔兹曼机 (RBM)[83] 隐藏层和逻辑回归层构成的网络结构. AngularQA[84] 将原子结构信息转化为二面角和键长, 并将序列信息通过 LSTM[85] 神经网络输入. 它使用每个残基作为时间步, 预测模型的质量, 并考虑 LSTM 单元的返回值. GraphQA[86] 使用图卷积网络并使用与 ProQ4 相同的特征, 将蛋白质分子转化为具有旋转不变性的图形来评估质量. tFold [87] 通过更改消息传递网络 (MPNN)[88]

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的图形通用架构, 学习了残基之间的相互作用对模型进行评分.

通过构建编解码可以更好地利用神经网络的模块, 以实现更准确的预测. Baker 课题组[27] 开发的 DeepAccNet 是基于一维、二维和三维特征的模型, 在不同层次上反映蛋白质模型. 它通过对三维原子网格在旋转不变的局部框架中对每个残基周围执行三维卷积操作来捕捉高分辨率原子空间结构. 二维特征提取了模型结构中所有残基对的信息, 包括 Rosetta 残基间的相互作用项, 进一步描述原子间相互作用的细节, 而残基与残基的距离和角度特征提供了较低分辨率的结构信息. 在每个残基水平上的一维特征包括氨基酸序列、主链扭转角和 Rosetta 残基能量项. 该网络使用三维卷积评估局部原子环境, 然后通过二维卷积提供全局环境来预测蛋白质的局部质量, 并预测每个残基的质量精度和蛋白质模型中残基间的距离误差, 并利用这些预测来指导蛋白质结构的精修和优化. 此外, AlphaFold2 通过 Evoformer 编码序列信息, 并在 Structure 模块解码中预测原子坐标和结构的质量.

4.5 复合物结构模型评估方法

在 CASP15 中, 模型质量评估从单体质量评估转移到复合物的质量评估. MULTICOM_qa 是结合了基于深度学习链间接触预测和界面接触概率评分的方法, 使用一个蛋白质目标的多聚体模型池作为输入, 预测它们的全局质量得分. 并使用 MMalign[89] 将多聚体模型相互比对, 并计算模型与池中其他模型之间的平均 TM-score 作为模型质量的度量. 此外, 对于每个多聚体目标蛋白质, 使用基于深度学习方法[18] 预测的多聚体残基间接触或距离, 计算链间残基接触的概率, 并将其平均值作为模型全局质量的另一个度量. 最后, 通过加权计算得到池中每个多聚物模型的最终预测质量得分. MULTICOM_egnn 基于 DProQA[90] 将多聚体模型作为输入并将其表示为三维图, 使用门控图 Transformer 架构预测 DockQ 质量分数. 此外, MULTICOM_deep 采用类似的方式.

McGuffin 课题组[91] 开发了 ModFOLDdock 的三种变体: ModFOLDdock, ModFOLDdockR 和 ModFOLDdockS. 这些变体结合了一系列单模型、聚类和深度学习方法形成共识来计算评估复合物质量. ModFOLDdock 优化了预测分数与参考分数的相关性, ModFOLDdockR 优化了挑选 Top 1 模型的能力, 而 ModFOLDdockS 使用 MultiFOLD 方法从输入序列生成参考模型集, 并使用多个评分方法将每个模型与参考集进行比较.

MUFold 和 MUFold2[32] 结合 AlphaFold-Multimer[92] 作为蛋白质复合物质量评估的方法. MUFold 采用了基于 AlphaFold-Multimer 预测结果的单阶段机器学习方法, 而 MUFold2 则采用了两阶段机器学习方法. 在 MUFold2 中, 首先使用 AlphaFold-Multimer 的输出结果训练一个模型进行初始预测, 然后使用第二个预训练的模型生成更准确的预测结果.

VoroIF-jury[93] 包含了两种界面评分方法: 一种是通用的基于原子间接触面积的能量势函数, 该势函数是从蛋白质界面的 VoroMQA 势能函数推导出来的; 另一种 VoroIF-GNN[93] 方法是基于接受由 Voronoi 镶嵌派生的蛋白质链间界面接触图的图注意力网络 (GAT) 预测复合物模型中的残基级别界面精度. 此外, APOLLO[94] 使用基于能量模型 (EBM) 来评估整体折叠、界面准确性以及界面残基的置信度得分.

4.6 DeepUMQA 系列

张贵军课题组在最近几年开发了 DeepUMQA 系列、GraphGPSM 等模型质量局部及全局评估方法. 基于 DeepUMQA[42–44] 系列算法开发的 GuijunLab-RocketX 服务器与基于 GraphGPSM[95] 算法开发的 GuijunLab-Threader 服务器首次参加了 2022 年举行 CASP15, 并表现出了不错的性能.

DeepUMQA[42] 基于超快速形状识别 (USR)[96] 来补充对于描述残基级别的拓扑信息可能不足的情况, 其能够与深度学习方法相结合进一步反映残基级别拓扑的特征来提高模型质量评估的性能. 体素化方法有效地描述了残基的局部结构信息, 但它并未完全反映残基与整体结构之间的拓扑关系. 此外, 体素化特征向量的计算和三维卷积非常复杂且耗时. 因此, 通过选择适当的一组原子间距离, 可以几乎不增加额外的计算成本快速捕捉蛋白质结构的拓扑信息. 具体而言, 考虑了四个参考位置有效代表蛋白质结构中心和边界关系, 并利用它们之间的距离子集构建蛋白质整体结构的拓扑关系.

DeepUMQA2[44] 是基于 DeepUMQA 的显著改进版本. 在基于之前特征基础上, 结合了来自多

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

序列比对的序列信息和同源模板的结构特征, 对模型的潜在属性进行表征. DeepUMQA2 首先根据输入模型的序列进行多序列比对 (MSA) 和同源模板搜索, 然后提取序列特征和模板结构特征, 并与输入模型相关特征结合, 形成初始残基对信息. 通过基于三角乘法更新和轴向注意机制的网络迭代更新残基对信息. 然后, 使用两个分支网络分别预测残基间距离偏差和接触图 (阈值为 15 Å), 进一步计算模型的每个残基的准确性.

DeepUMQA3[97] 适用于评估蛋白质复合物模型质量的方法. 在 DeepUMQA 和 DeepUMQA2 的基础上, 为复合物结构设计了新的特征, 并使用改进的深度神经网络预测了每个残基的 lDDT 和界面残基的准确性. DeepUMQA3 在 CASP15 的蛋白质复合物界面残基准确性估计中名列第一, 参见图 3. 其 Web 服务器为蛋白质复合物提供了快速准确的界面残基准确性预测和每个残基的 lDDT 预测服务. 对于待评估的复合物结构, DeepUMQA3 从三个层次描述它: 整体复合物特征、单体内特征和单体间特征. 在整体复合物层次上, 将整个复合物视为一个大的单体结构. 考虑到蛋白质复合物在序列上是不连续的, 提取了与残基顺序无关的特征, 包括整体 USR、残基体素化、残基间距离和方向以及氨基酸性质. 在单体内层次上, 分别提取了每个单体的特征, 包括由 ESM-1b[98] 生成的序列嵌入、二级结构和 Rosetta 能量项. 在单体间层次上, 使用单体间成对序列的注意力图描述了单体之间的序列关系. 此外, 设计了单体间 USR 来描述一个单体中残基与其他单体的拓扑关系. 这三个层次的特征被输入带有三角形更新和轴向注意力的深度卷积神经网络, 以预测残基间距离偏差和阈值为 15 Å的残基间接触图, 从而计算每个残基的 lDDT 和界面残基准确性.

在 DeepUMQA 系列算法基础上, 张贵军课题组 [99] 进一步结合图耦合网络开发了 GraphCPLMQA 算法. 算法利用蛋白质语言模型的嵌入来评估残基级别的蛋白质模型质量. GraphCPLMQA 由图编码模块和基于变换的卷积解码模块组成. 在编码模块中, 利用具有 ESM 蛋白质语言模型提取序列和高维几何结构的潜在关系表示, 能够捕捉蛋白质模型的序列和结构特征的重要信息. 在解码模块中, 利用提取的嵌入表示和低维特征推断蛋白质结构与质量之间的映射关系. 为了增强局部结构和整体拓扑之间的关联性, 设计了三角定位和残基级别接触顺序特征. 其中, 三角定位基于 DeepUMQA 中的 USR 引入了残基之间方向的信息, 可以更为充分地描述蛋白质局部空间的结构. 接触序 (contact order)[100] 用于描述整体拓扑的复杂性, 并扩

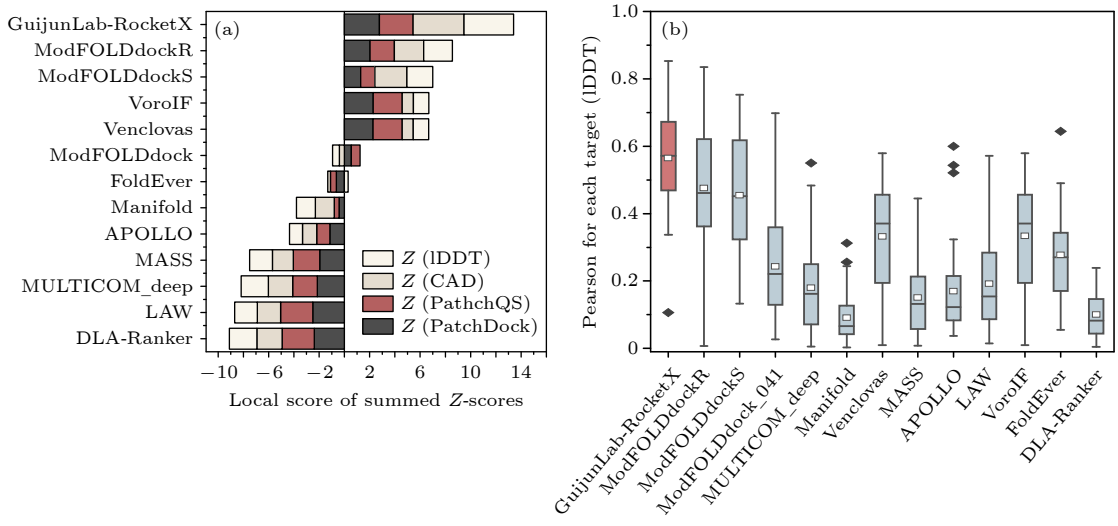

图 3 (a) lDDT, CAD, PatchDockQ 和 PatchQS 的平均 Z 分数之和, CASP15 官方公布各个小组在界面残基精确度估计排名 (数据来自 https://predictioncenter.org/casp15). CASP15 中 DeepUMQA3 的组名称为“GuijunLab-RocketX”; (b) 针对 CASP15, 每个蛋白质目标上的预测的 lDDT 质量与真实 lDDT 质量的 Pearson 相关性, 其中, 白色方框是均值, 中间横线是中位数

Fig. 3. (a) The sum of average Z-scores of lDDT, CAD, PatchDockQ and PatchQS, CASP15 officially announces the ranking of each group in the interface residue accuracy estimation (data from https://predictioncenter.org/casp15). The group name of DeepUMQA3 in CASP15 is “GuijunLab-RocketX”. (b) Pearson correlation of predicted and true lDDT quality on each protein target. The white box is the mean and the middle horizontal line is the median.

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 2 CAMEO-QE: 模型质量评估性能 (数据来自官网 2022-6-24—2023-6-17)

Table 2. CAMEO-QE: Model Quality Evaluation Performance (Data from official website 2022-6-24–2023-6-17).

Predictor Name	$ROC^{normalized}$		$PR^{normalized}$		Models
	$AUC_{0,1}$	$AUC^{*}_{0,0.2}$	$AUC_{0,1}$	$AUC^{*}_{0.8,1}$	Received
ZJUT-GraphCPLMQA	0.82	0.73	0.79	0.54	5143
DeepUMQA2	0.72	0.62	0.68	0.47	4468
DeepUMQA	0.73	0.60	0.67	0.45	4611
ModFOLD9	0.63	0.52	0.59	0.36	4309
QMEANDisCo3	0.9	0.66	0.79	0.49	6348
ProQ3D_LDDT	0.74	0.55	0.67	0.43	5171
QMEAN3	0.88	0.65	0.77	0.43	6348
ProQ3	0.72	0.53	0.66	0.39	5126
VoroMQA_v2	0.89	0.64	0.77	0.45	6350
ProQ2	0.86	0.59	0.74	0.39	6337
ProQ3D	0.70	0.47	0.61	0.35	5119
ModFOLD7_lDDT	0.84	0.53	0.69	0.41	6191
ModFOLD8	0.79	0.50	0.65	0.38	5802
Baseline Potential	0.80	0.51	0.66	0.32	6350
VoroMQA_sw5	0.82	0.50	0.65	0.36	6349
ModFOLD6	0.73	0.42	0.57	0.35	5380

展到残基级别特征以描述局部结构之间的复杂性. 这些特征有助于捕捉蛋白质模型的局部结构元素与全局折叠模式之间的关系. 通过结合图编码模块和基于变换的卷积解码模块, 能够评估蛋白质模型的残基级别的质量. GraphCPLMQA 持续参加了一年的 CAEMO (https://www.cameo3d.org), 结果如下表 2 所列.

此外, 本课题组[95] 还开发了全局质量评估模型 GraphGPSM, 该模型利用高斯径向基函数对原子级别的主链特征进行编码, 基于 DeepUMQA 的 USR, Rosetta 能量项、距离和方向、序列的独热编码以及残基的位置嵌入来描述蛋白质结构. 这些特征被配置到初始图的节点和边上, 并与坐标嵌入相结合, 构建了 EGNN[101] 的初始架构. 通过堆叠 EGNN 架构形成了一个密集的消息传递网络. 最后, 通过多层感知器 (由 Dropout 层、激活函数和线性层组成) 生成结构模型的全局评分. 特别地, GraphGPSM(GuijunLab-Threader) 在CASP15 性能如表 3 所列.

深度学习在蛋白质模型质量评估领域得到广泛应用, 并成为主流技术, 评估质量的效果也显著提升. 回顾模型质量评估方法, 可以得出以下几点结论:

1) 近三年来开发出的单模型方法大多都是基于深度学习. 尤其, 与之前 CASP 中最佳的单模型方法以及 CASP 中最佳的多模型方法相比, CASP14 上最佳单模型方法 (DeepAccNet 和 DeepAccNet-MSA) 在全局结构准确性评估方面取得显著的提升. 虽然, 在 CASP15 全局质量评估和接口界面评估中最好的两种方法分别是 MULTICOM_qa 和 ModFOLDdock 这两种共识方法. 但是, 在局部接触界面的质量评估方法基于深度学习的 DeepUMQA3 相比于排名第二的共识方法具有显著的优势, 单模型方法依然是未来的发展趋势.

表 3 在所有蛋白质目标与 CASP15 服务器的性能比较 (数据来自 GraphGPSM)

Table 3. Performance comparison with CASP15 server on all protein targets (data from GraphGPSM).

Method	Average TM-score	Average Pearson	Average bias
GraphGPSM	0.730	0.633	0.126
MULTICOM_qa	0.485	0.715	0.258
ModFOLDdock	0.515	0.636	0.241
ModFOLDdockR	0.666	0.635	0.165
Venclovas	0.449	0.494	0.339
Manifold	0.582	0.541	0.179
Bhattacharya	0.387	0.474	0.361
*Real value	0.716	None	None

注: *Real value 代表CASP15中所有蛋白质目标所有模型的真实平均TM-score分数.

Note: *Real value represents the real average T-score of all targets in CASP15.

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2) 从 CASP13—CASP15 模型质量评估的参赛组可以看出: 在 CASP13 中分别有 51 个和 29 个参赛组提交了全局和局部精度估计; 在 CASP14 中分别有 72 个和 38 个参赛组提交了对全局和局部精度估计; 在 CASP15 中分别有 22 个, 13 个和 17 个参赛组提交了全局, 局部和接触界面精度估计. 从 CASP13 至 CASP14 对于评估质量的参赛组的数量呈现上升的趋势, 但是从 CASP14 至 CASP15 的参赛数量非常明显的减少. 这可能的原因是: ①对于复合物的模型质量评估, 很多之前的参赛组并没有开发出相应的方法. ②现阶段复合物的结构模型质量评估依旧存在挑战.

3) 通过深度学习的发展历程可以看出, 在网络层面, 从 ProQ3D 简单的几层神经网络逐步引入了更加复杂的模型, 即 3DCNN 的 3 维卷积网络、AngularQA 的 LSTM 网络、GraphQA 的图神经网络、GraphGPSM 的等变图网络, DeepUMQA2 的注意力机制网络以及编解码模块 AlphaFold2 或者 GraphCPLMQA. 在特征层面, 距离图的特征和序列编码向表征局部空间结构, 全局拓扑结构和进化信息设计特征描述蛋白质模型, 如 USR, 体素化, MSA 多序列比对信息等. 这表明深度网络的架构和蛋白质特征对网络模型性能的提升产生关键作用.

5 模型质量评估方法的挑战与发展趋势

模型质量评估方法在蛋白质结构预测中扮演着关键角色, 并持续成为该领域的研究热点. 然而, 这一领域依然面临许多挑战, 以下从单体模型评估、复合物模型评估和模型评估的共性问题三个方面进行讨论.

在单体模型评估方面, 尽管 AlphaFold2 已经取得了卓越的精度, 但对于缺乏多序列比对 (MSA) 数据或模板质量较低的情况, 建模精度仍存在局限性. 目前关键问题在于如何区分高质量模型 (如 AlphaFold2 生成的模型) 和低质量模型, 并评估高质量模型中需要改进的相对不正确区域. 此外, 目前蛋白质预测的结构数据库规模庞大, 如 AlphaFold Protein Structure Database (~2 亿) 和 ESM Metagenomic Atlas (~7 亿). 虽然这些预测结构有自评估的质量分数, 但是这些分数与预测的结构相关性依然需要提升, 特别是在局部区域. 如何通过模型质量评估合理利用这些预测数据促进生物学研究值得深思.

在复合物评估方面, 研究者们面临着许多需要进一步探索的问题, 这些问题源于复合物结构的复杂性和多样性. 首先, 复合物的质量评估需要解决基于深度学习的方法如何构建适当的训练数据集的问题. 由于复合物模型可能包含多个链, 而蛋白质结构数据库中主要以双链结构为主, 如何有效地收集和组织复合物结构数据, 以便用于训练深度学习模型. 其次, 复合物的结构通常比单体结构更加复杂和庞大, 其复杂性意味着在网络训练过程中需要更大的计算和内存资源, 并且训练时间可能会显著增加. 最后, 复合物评估指标体系的建立和应用也需要进一步发展. 目前, 许多复合物的评估指标仍在沿用单体结构的评估方法, 然而复合物具有独特的结构和功能特征, 需要开发适用于复合物质量评估的专用指标, 以更好地反映复合物的质量和功能特性, 并促进复合物结构预测领域的进一步发展.

除了在单体和复合物评估中面临的挑战之外, 模型评估中还存在一些共性问题需要解决. 首先, 对于模型的质量评估, 传统上常常依赖于多序列比对 (MSA) 和模板的信息来提高评估的准确性. 然而, 在某些情况下, 蛋白质的序列可能缺乏足够的相关信息或者没有相关的模板结构可供参考. 因此, 如何仅仅利用蛋白质的单序列和结构本身的信息来评估模型的质量成为一个重要的问题. 其次, 在模型评估中, 有时会发现模型的结构在局部区域被认为是较低质量的, 然而却缺乏对这些局部结构进一步处理的方法. 如何在模型评估的基础上进行结构的精修成为一个需要关注的问题.

综上所述, 未来模型质量评估的趋势将聚焦于复合物模型结构的评估. 借助深度学习网络和最新技术的融合, 以及对复合物模型的结构和序列特征进行工程化的探索, 以揭示不同类型复合物的互作方式. 同时, 引入更加全面和合理的评估指标体系, 将进一步推动复合物结构预测的发展, 并为模型评估提供更加可靠和准确的基础. 这一努力的成果将为蛋白质领域带来更为深入的认知和应用前景, 为研究者揭示复合物结构的复杂性和功能特征提供更精准的工具和方法.

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参考文献

[1] Thompson M C, Yeates T O, Rodriguez J A 2020 *F1000 Research* **9** 667

[2] Bai X C, McMullan G, Scheres S H 2015 *Trends Biochem. Sci.* **40** 49

[3] Wüthrich K 2001 *Nat. Struct. Biol.* **8** 923

[4] Steinegger M, Mirdita M, Söding J 2019 *Nat. Methods* **16** 603

[5] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl S A, Ballard A J, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman1 D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior A W, Kavukcuoglu K, Kohli P, Hassabis D 2021 *Nature* **596** 583

[6] Rohl C A, Strauss C E, Misura K M, Baker D 2004 *Methods in Enzymology* (Amsterdam: Elsevier) pp66–93

[7] Zhang Y 2008 *BMC Bioinf.* **9** 40

[8] Källberg M, Wang H P, Wang S, Peng J, Wang Z Y, Lu H, Xu J B 2012 *Nat. Protoc* **7** 1511

[9] Yang J Y, Anishchenko I, Park H, Peng Z L, Ovchinnikov S, Baker D 2020 *PNAS* **117** 1496

[10] Zhao K L, Xia Y H, Zhang F J, Zhou X G, Li S Z, Zhang G J 2023 *Commun. Biol.* **6** 243

[11] Lin Z M, Akin H, Rao R, Hie B, Zhu Z K, Lu W T, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Costa S D A, Zarandi F M, Sercu T, Candido S, Rives S 2023 *Science* **379** 1123

[12] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A 2022 *Nucleic Acids Res.* **50** D439

[13] Chen J R, Siu S W 2020 *Biomolecules* **10** 626

[14] Zemla A J 2003 *Nucleic Acids Res.* **31** 3370

[15] Zhang Y, Skolnick J 2004 *Proteins Struct. Funct. Bioinf.* **57** 702

[16] Mariani V, Biasini M, Barbato A, Schwede T J 2013 *Bioinformatics* **29** 2722

[17] Olechnovič K, Kulberkytė E, Venclovas Č 2013 *Proteins Struct. Funct. Bioinf.* **81** 149

[18] Antczak P L M, Ratajczak T, Lukasiak P, Blazewicz J 2015 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* Washington D. C, November 9–12, 2015 p665

[19] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A 2016 *Proteins Struct. Funct. Bioinf.* **84** 4

[20] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J 2019 *Proteins Struct. Funct. Bioinf.* **87** 1011

[21] Moult J, Pedersen J T, Judson R, Fidelis K 1995 *Proteins Struct. Funct. Bioinf.* **23** R2

[22] Robin X, Haas J, Gumienny R, Smolinski A, Tauriello G, Schwede T 2021 *Proteins Struct. Funct. Bioinf.* **89** 1977

[23] Fowler N J, Williamson M P 2022 *Structure* **30** 925

[24] Kryshtafovych A, Antczak M, Szachniuk M, Zok T, Kretsch R C, Rangan R, Pham P, Das R, Robin X, Studer G, Durairaj J, Eberhardt J, Sweeney A, Topf M, Schwede T, Fidelis K, Moult J 2023 *Proteins Struct. Funct. Bioinf.* **91** 1550

[25] Basu S, Wallner B 2016 *PLoS One* **11** e0161879

[26] Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T 2017 *Sci. Rep.* **7** 10480

[27] Hiranuma N, Park H, Baek M, Anishchenko I, Dauparas J Baker D 2021 *Nat. Commun.* **12** 1340

[28] Wang Z, Eickholt J, Cheng J L 2010 *Bioinformatics* **26** 882

[29] Cheng J L, Wang Z, Tegge A N, Eickholt J 2009 *Proteins Struct. Funct. Bioinf.* **77** 181

[30] Wu T Q, Guo Z Y, Hou J, Cheng J L 2021 *BMC Bioinf.* **22** 1

[31] Wang J L, Wang W B, Shang Y, Xu D 2022 *IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI)* Las Vegas, NV, USA & Changsha, China, December 6–8, 2022 p84

[32] Wang W B, Li Z Y, Wang J L, Xu D, Shang Y 2019 *Nucleic Acids Res.* **47** W443

[33] McGuffin L J, Aldowsari F M, Alharbi S M, Adiyaman R 2021 *Nucleic Acids Res.* **49** W425

[34] McGuffin L J, Buenavista M T, Roche D B 2013 *Nucleic Acids Res.* **41** W368

[35] McGuffin L J 2008 *Bioinformatics* **24** 586

[36] Uziela K, Wallner B 2016 *Bioinformatics* **32** 1411

[37] Uziela K, Shu N, Wallner B, Elofsson A 2016 *Sci. Rep.* **6** 33509

[38] Olechnovič K, Venclovas Č 2017 *Proteins Struct. Funct. Bioinf.* **85** 1131

[39] Olechnovič K, Venclovas Č 2019 *Nucleic Acids Res.* **47** W437

[40] Igashov I, Olechnovič K, Kadukova M, Venclovas Č, Grudinin S 2021 *Bioinformatics* **37** 2332

[41] Ye L S, Wu P K, Peng Z L, Gao J Z, Liu J, Yang J Y 2021 *Bioinformatics* **37** 3752

[42] Guo S S, Liu J, Zhou X G, Zhang G J 2022 *Bioinformatics* **38** 1895

[43] Liu J, Liu D, He G X, Zhang G J 2023 *Proteins Struct. Funct. Bioinf.* **91** 1861

[44] Liu J, Zhao K L, Zhang G J 2023 *Brief. Bioinform.* **24** bbac507

[45] Kryshtafovych A, Barbato A, Fidelis K, Monastyrskyy B, Schwede T, Tramontano A 2014 *Proteins Struct. Funct. Bioinf.* **82** 112

[46] Kryshtafovych A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A 2018 *Proteins Struct. Funct. Bioinf.* **86** 345

[47] Won J, Baek M, Monastyrskyy B, Kryshtafovych A, Seok C 2019 *Proteins Struct. Funct. Bioinf.* **87** 1351

[48] Haas J, Barbato A, Behringer D, Studer G, Roth S, Bertoni M, Mostaguir K, Gumienny R, Schwede T 2018 *Proteins Struct. Funct. Bioinf.* **86** 387

[49] Jones T A, Kleywegt G J 1999 *Proteins Struct. Funct. Bioinf.* **37** 30

[50] Martin A C, MacArthur M W, Thornton J M 1997 *Proteins Struct. Funct. Bioinf.* **29** 14

[51] Keedy D A, Williams C J, Headd J J, Arendall III W B, Chen V B, Kapral G J, Gillespie R A, Block J N, Zemla A, Richardson D C, Richardson 2009 *Proteins Struct. Funct. Bioinf.* **77** 29

[52] Janin J, Henrick K, Moult J, Eyck T L, Sternberg G E, Vajda S, Vakser L, Wodak S J 2003 *Proteins Struct. Funct. Bioinf.* **52** 2

[53] Lipton Z C, Elkan C, Narayanaswamy B 2014 *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014*, Nancy, France, September 15–19, 2014 p225

[54] Ozden B, Kryshtafovych A, Karaca E 2021 *Proteins Struct. Funct. Bioinf.* **89** 1787

[55] Kwon S, Won J, Kryshtafovych A, Seok C 2021 *Proteins Struct. Funct. Bioinf.* **89** 1940

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[56] Lobo J M, Jiménez-Valverde A, Real R 2008 *Global Ecol. Biogeogr.* **17** 145

[57] Spearman correlation coefficients, differences between, Myers L, Sirois M J https://doi.org/10.1002/0471667196.ess5050.pub2 [2023-11-21]

[58] Ron K, Foster P 1998 *J. Mach. Learn.* **30** 271

[59] Wang W B, Wang J L, Li Z Y, Xu D, Shang Y 2021 *Comput. Struct. Biotechnol. J.* **19** 6282

[60] McGuffin L J, Roche D B 2010 *Bioinformatics* **26** 182

[61] McGuffin L J 2009 *Proteins Struct. Funct. Bioinf.* **77** 185

[62] Ben-David M, Noivirt-Brik O, Paz A, Prilusky J, Sussman J L, Levy Y 2009 *Proteins Struct. Funct. Bioinf.* **77** 50

[63] Alapati R, Bhattacharya D 2018 *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* Washington DC, USA, August 29–September 1, 2018 p307

[64] Cheng J L, Choe M H, Elofsson A, Han K S, Hou J, Maghrabi A H, McGuffin L J, Menéndez-Hurtado D, Olechnovič K, Schwede T , Studer G, Uziela K, Venclovas Č, Wallner B 2019 *Proteins Struct. Funct. Bioinf.* **87** 1361

[65] Bitton M, Keasar C 2022 *Sci. Rep.* **12** 14074.

[66] Ke G L, Meng Q, Finley T, Wang T F, Chen W, Ma W D, Ye Q W, Liu T Y 2017 *Adv. Neural Inf. Process. Syst.* **30** 3149

[67] Maghrabi A H, McGuffin L J 2017 *Nucleic Acids Res.* **45** W416

[68] Maghrabi A H, McGuffin L J 2020 *Protein Struct. Prediction* **2165** 69

[69] McGuffin L J, Shuid A N, Kempster R, Maghrabi A H, Nealon J O, Salehe B R, Atkins J D, Roche D B 2018 *Proteins Struct. Funct. Bioinf.* **86** 335

[70] Studer G, Rempfer C, Waterhouse A M, Gumienny R, Haas J, Schwede T 2020 *Bioinformatics* **36** 1765

[71] Benkert P, Tosatto S C, Schomburg D 2008 *Proteins Struct. Funct. Bioinf.* **71** 261

[72] Manavalan B, Lee J 2017 *Bioinformatics* **33** 2496

[73] Derevyanko G, Grudinin S, Bengio Y, Lamoureux G 2018 *Bioinformatics* **34** 4046

[74] Pagès G, Charmettant B, Grudinin S 2019 *Bioinformatics* **35** 3313

[75] Uziela K, Menéndez Hurtado D, Shu N, Wallner B, Elofsson A 2017 *Bioinformatics* **33** 1578

[76] Rother K, Hildebrand PW, Goede A, Gruening B, Preissner R 2009 *Nucleic Acids Res.* **37** D393

[77] Krivov G G, Shapovalov M V, Dunbrack Jr R L 2009 *Proteins Struct. Funct. Bioinf.* **77** 778

[78] Hurtado D M, Uziela K, Elofsson A 2018 arXiv:1804.06281 [q-bio.BM]

[79] Shuvo M H, Bhattacharya S, Bhattacharya D 2020 *Bioinformatics* **36** i285

[80] Laine E, Karami Y, Carbone A 2019 *Mol. Biol. Evol.* **36** 2604

[81] Dapkūnas J, Olechnovič K, Venclovas Č 2021 *Proteins Struct. Funct. Bioinf.* **89** 1834

[82] Cao R Z, Bhattacharya D, Hou J, Cheng J L 2016 *BMC Bioinf.* **17** 495

[83] Fischer A, Igel C 2012 *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012*, Buenos Aires, Argentina, September 3–6, 2012 p14

[84] Conover M, Staples M, Si D, Sun M, Cao R Z 2019 *Comput. Math. Biophys.* **7** 1

[85] Yu Y, Si X S, Hu C H, Zhang J X 2019 *Neural Comput.* **31** 1235

[86] Baldassarre F, Menéndez Hurtado D, Elofsson A, Azizpour H 2021 *Bioinformatics* **37** 360

[87] Shen T, Wu J X, Lan H D, Zheng L Z, Pei J G, Wang S, Liu W, Huang J Z 2021 *Proteins Struct. Funct. Bioinf.* **89** 1901

[88] Gilmer J, Schoenholz S S, Riley P F, Vinyals O, Dahl G 2017 *International Conference on Machine Learning* Sydney, Australia, August 6–11, 2017 p1263

[89] Mukherjee S, Zhang Y 2009 *Nucleic Acids Res.* **37** e83

[90] Chen X, Morehead A, Liu J, Cheng J L 2023 *Bioinformatics* **39** i308

[91] McGuffin L J, Edmunds N S, Genc A G, Alharbi S, Salehe B R, Adiyaman R 2023 *Nucleic Acids Res.* **51** W274

[92] Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S, Yim J, Ronneberger O, Bodenstein1 S, Zielinski1 M, Bridgland A, Potapenko A, Cowie A, Tunyasuvunakool K, Jain R, Clancy E, Kohli1 P, Jumper J, Hassabis D 2022 bioRxiv 2021.10.04.463034

[93] Olechnovic K, Venclovas Č 2023 *Proteins Struct. Funct. Bioinf.* **91** 1879

[94] Wang Z, Eickholt J, Cheng J L 2011 *Bioinformatics* **27** 1715

[95] He G, Liu J, Liu D, Zhang G 2023 *Brief. Bioinform.* **24** 4

[96] Ballester P J, Richards W G 2007 *J. Comput. Chem.* **28** 1711

[97] Liu J, Liu D, Zhang G 2023 bioRxiv 2023.04.24.538194

[98] Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A 2021 *Adv. Neural Inf. Process. Syst.* **34** 29287

[99] Ivankov D N, Garbuzynskiy S O, Alm E, Plaxco K W, Baker D, Finkelstein A V 2003 *Protein Sci.* **12** 2057

[100] Liu D, Zhang B, Liu J, Li H, Song L, Zhang G 2023 bioRxiv 2023.05.16.540981

[101] Satorras V G, Hoogeboom E, Welling M 2021 *International Conference on Machine Learning* Vienna, Austria, July 18–24, 2021 p9323

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

SPECIAL TOPIC—Machine learning in biomolecular simulations

Recent advances in estimating protein structure model accuracy*

Liu Dong Cui Xin-Yue Wang Hao-Dong Zhang Gui-Jun †

(*School of Information Engineering, Zhejiang University of Technology, Hangzhou* 310014, *China*)

(Received 30 June 2023; revised manuscript received 1 August 2023)

Abstract

The quality assessment of protein models is a key technology in protein structure prediction and has become a prominent research focus in the field of structural bioinformatics since advent of CASP7. Model quality assessment method not only guides the refinement of protein structure model but also plays a crucial role in selecting the best model from multiple candidate conformations, offering significant value in biological research and practical applications. This study begins with reviewing the critical assessment of protein structure prediction (CASP) and continuous automated model evaluation (CAMEO), and model evaluation metrics for monomeric and complex proteins. It primarily summarizes the development of model quality assessment methods in the last five years, including consensus methods (multi-model methods), single-model methods, and quasi-single-model methods, and also introduces the evaluation methods for protein complex models in CASP15. Given the remarkable progress of deep learning in protein prediction, the article focuses on the in-depth application of deep learning in single-model methods, including data set generation, protein feature extraction, and network architecture construction. Additionally, it presents the recent efforts of our research group in the field of model quality assessment. Finally, the article analyzes the limitations and challenges of current protein model quality assessment technology, and also looks forward to future development trends.

Keywords: protein model quality assessment, deep learning, single-model methods, complex model evaluation

PACS: 87.10.Vg, 87.14.E–, 87.16.A–, 87.55.de **DOI:** 10.7498/aps.72.20231071

* Project supported by the Scientific and Technological Innovation 2030— "New Generation Artificial Intelligence", China (Grant No. 2022ZD0115103), the National Nature Science Foundation of China (Grant No. 62173304), and the Key Project of Zhejiang Provincial Natural Science Foundation of China (Grant No. LZ20F030002).

† Corresponding author. E-mail: zgj@zjut.edu.cn

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net