

四川大學

本科生毕业论文（设计）



题 目 AlphaFold2 的蛋白结构预测与
传统同源建模预测准确性的对比研究

学 院 生命科学学院

专 业 生物科学（基地班）

学生姓名 李金汕

学 号 2018141475053 年级 2019 级

指导教师 曹洋

教务处制表

2023 年 4 月 25 日

AlphaFold2 的蛋白结构预测与传统同源建模预测准确性的对比研究

专业：生物科学（基地班）

学生：李金汕 指导老师：曹洋

【摘要】 随着时代发展，越来越多的蛋白质结构预测工具涌现，AlphaFold2 和 SWISS-MODEL 是其中的代表工具，面对未知的蛋白，如何选取合适的工具去预测它的结构成为一个值得探索的问题。本研究从多个维度考察工具优劣，以 SCOP2 数据库中的蛋白样本为数据集，评估了 AlphaFold2 和 SWISS-MODEL 的表现能力。通过 TM-score 指标对预测结果进行比对，构建出了一套完整的评估体系。该研究提供了全面的评估方法，可应用于寻找适合自身研究的预测工具，促进蛋白质结构研究的进步并有望在药物研发、生物技术和医学领域中取得新的突破。未来需不断更新完善评估方法，以精准应用于实践，推动科学研究和人类生命健康的进步。

【关键词】 AlphaFold2; SWISS-MODEL; TM-score; 蛋白结构预测

A comparison study of the protein structure prediction accuracy of AlphaFold2 and traditional homology modeling

Major: Biological Science

Student: LI Jin-shan Tutor: CAO Yang

[Abstract] With the continuous development of technology, more and more protein structure prediction tools have emerged, among which AlphaFold2 and SWISS-MODEL are representative tools. However, it remains a challenge to select appropriate tools for predicting the structure of unknown proteins. This study evaluated the performance of AlphaFold2 and SWISS-MODEL from multiple perspectives, using the protein samples in the SCOP2 database as the dataset. A comprehensive evaluation system was established by comparing the prediction results with the original structures using the TM-score index. This study provided a comprehensive evaluation method for selecting suitable prediction tools for individual research, promoting the progress of protein structure research, and potentially achieving new breakthroughs in drug research, biotechnology, and medicine. In the future, it is necessary to continuously update and improve the evaluation methods for accurate application in practice, promoting scientific research and improving human health.

[Key words] AlphaFold2; SWISS-MODEL; TM-score; protein structure prediction

目 录

1 研究背景综述	6
1.1 蛋白质结构预测	6
1.1.1 研究历史背景	6
1.1.2 研究意义与展望	7
1.2 蛋白结构预测工具 AlphaFold2	8
1.2.1 原理与算法	8
1.2.2 目前表现及面对的挑战	10
1.3 蛋白结构预测工具 SWISS-MODEL	10
1.3.1 原理与算法	11
1.3.2 目前表现及面对的挑战	12
1.4 研究蛋白质结构差异的指标	13
1.4.1 Root-Mean-Square Deviation (RMSD)	13
1.4.2 Template Modeling score (TM-score)	13
1.4.3 Global Distance Test (GDT)	14
1.5 SCOP 数据库	14
1.6 本研究的目的和意义	15
2 材料与方法	16
2.1 研究对象	16
2.1.1 蛋白质数据的分类与筛选	16
2.1.2 测试数据集的建立	16
2.2 研究思路及步骤	24
2.2.1 数据采集	24
2.2.2 蛋白质结构差异的量化	24
2.2.3 两类工具表现差异的统计学研究	25
3 结果	25
3.1 不同种类蛋白表现	25
3.1.1 正向思考	30
3.1.2 反向思考	39

3.1.3 AlphaFold2 倾向性探究	40
3.2 不同结构分类蛋白表现	41
3.2.1 正向思考	45
3.2.2 AlphaFold2 倾向性探究	51
3.3 不同物种来源蛋白表现	52
3.4 结果概览	55
4 讨论	56
参考文献	58

1 研究背景综述

1.1 蛋白质结构预测

蛋白质结构预测领域是生物信息学的一个重要分支，它旨在通过计算方法推断出蛋白质序列对应的三维结构。蛋白质结构预测领域具有重要的理论意义和实际应用价值，它可以帮助揭示蛋白质的功能、进化、相互作用和设计，也可以为药物发现、疾病诊断和治疗提供有力的工具。蛋白质结构预测领域面临着许多挑战，如序列与结构之间的复杂映射关系、结构空间的高维度和多样性、实验数据的稀缺和噪声等。为了解决这些问题，蛋白质结构预测领域发展了多种方法，包括同源建模、折叠识别、模板自由建模、混合建模等。近年来，随着深度学习技术的快速发展，蛋白质结构预测领域也取得了突破性的进展，如 AlphaFold、RoseTTAFold 等方法在国际蛋白质结构预测评测（Critical Assessment of Structure Prediction, CASP）竞赛^[1]中展现了超越人类水平的性能。蛋白质结构预测领域仍然是一个活跃和前沿的研究方向，它有望为生命科学和医学带来革命性的变化。

1.1.1 研究历史背景

蛋白质结构是蛋白质研究的重要方向。传统的蛋白质结构研究主要依靠实验方法，如 X 射线衍射晶体学、核磁共振技术、电子晶体学和波普技术，但得到蛋白质结构的实验方法非常耗时，效率低下。近年来，理论计算和统计预测方法被广泛应用于蛋白质结构预测，为分子设计和蛋白质功能研究提供了可靠依据。这些方法包括计算图像分析、计算机模拟技术、遗传算法、人工神经网络方法、统计物理学方法、蒙特卡洛方法、弹性网算法和模拟退火算法等，其中人工神经网络方法和遗传算法都被广泛应用于蛋白质结构预测领域。

目前，主要有三种方法用于预测蛋白质的三维结构：同源建模法、折叠识别法和从头预测法。同源建模法适用于序列与模板序列匹配后的同源性在 30% 以上的蛋白质^[2]，常见的工具有 Modeller、SWISS-MODEL。折叠识别法适用于同源性低于 30% 的蛋白质，这种方法主要依赖于采用进化信息，如 PSI-BLAST^[3]和基于隐马尔可夫模型的方法^[4]。Threading 法^[5]是一种基于序列的折叠识别方法，通过将未知蛋白质序列与已知蛋白质结构进行比对，找出合适的匹配结构作为未知蛋白质的预测结构。但该方法的适用性取决于自然界中存在蛋白质折叠类型的数量。如果找不到合适的模板结构，只能根据序列信息进行从头预测蛋白质结构，其中 Rosetta 是目前最成功的从头预测方法。

随着深度学习技术的发展，最近出现了融合深度学习技术的蛋白质结构预测工具 AlphaFold2^[6]，其在预测方面表现出色。AlphaFold2 通过神经网络对蛋白质序列进行训练，然

后对未知蛋白质进行预测，从而获得高精度的蛋白质结构预测。

1.1.2 研究意义与展望

我们研究蛋白质结构预测是因为它是一个具有许多实际应用和科学意义的基本问题。然而，这也是一个非常困难的问题，需要复杂的方法和算法来解决。主要的挑战是，有许多可能的方式将一个蛋白质折叠成一个稳定的结构，而在这些方式中寻找最佳的方式在计算上是昂贵的，而且往往是难以解决的。因此，我们需要开发高效的启发式方法和近似技术，以便在合理的时间产生准确和可靠的预测结果。

在过去几十年里我们已经取得了非常重要的进展，特别是在利用深度学习技术和大规模数据分析的帮助下，所预测的最为精确的结构能够充分地在分子尺度上阐明蛋白质作用的生物学机制，例如催化反应、信号传导、分子识别等。这些结构信息对于药物开发等领域具有重要的指导意义，可以帮助设计和优化与蛋白质相互作用的小分子药物，从而提高药物的效果和安全性。同时，这些结构信息也可以促进生物化学研究的发展，揭示蛋白质之间的相互作用网络和功能调控机制，增进对生命现象的理解。然而，尽管蛋白质结构预测已经取得了显著的成就，但是仍然存在巨大的挑战和困难。首先，对于大多数的蛋白质结构预测的精度远远不能达到实际应用的程度，尤其是对于没有同源模板或者具有多种构象变化的蛋白质。其次，对于预测结果的评价和选择也缺乏统一和有效的标准和方法，很难判断预测结果是否可靠和有用。再有，至今为止只有 10% 被预测的结构可以用来指导实际的药物合成等生物化学方面的研究^[7]，这意味着大部分预测结果还没有得到实验验证或者应用。因此，为了提高蛋白质结构预测的准确性、可靠性和实用性，还需要不断地改进预测算法、评价方法和应用场景。

此外，如何评价不同的预测方法的优劣，目前还没有一个公认的标准。CASP 是一个每两年举办一次的蛋白质结构预测竞赛，具备强大的国际公信力，它为预测方法提供了一个公平的评估平台。CASP 使用实验测定的蛋白质结构与预测模型之间的原子距离作为评分标准，这虽是一种直观和客观的指标，但却存在一些局限性和争议^[8]。近年来，有些研究者质疑 CASP 的评分标准是否能够充分反映蛋白质结构的复杂性和多样性，因为蛋白质结构并不是刚性的，而是具有柔性和动态性，这些特征对于蛋白质执行其生物学功能是必不可少的。因此，CASP 制定的标准是否适用于描述蛋白质的动态结构还需要进一步探讨。但是目前学术界还没有提出更好的替代方案，或者说没有一个能够得到广泛认可和应用的方案。因此，寻找更合理和有效的蛋白质结构预测评价标准仍然是一个有待解决的挑战。

目前所有的预测程序都没有很好的避免过拟合问题，即预测模型过于依赖训练数据而失去泛化能力，导致在新的蛋白质序列上预测效果不佳。因为蛋白质结构预测涉及到多种复杂的因素，如序列特征、进化信息、物理化学性质、相互作用力场等，很难找到一个能够平衡各种

因素的优化目标和损失函数。目前的很多方法，包括深度学习，都不能很好地处理这些因素，所以对于不同的蛋白质，预测效果差异很大。如果要预测的蛋白质有和已知结构相似的同源蛋白，那么预测准确度会很高，因为可以利用同源蛋白的结构信息来指导预测；但是如果要预测的蛋白质没有和已知结构相似的同源蛋白，那么预测准确度就会很低，因为没有足够的信息来约束预测^[7-9]。

1.2 蛋白结构预测工具 AlphaFold2

AlphaFold 是一个人工智能（AI）程序，由 Alphabet 旗下的 DeepMind 公司开发，对蛋白质结构进行预测。该程序被设计成一个深度学习系统，使用神经网络从一个大型蛋白质序列和结构数据库中学习。AlphaFold 可以从蛋白质的氨基酸序列中预测其三维形状，精确度和速度都是前所未有的。这种能力对生物研究、药物发现和人类健康具有重大意义。

1.2.1 原理与算法

在进化过程中，蛋白质往往是保守的。例如，人类、马和鱼的血红蛋白都是由同一种蛋白质演变而来。这种同一蛋白质出现在不同物种中的现象被称为同源性（Homology）。

蛋白质的进化在很大程度上是中性的，意味着大多数突变不影响蛋白质的功能。蛋白质结构比其序列更保守。通常情况下，在很长一段时间内，物种中的氨基酸序列分化了 70%，但三维结构基本保持不变。比较来自不同物种的同源蛋白提供了有价值的信息，通常用一个二维表表示，称为多序列比对（Multiple Sequence Alignment, MSA）。来自不同物种的蛋白质的氨基酸序列被写成行，这样相应的残基就在同一列中。多序列比对中位置的一致性通常表明它们对蛋白质折叠或功能的重要性（如催化活性、与配体的相互作用、识别结合点）。蛋白质中两个氨基酸残基的共同进化往往意味着这些氨基酸之间的相互作用。这一信息可以作为预测蛋白质三维结构的基础。

AlphaFold2 的原理是利用多序列比对（MSA）来捕捉蛋白质序列之间的进化关系，以及结合物理和生物学知识来约束蛋白质结构的可能性。这种方法可以帮助算法更好地理解蛋白质序列之间的相似性和差异性，从而更准确地预测蛋白质结构。

AlphaFold2 在广义上是一个综合利用多个外部开源程序和数据库，通过蛋白质序列预测其三维结构的系统；狭义上，它是指该系统中的核心神经网络模型部分。

AlphaFold2 作为一个系统，能够根据输入的蛋白质序列预测其三维结构。为了实现这一目标，它借助了多个蛋白质数据库和开源程序。首先，AlphaFold2 使用 HMMER^[10] 软件（<http://hmmer.org/>）在 Uniprot^[11]（<https://www.uniprot.org/>）和 MGnify^[12]（<https://www.ebi.ac.uk/>）这两个序列数据库中寻找输入序列的同源物。HMMER 是一个基于

马尔可夫链/隐马尔可夫模型的软件包，这些方法在 21 世纪初的语音识别/合成领域占据主导地位。HMMER 构建并返回给定蛋白质序列及其同源物的多序列比对（MSA）。

AlphaFold2 还使用 HH-suite 包^[13]来检查是否存在已知的具有三维结构的同源物。如果存在这样的同源物，那么只需要将其作为模板，并在此基础上构建预测模型，不过，只有不到 0.1% 的蛋白质有这样的模板可用。在有可用的三维结构模板的情况下，AlphaFold2 会根据蛋白质残基与该模板之间的距离构建 pair representation；如果没有，则初始化一个带有合理默认值的 pair representation。

接下来，AlphaFold2 从每个氨基酸残基（MSA 部分）和氨基酸残基对（residue pair）中生成向量嵌入（embedding）。如果比对的数量太少，少于 30 个序列，AlphaFold2 就无法有效地工作。同时，如果序列的数量超过大约 100，考虑到这会降低训练速度，因此，比对中的序列是经过聚类得到的。

AlphaFold2 的核心部分是一个端到端的 Transformer 模型。这个模型以 MSA embedding 和 pair representation 作为输入，并输出相应的三维结构。这部分可以概括为 Evoformer 和 Structure 两个子模块。Evoformer 模块负责更新 MSA embedding 和 pair representation。本质上，它用来检测蛋白质中氨基酸之间相互作用的模式；Structure 模块根据 Evoformer 模块得到的 embedding 信息预测蛋白质的三维结构。

在 Structure 模块生成一些三维结构后，再使用 OpenMM 软件对生成的三维结构进行基于物理的弛缓。这个过程会重复三次，被称为 recycling。

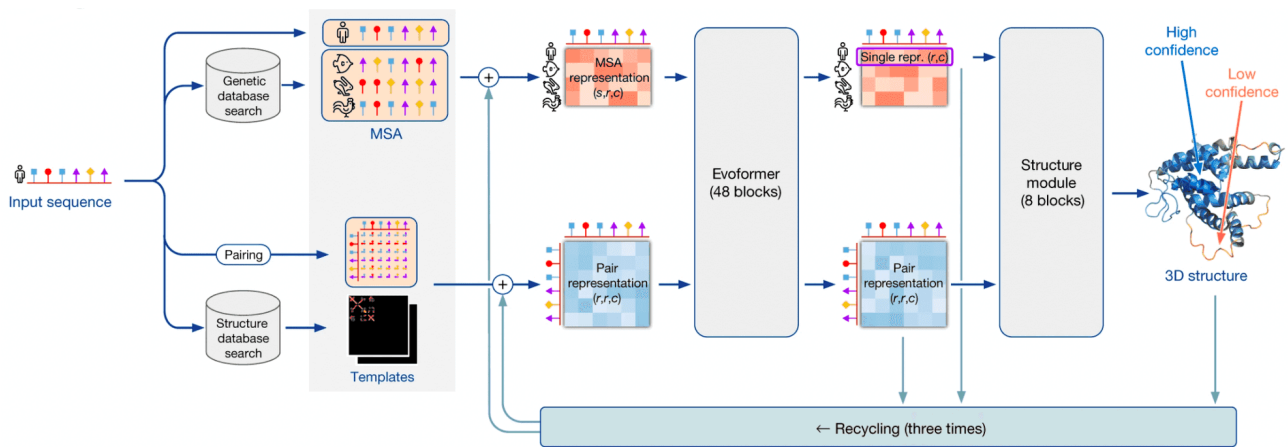


图 1 AlphaFold2 的模型架构^[6]。箭头显示了 AlphaFold2 原文^[6]描述的各个组件之间的信息流。数组的形状用括号表示，其中 s 表示序列数（原文中的 N_{seq} ），r 表示残基数（原文中的 N_{res} ），c 表示通道数。

1.2.2 目前表现及面临的挑战

AlphaFold 于 2018 年首次推出^[14], 当时它参加了蛋白质结构预测关键评估(CASP)竞赛, 这是一个两年一次的评估蛋白质结构预测方法性能的活动。AlphaFold 取得了令人瞩目的成绩, 以较大的优势超过了所有其他方法。2020 年, AlphaFold 参加了 CASP14^[6], 在那里它取得了更高的准确性, 并解决了几个多年来无法通过实验确定的挑战性蛋白质结构。

2021 年, DeepMind 和 EMBL 的欧洲生物信息学研究所(European Molecular Biology Laboratory's European Bioinformatics Institute, EMBL-EBI)推出了 AlphaFold 蛋白质结构数据库^[6, 15] (<https://www.alphafold.ebi.ac.uk/>), 这是一个公共资源, 为科学界提供了超过 2 亿个蛋白质结构预测的访问。该数据库涵盖了大部分已知的蛋白质序列空间, 包括对人类蛋白质组和其他 47 种对研究和全球健康有重要意义的关键生物的预测。该数据库还允许用户提交自己的蛋白质序列, 并在几分钟内获得 AlphaFold 预测结果。2022 年 7 月 28 日, 该团队将来自 100 万个物种的约 2 亿个蛋白质的结构上传到数据库, 几乎涵盖了地球上所有已知的蛋白质^[16]。

但值得一提的是, AlphaFold DB(Data Base)使用的是与 CASP14 版本类似的单体模型, 这意味着 AlphaFold2 只能预测单个蛋白质链的结构, 而不能预测它们与其他蛋白质或分子形成的复合体^[15]。此外, AlphaFold2 预测无序的蛋白来说是不可靠的, 但它还是会输出一个低置信度分数。由于其在某种程度上依赖于类似蛋白质之间的协同进化信息, 因此在合成蛋白质或与数据库中任何蛋白质的同源性非常低的蛋白质上可能表现不佳^[6]。遇到具有多种构象蛋白质的情况, AlphaFold2 也只能输出一种构象, 而无法控制其中的构象^[15]。该工具也只适用于不含辅因子和翻译后修饰的蛋白质结构, 这对于许多生物学相关系统来说是一个重大缺陷^[17]: 如果没有共价连接的聚糖, 人类蛋白质组中 50%至 70%的结构是不完整的^[18]。另一方面, 由于该模型是基于 PDB 模型训练的, 这些模型通常包含了这些修饰, 所以该模型预测的结构在存在离子或辅因子时经常与预期结构一致^[15]。有报道试图将 AlphaFold 应用于破坏蛋白质自然结构的各种突变, 包括与早期乳腺癌有关的突变, 证实该软件不具备预测蛋白质新突变后果的能力, 因为没有与进化有关的序列可供研究^[15]。

即使存在一些局限性, AlphaFold 仍是人工智能如何推动科学知识和创新的一个典范。通过提供对蛋白质结构和功能的准确和全面的见解, AlphaFold 可以帮助研究人员了解生命和疾病的分子基础^[19], 并加速新疗法和诊断方法的开发。

1.3 蛋白结构预测工具 SWISS-MODEL

SWISS-MODEL^[20]是一个基于网络的蛋白质结构同源建模服务器, 它的目的是让全世界的生命科学研究人员都能方便地进行蛋白质建模。它提供了三种操作模式, 分别是自动模式、

比对模式和项目模式。它还有一个存储库^[21]，每周根据最新的 UniProtKB (Knowledge Base) 蛋白组^[11]数据，对 13 个核心物种的所有蛋白质序列进行建模。

1.3.1 原理与算法

同源建模是目前最流行的蛋白质三维结构预测方法之一。该方法基于一个观察结果：当两个蛋白质的序列相似度超过 30% 时，它们的三维结构通常也是相似的。这是因为蛋白质的结构比序列更加保守，当两个蛋白质的氨基酸序列有 50% 相同时，它们 90% 的 α 碳原子的位置偏差大约只有 3Å。这为同源建模提供了可靠的依据。同源建模方法通过进行序列同源分析，找到一个已知结构的同源蛋白质模板，并以此模板构建目标蛋白质的结构模型。关键在于找到一个合适的同源模板，这可以通过检索 PDB 等蛋白质结构数据库来实现。同源建模方法是目前最为成功的蛋白质三维结构预测方法之一。但是，同源蛋白质在结构上相似的前提是它们的氨基酸序列具有相似性，因此，在序列信息不足的情况下，同源建模的精度可能会受到影响。

在 SWISS-MODEL 中默认工作流程包括以下主要步骤：首先，输入目标蛋白质的氨基酸序列或者 UniProtKB 编号^[11]。如果目标蛋白质是由不同的亚基组成的，需要为每个亚基提供相应的氨基酸序列或 UniProtKB 编号。其次，利用 BLAST^[3,22]或 HHblits^[23]两种搜索方法，在 SWISS-MODEL 模板库^[24] (SWISS-MODEL template library, SMTL) 中寻找与目标蛋白质有进化关系的结构作为模板。BLAST 作为典型的序列-序列联配算法对于密切相关的模板来说是快速且足够准确的，HHblits 则为远距离同源的情况增加了准确性。接着，模板搜索完成后，根据全球模型质量估计^[24] (Global Model Quality Estimate, GMQE) 和四级结构质量估计^[25] (Quaternary Structure Quality Estimate, QSQE) 估计的结果模型的预期质量对模板进行排名。对排名靠前的模板和排列进行比较，以验证它们是否代表替代的构象状态或覆盖目标蛋白的不同区域。在这种情况下，服务器还会自动选择多个模板，并相应建立不同的模型，呈现交互式图形视图，有助于用户分析和比较可用模板的三维结构、序列相似性和四级结构特征。然后，SWISS-MODEL 用选好的模板来建立目标蛋白质的三维结构模型，它会先把模板上和目标序列对应的部分 (Target-Template Alignment) 复制过来，然后再用一些算法来填补缺失或不同的部分，比如插入或删除一些氨基酸的区域，或者需要改变一些侧链的方向。SWISS-MODEL 依靠 OpenStructure 计算结构生物学框架^[26]和 ProMod3 建模引擎^[20]来完成这一步骤。最后，为了量化建模误差并对预期的模型精度进行估计，SWISS-MODEL 依赖了 QMEAN 评分函数^[27]。QMEAN 会考虑整个模型和每个残基位置的情况，还会用一些统计方法来比较模型和其他已知结构之间的差异。

像 ProMod-II^[28]一样，ProMod3 从笛卡尔空间 (Cartesian Space) 的对齐模板结构中提取

结构信息。它会先把目标蛋白质和模板蛋白质的序列对齐，然后根据对齐结果，把模板蛋白质的坐标复制到目标蛋白质上。如果目标蛋白质和模板蛋白质有不同的部分，比如多了或少了一些氨基酸，它会在一个结构数据库里找到合适的片段来填补空缺或替换多余的部分。如果没有找到合适的片段，则使用蒙特卡洛抽样（Monte Carlo Sampling）来完善构象。对于非保守的侧链，ProMod3 使用了 Dunbrack 实验室的侧链旋转异构体库^[29]（Backbone Dependent Rotamer Library）进行建模，以及使用基于图形的 TreePack 算法^[30]通过最小化 SCWRL4 能量函数^[31]来优化旋转体的位置，使得能量最低。最后，在建模过程中引入的小的结构畸变、不利的相互作用或冲突，都可以通过能量最小化来解决。ProMod3 使用 OpenMM 库^[32]来执行能量最小化的计算，并使用 CHARMM27 力场^[33]来参数化蛋白质的原子间相互作用。

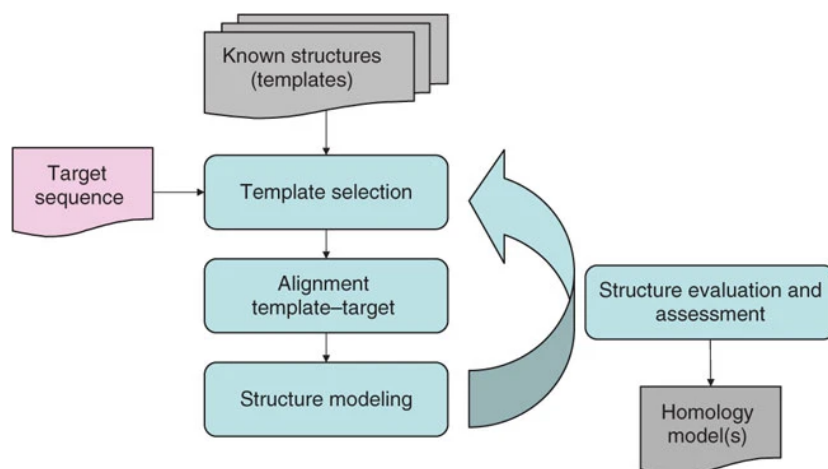


图 2 使用 SWISS-MODEL 进行蛋白质结构同源建模的工作流^[34]

1.3.2 目前表现及面对的挑战

同源建模已经成熟，成为结构生物学中的一项重要技术，大大有助于缩小已知蛋白质序列与实验确定的结构之间的差距。全自动工作流程和服务简化了同源建模过程，还允许没有特定计算专业知识的用户生成可靠的蛋白质模型，并轻松获得建模结果、可视化和注释。自从 1993 年 SWISS-MODEL 开创自动建模领域后，它一直在不断发展和改进。最近，它的功能已经扩展到同源和异源复合物的建模。从互作蛋白的氨基酸序列开始，复合物的化学计量和整体结构都是通过同源建模推断出来的。其他主要改进包括实施新的建模引擎 ProMod3^[20]和引入新的局部模型质量估计方法 QMEANDisCo^[20]。

然而，SWISS-MODEL 在预测精度和序列信息的提取等方面仍有不足。模型的准确性取决于模板结构的质量。如果没有合适的模板可用，那么生成的模型可能不够准确。它可能无法准确预测蛋白质中的某些结构特征，如非规则二级结构和某些类型的环状结构。也可能无法

准确预测蛋白质之间的相互作用。此外，有些算法仍停留在理论研究层面，预测精度还无法达到实际应用的水平，更无法解答蛋白质结构中的折叠现象，面对蛋白质结构与序列数据间的巨大差距，有必要设计更高效的算法。

尽管存在这些局限性，但不可否认 SWISS-MODEL 仍然是一个非常有用的工具，可以帮助研究人员快速生成蛋白质同源模型。在使用 SWISS-MODEL 时，应注意这些局限性带来的风险，并根据需要对生成的模型进行验证和优化。

1.4 研究蛋白质结构差异的指标

1.4.1 Root-Mean-Square Deviation (RMSD)

RMSD (Root-Mean-Square Deviation, 均方根偏差)^[35]是一种常用的度量方法，用于衡量两个分子结构之间的差异。它通过计算两个结构中相应原子之间的距离的平方和的平均值的平方根来实现。RMSD值越小，表示两个结构之间的差异越小。

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2}$$

其中 δ_i 为对应原子间的距离， i 为原子数目。在蛋白质结构预测中，RMSD 常用于衡量预测模型与实验测定结构之间的差异。如果预测模型与实验测定结构之间的 RMSD 值小于 6 Å，则表示预测模型较为准确^[36]。然而，需要注意的是，RMSD 并不是唯一衡量预测模型准确性的方法，也不是最佳方法。

1.4.2 Template Modeling score (TM-score)

TM-score^[37, 38]是一种用于评估蛋白质结构拓扑相似性的度量标准。

$$TM - score = \text{Max} \left[\frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2} \right]$$

其中 L_N 是模拟蛋白构型的长度， L_T 是与目标蛋白构型对齐的残基长度， d_i 是第 i 对对齐的残基之间的距离，而 d_0 用于归一化匹配差异‘Max’表示最佳空间重叠后的最大值。它旨在解决传统度量标准（如均方根偏差 RMSD）中的两个主要问题：

其一，RMSD 通过计算两个结构中相应原子之间的距离的平方和的平均值的平方根来实现。由于 RMSD 是对所有残基对的距离误差 δ_i 进行平均，因此几个残基对上的较大局部误差

可能导致 RMSD 值较大。另一方面, TM-score 通过在分母中引入 d_0 来自然地将较小的距离误差加权得比较大的距离误差更强。因此, 与 RMSD 相比, TM-score 值对全局结构相似性更敏感, 而不是对局部结构误差敏感。其二, TM-score 引入了一个尺度 d_0 , 使得随机结构对之间的 TM-score 大小与蛋白质长度无关, 而 RMSD 和大多数其他度量都是长度相关的指标。

RMSD 和 TM-score 都可以用来衡量两个蛋白质结构之间的相似性, 但它们之间存在一些重要的区别。在选择度量方法时, 应考虑这些区别。

1.4.3 Global Distance Test (GDT)

全局距离测试 (Global Distance Test, GDT), 也写作 GDT_TS, 表示“全局距离测试-总分”, 是一种用于衡量两个蛋白质结构之间相似性的度量计算方法。它通常用于比较蛋白质结构预测的结果与实验测定的结构。这种度量标准由 Adam Zemla 在劳伦斯利弗莫尔国家实验室开发, 并最初在 Local-Global Alignment (LGA) 程序^[39]中实现。它旨在比常用的均方根偏差 (RMSD) 度量更精确, GDT_TS 通过计算模型结构中最大的一组氨基酸残基的 α 碳原子与实验结构中相应位置之间的距离来实现。GDT_TS 得分越高, 表示模型结构与参考结构越接近。

与 RMSD 不同, GDT_TS 不容易受到离群区域 (例如, 由于单个环区建模不佳而导致的问题) 的影响。因此, 它被认为是一种比 RMSD 更准确的度量方法。GDT-TS 常用作蛋白质结构预测评估标准, 例如 CASP 竞赛中。

1.5 SCOP 数据库

研究蛋白质形态是蛋白质结构领域的重要方向, 也是构建三维结构数据库的关键。蛋白质结构的分类可以从多个层次进行, 例如折叠类型、拓扑结构、家族、超家族、结构域、二级结构以及超二级结构等。目前存在许多蛋白质分类数据库, 其中最具代表性的是由英国两个研究小组建立的数据库: 由剑桥大学的 Chothia 教授小组建立的 SCOP^[40] (<https://scop.mrc-lmb.cam.ac.uk/>) 数据库和由伦敦大学 Thornton 小组建立的 CATH^[41] (<http://cathdb.info/>) 数据库。SCOP 和 CATH 的分类方式基本相似, 但它们最大的不同在于 SCOP 主要依靠人工分类, 而 CATH 则主要采用程序进行自动化计算。

SCOP (Structural Classification of Proteins) 是由英国医学研究委员会 (Medical Research Council, MRC) 的分子生物学实验室和蛋白质工程研究中心创建和维护的。该数据库通过蛋白质的进化关系和折叠原理对已知蛋白质的三维结构进行分类, 并描述它们之间的结构和进化关系。SCOP 数据库的建立主要依靠人工注释, 同时也运用计算机程序进行支持, 以确保最大的准确性和可靠性。

SCOP 数据库按照四个层次对蛋白质结构进行分类, 包括结构类型(class)、折叠模式(fold)、

超家族 (superfamily)、家族 (family)。结构类型分为六类，分别为全 α 蛋白、全 β 蛋白、 α/β 蛋白、 $\alpha + \beta$ 蛋白、多结构域蛋白和其他类型，如膜蛋白、细胞表面蛋白、多肽、小蛋白质以及人工设计的蛋白质等。以全 β 蛋白为例，SCOP 将其折叠模式分为 61 种。其中，全 β 蛋白的第一种折叠模式为免疫球蛋白样 β 三明治结构，该折叠模式所包含的超家族数为 9 个。其中第一个超家族为免疫球蛋白，含有 5 种蛋白质家族。该超家族中的一个家族包含 C2set 结构域，该结构域又包含四种蛋白质结构域。

SCOP 数据库作为一个重要的工具，它通过蛋白质序列和结构之间的关系，帮助科学家理解蛋白质的进化关系以及判断未知蛋白质结构与已知蛋白质的相关性。SCOP 数据库的分类方法旨在创建一个全面、系统性的分类方案，以便科学家们能够更好地理解蛋白质的结构和功能。因此，该数据库对研究蛋白质发育和进化以及预测蛋白质结构和功能具有重要意义。

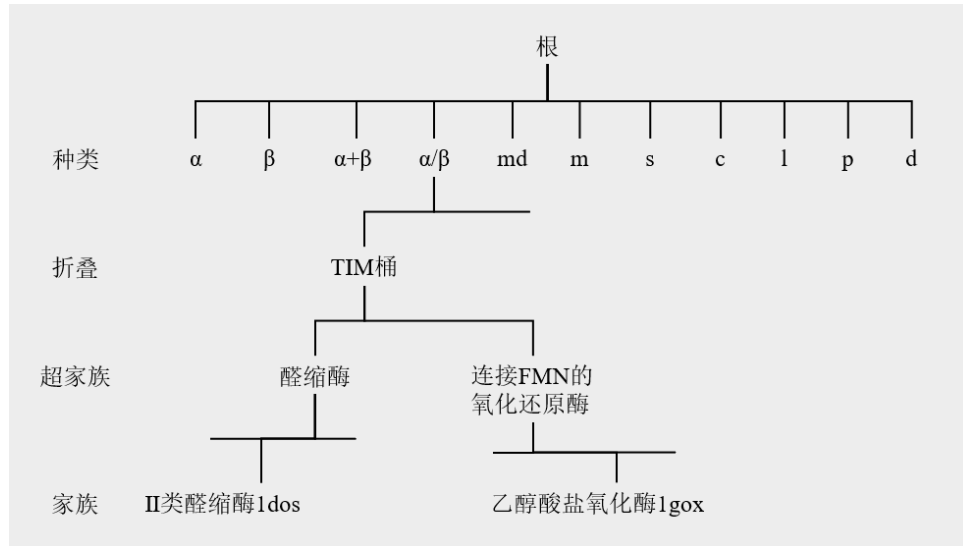


图 3 SCOP 分类的结构。显示了连接 FMN 的氧化还原酶和醛缩酶超家族的位置，lgox 和 ldos 的分类位置。

α 为所有 α ； β 为所有 β ； $\alpha + \beta$ 为 α 和 β 分离； α/β 为 α 和 β 交替；md 为多结构域；m 为膜和细胞表面；s 为小蛋白；c 为卷曲螺旋蛋白；l 为低分辨率结构；p 为多肽；d 为人工设计蛋白质。

1.6 本研究的目的是和意义

随着科技的发展，蛋白质结构预测的工具不断涌现，但预测质量却存在较大的差异，因此如何选择适合的工具是一个值得探究的问题。对于不同的蛋白质序列和研究目的，不同的预测工具可能会有不同的优劣。例如，有些工具在预测以 α -螺旋为主的蛋白质结构时表现出较高的效率和准确性。虽然也有其他更先进的预测工具，但如果只需要在有限的时间内获得大量的预测结果来进行概览性的分析，那么使用这类工具会更加合理，由此来看探究这类工具

的特点和机制是有意义的。另一方面,有些未知蛋白质在数据库中缺少相应的模板,这种情况下就需要根据预测工具在类似情况下的综合性能来进行评价,以判断其预测结果的可信度和可靠性。目前还没有建立这样的评估系统,因此填补这一空白是有价值的。

大多数科研工作者倾向于将蛋白结构预测工具用在自己的研究中,以达到提高实验效率或是增加实验证据等目的,该毕业设计项目也将从使用者的角度评估各类工具,给更多科研工作者带来一定参考价值。

2 材料与方法

2.1 研究对象

2.1.1 蛋白质数据的分类与筛选

本项研究的蛋白质数据均来自目前全球最大的实验获取的生物大分子结构 (Research Collaboratory for Structural Bioinformatics Protein Data Bank, RCSB PDB) 数据库^[42] (<https://www.rcsb.org/>)。我们根据提供了绝大多数已知结构蛋白之间结构和进化关系的蛋白结构分类 (Structural Classification of Proteins 2, SCOP2) 数据库^[40] (<https://scop.mrc-lmb.cam.ac.uk/>), 从结构层面、蛋白种类层面、以及物种来源层面进行了划分,接着在每一分类标准下筛选了比较突出的4个超家族成员,每个超家族中再筛选发表较少的6种蛋白,基于此原则,总共计入了168个(由于存在重复成员,实际数量132个)蛋白入数据集中。

2.1.2 测试数据集的建立

数据集中录入了蛋白分类标准、折叠种类、超家族、家族、PDB ID、蛋白长度信息,在获得AlphaFold2及传统同源建模工具预测结果后也将录入其中。

数据集的初步建立详见表1:

表 1 蛋白质结构预测情况预备数据集¹

Classification Criteria	Folds	Superfamilies	Families	PDB ID
All Alpha proteins	Left-handed	Leucine zipper-like	C-Jun-amino-	6EJN
All Alpha proteins	Left-handed	Leucine zipper-like	C-Jun-amino-	2W83
All Alpha proteins	Left-handed	Leucine zipper-like	Viral jun-	2H7H
All Alpha proteins	Left-handed	Rotavirus	Non-structural	1LJ2

¹ 表中的单元格内容部分由于篇幅过长进行了省略

Classification Criteria	Folds	Superfamilies	Families	PDB ID
All Alpha proteins	Left-handed	G protein-binding	Rho-associated	1UIX
All Alpha proteins	Left-handed	G protein-binding	Rho-associated	1S1C
All Alpha proteins	Left-handed	C-terminal domain of	1-	1JAD
All Alpha proteins	Left-handed	Myosin- dimerisation	Unconventional	2LW9
All Alpha proteins	Left-handed	SOAR domain-like	Stromal interaction	3TER
All Alpha proteins	Left-handed	Variant surface	Variant surface	2VSG
All Alpha proteins	Left-handed	SARAH	Ras association	2YMY
All Alpha proteins	Left-handed	LAP2alpha, C-	Lamina-associated	2V0X
All Alpha proteins	alpha-alpha	TPR-like	Tetratricopeptide	4ABN
All Alpha proteins	alpha-alpha	TPR-like	Small glutamine-rich	2VYI
All Alpha proteins	alpha-alpha	TPR-like	CRISPR system	5FSH
All Alpha proteins	alpha-alpha	Ras GEF	Ras-specific guanine	3QXL
All Alpha proteins	alpha-alpha	ARM repeat-like	Eukaryotic	1UG3
All Alpha proteins	alpha-alpha	ARM repeat-like	Ycf53 protein	1Z3X
All Alpha proteins	Long alpha-hairpin	Type III secretion	AscE	2Q1K
All Alpha proteins	Long alpha-hairpin	Homodimerisation	Protein quaking	4DNN
All Alpha proteins	Long alpha-hairpin	Homodimerisation	Protein quaking-A	2YMJ
All Alpha proteins	Long alpha-hairpin	Metal-sensitive	Sll0176 protein	5FMN
All Alpha proteins	Long alpha-hairpin	Chaperone J-domain	Heat shock protein	6RZY
All Alpha proteins	Long alpha-hairpin	Chaperone J-domain	DnaJ homolog	2LO1
All Beta proteins	Single-stranded	Pectin lyase-like	Detranase	1OGM
All Beta proteins	Single-stranded	Pectin lyase-like	Endopolygalacturona	1CZF
All Beta proteins	Single-stranded	Pectin lyase-like	Pectinesterase	5C1C

Classification Criteria	Folds	Superfamilies	Families	PDB ID
All Beta proteins	Single-stranded	Ice-binding protein-	Ice-binding protein	3VN3
All Beta proteins	Single-stranded	Ice-binding protein-	Ice-binding protein 1	6BG8
All Beta proteins	Single-stranded	Ice-binding protein-	Ice-binding protein	5UYT
All Beta proteins	beta-clip	SET domain	PR domain zinc	3DB5
All Beta proteins	beta-clip	SET domain	PR domain zinc	3DAL
All Beta proteins	beta-clip	SET domain	PR domain zinc	3EP0
All Beta proteins	beta-clip	dUTPase-like	BcDNA.LD08534	3ECY
All Beta proteins	beta-clip	dUTPase-like	Deoyuridine 5-	6MAI
All Beta proteins	beta-clip	dUTPase-like	Pol polyprotein	1DUN
All Beta proteins	Split barrel-like	FMN-binding split	Transcriptional	2OL5
All Beta proteins	Split barrel-like	FMN-binding split	Pyridoamine 5-	5BNC
All Beta proteins	Split barrel-like	FMN-binding split	Flavin-nucleotide-	4YBN
All Beta proteins	Split barrel-like	PilZ domain-like	Cellulose synthase 1	4I86
All Beta proteins	Split barrel-like	PilZ domain-like	Flagellar brake	3KYF
All Beta proteins	Split barrel-like	PilZ domain-like	Uncharacterized	5VX6
All Beta proteins	Immunoglobulin-	Immunoglobulin (Ig)	Triggering receptor	1U9K
All Beta proteins	Immunoglobulin-	Cytokine	Fibroblast growth	2P39
All Beta proteins	Immunoglobulin-	Cytokine	Multifunctional	3HAL
All Beta proteins	Immunoglobulin-	Cytokine	Fibroblast growth	1Q1U
All Beta proteins	Immunoglobulin-	PapD-like	Chaperone protein	1L4I
All Beta proteins	Immunoglobulin-	PapD-like	Putative fimbriae	2CO6
Alpha and beta proteins (a+b)	Ferredoin-like	CRISPR-associated	CRISPR system Cmr	4W8Z
Alpha and beta proteins (a+b)	Ferredoin-like	CRISPR-associated	CRISPR system Cmr	4W8V

Classification Criteria	Folds	Superfamilies	Families	PDB ID
Alpha and beta proteins (a+b)	Ferredoin-like	CRISPR-associated	BH0337 protein	4F3M
Alpha and beta proteins (a+b)	Ferredoin-like	Dimeric alpha+beta	Regulatory protein	2CG4
Alpha and beta proteins (a+b)	Ferredoin-like	Dimeric alpha+beta	Dyp-type heme-	2IIZ
Alpha and beta proteins (a+b)	Ferredoin-like	Dimeric alpha+beta	MmII domain-	2IFX
Alpha and beta proteins (a+b)	IF3-like	AlbA-like	DNA/RNA-binding	3TOE
Alpha and beta proteins (a+b)	IF4-like	EPT/RTPC-like	UDP-N-	2YVW
Alpha and beta proteins (a+b)	IF5-like	EPT/RTPC-like	UDP-N-	3R38
Alpha and beta proteins (a+b)	IF6-like	R3H domain	DNA-binding protein	1MSZ
Alpha and beta proteins (a+b)	IF7-like	YhbY-like	RNA-binding protein	1JO0
Alpha and beta proteins (a+b)	IF8-like	SepF C-terminal	Uncharacterized	3ZIG
Alpha and beta proteins (a+b)	TBP-like	TATA-bo binding	TATA-bo-binding	1MP9
Alpha and beta proteins (a+b)	TBP-like	Bet v1-like	Major allergen Pru	1E09
Alpha and beta proteins (a+b)	TBP-like	Bet v1-like	Activator of 90 kDa	1X53
Alpha and beta proteins (a+b)	TBP-like	Bet v1-like	CoG homolog	2NS9
Alpha and beta proteins (a+b)	TBP-like	Dmd-like	Dmd discriminator of	5I8J
Alpha and beta proteins (a+b)	TBP-like	YugN-like	Hypothetical	2R5X
Alpha and beta proteins (a+b)	Cystatin-like	Cystatin/monellin	Protegrin-4	1N5H
Alpha and beta proteins (a+b)	Cystatin-like	Cystatin/monellin	Colicin-M immunity	4AEQ
Alpha and beta proteins (a+b)	Cystatin-like	Cystatin/monellin	Cystatin-D	1ROA
Alpha and beta proteins (a+b)	Cystatin-like	NTF2-like	NTF2 domain-	1ZO2
Alpha and beta proteins (a+b)	Cystatin-like	NTF2-like	Protein virB8	2CC3
Alpha and beta proteins (a+b)	Cystatin-like	NTF2-like	SnoL	2GEX
Globular proteins	Ferredoin-like	CRISPR-associated	CRISPR system Cmr	4W8Z

Classification Criteria	Folds	Superfamilies	Families	PDB ID
Globular proteins	Ferredoin-like	CRISPR-associated	CRISPR system Cmr	4W8V
Globular proteins	Ferredoin-like	CRISPR-associated	BH0337 protein	4F3M
Globular proteins	Ferredoin-like	Dimeric alpha+beta	Regulatory protein	2CG4
Globular proteins	Ferredoin-like	Dimeric alpha+beta	Dyp-type heme-	2IIZ
Globular proteins	Ferredoin-like	Dimeric alpha+beta	MmII domain-	2IFX
Globular proteins	alpha-alpha	TPR-like	Tetratricopeptide	4ABN
Globular proteins	alpha-alpha	TPR-like	Small glutamine-rich	2VYI
Globular proteins	alpha-alpha	TPR-like	CRISPR system	5FSH
Globular proteins	alpha-alpha	Ras GEF	Ras-specific guanine	3QXL
Globular proteins	alpha-alpha	ARM repeat-like	Eukaryotic	1UG3
Globular proteins	alpha-alpha	ARM repeat-like	Ycf53 protein	1Z3X
Globular proteins	Immunoglobulin-	Immunoglobulin (Ig)	Triggering receptor	1U9K
Globular proteins	Immunoglobulin-	Cytokine	Fibroblast growth	2P39
Globular proteins	Immunoglobulin-	Cytokine	Multifunctional	3HAL
Globular proteins	Immunoglobulin-	Cytokine	Fibroblast growth	1Q1U
Globular proteins	Immunoglobulin-	PapD-like	Chaperone protein	1L4I
Globular proteins	Immunoglobulin-	PapD-like	Putative fimbriae	2CO6
Globular proteins	Long alpha-hairpin	Type III secretion	AscE	2Q1K
Globular proteins	Long alpha-hairpin	Homodimerisation	Protein quaking	4DNN
Globular proteins	Long alpha-hairpin	Homodimerisation	Protein quaking-A	2YMJ
Globular proteins	Long alpha-hairpin	Metal-sensitive	Slh0176 protein	5FMN
Globular proteins	Long alpha-hairpin	Chaperone J-domain	Heat shock protein	6RZY
Globular proteins	Long alpha-hairpin	Chaperone J-domain	DnaJ homolog	2LO1

Classification Criteria	Folds	Superfamilies	Families	PDB ID
Membrane proteins	Single	PetL subunit of the	Cytochrome b6-f	2E74
Membrane proteins	Single	PetM subunit of the	Cytochrome b6-f	2ZT9
Membrane proteins	Single	PetG subunit of the	Cytochrome b6-f	1Q90
Membrane proteins	Single	Sec-beta subunit	Preprotein	1RH5
Membrane proteins	Single	Oligosaccharyltransfer	Dolichyl-	1RKL
Membrane proteins	Single	Bacterial ba3 type	Cytochrome c oxidase	1XME
Membrane proteins	Transmembrane	F1F0 ATP synthase	ATP synthase subunit	1C99
Membrane proteins	Transmembrane	Htr2 transmembrane	Sensory rhodopsin II	1H2S
Membrane proteins	Transmembrane	YgaP, transmembrane	Inner membrane	2MPN
Membrane proteins	Transmembrane	Preprotein translocase	Protein-export	5AWW
Membrane proteins	Transmembrane	MerF-like	MerF	2H3O
Membrane proteins	Transmembrane	Monoxygenase subunit	Particulate methane	6CXH
Membrane proteins	Transmembrane	Porins	Porin	2POR
Membrane proteins	Transmembrane	Porins	Ferrienterobactin	1FEP
Membrane proteins	Transmembrane	Porins	Colicin I receptor	2HDF
Membrane proteins	Transmembrane	OMPA barrel-like	Lipid A deacylase	2ERV
Membrane proteins	Transmembrane	OMPA barrel-like	Photosystem II	5G3A
Membrane proteins	Transmembrane	OMPA barrel-like	Outer membrane	2X27
Membrane proteins	Toins' membrane	Colicin	Colicin-A	1COL
Membrane proteins	Toins' membrane	Diphtheria toxin,	Diphtheria toxin	1DDT
Membrane proteins	Toins' membrane	delta-Endotoxin	Pesticidal crystal	1DLC
Membrane proteins	Toins' membrane	delta-Endotoxin	Pesticidal crystal	1I5P
Membrane proteins	Toins' membrane	delta-Endotoxin	Pesticidal crystal	1CIY

Classification Criteria	Folds	Superfamilies	Families	PDB ID
Membrane proteins	Toins' membrane	Eotoin A, middle	Eotoin A	1IKQ
Fibrous proteins	Left-handed	Leucine zipper-like	C-Jun-amino-	6EJN
Fibrous proteins	Left-handed	Leucine zipper-like	C-Jun-amino-	2W83
Fibrous proteins	Left-handed	Leucine zipper-like	Viral jun-	2H7H
Fibrous proteins	Left-handed	Rotavirus	Non-structural	1LJ2
Fibrous proteins	Left-handed	G protein-binding	Rho-associated	1UIX
Fibrous proteins	Left-handed	G protein-binding	Rho-associated	1S1C
Fibrous proteins	Left-handed	C-terminal domain of	1-	1JAD
Fibrous proteins	Left-handed	Myosin- dimerisation	Unconventional	2LW9
Fibrous proteins	Left-handed	SOAR domain-like	Stromal interaction	3TER
Fibrous proteins	Left-handed	Variant surface	Variant surface	2VSG
Fibrous proteins	Left-handed	SARAH	Ras association	2YMY
Fibrous proteins	Left-handed	LAP2alpha, C-	Lamina-associated	2V0X
Fibrous proteins	Right-handed	Tetrabrachion	Tetrabrachion	1FE6
Fibrous proteins	Right-handed	Vasodilator-stimulated	Vasodilator-	1USE
Fibrous proteins	Right-handed	Microvirus pilot	Minor spike protein	4JPN
Fibrous proteins	Right-handed	VAPG-like	V-type ATP synthase,	3V6I
Fibrous proteins	Right-handed	Tetramerization	Regulatory protein	1QEY
Fibrous proteins	Right-handed	HDAC4 glutamine-	Histone deacetylase	2H8N
Fibrous proteins	Synuclein	Synuclein	Alpha-synuclein	1XQ8
Fibrous proteins	Triple HTH-repeat	phi29 head fiber stem-	Capsid fiber protein	3QC7
Fibrous proteins	Stalk segment of	Influenza	Hemagglutinin	3S12
Fibrous proteins	Stalk segment of	Influenza	Hemagglutinin	3EYJ

Classification Criteria	Folds	Superfamilies	Families	PDB ID
Fibrous proteins	Stalk segment of	Virus ectodomain	Envelope	4JPR
Fibrous proteins	Stalk segment of	Virus ectodomain	Envelope	2EBO
Non-globular/Intrinsically	Non-globular all-	Fe-only hydrogenase	Periplasmic [Fe]	1HFE
Non-globular/Intrinsically	Non-globular all-	Transducin	Guanine nucleotide-	6RMV
Non-globular/Intrinsically	Non-globular all-	Proteinase A inhibitor	Protease A inhibitor	1DPJ
Non-globular/Intrinsically	Non-globular all-	Quinohemoprotein	Quinohemoprotein	1PBY
Non-globular/Intrinsically	Non-globular all-	Lag-3 N-terminal	Protein lag-3	2FO1
Non-globular/Intrinsically	Non-globular all-	RelB-like	Uncharacterized	1WMI
Non-globular/Intrinsically	Nuclear receptor	Nuclear receptor	CREB-binding	1ZOQ
Non-globular/Intrinsically	Nuclear receptor	Nuclear receptor	Histone lysine	2C52
Non-globular/Intrinsically	Nuclear receptor	Nuclear receptor	Nuclear receptor	1KBH
Non-globular/Intrinsically	Epresed protein	Epresed protein	Uncharacterized	1WVK
Non-globular/Intrinsically	DNA ligase IV	DNA ligase IV rcc4-	DNA ligase 4	1IK9
Non-globular/Intrinsically	MTA1-R1 and -R2	MTA1-R1 and -R2	Metastasis-associated	4PC0
Non-globular/Intrinsically	Helical binder	Calmodulin binding	Small conductance	5V02
Non-globular/Intrinsically	Helical binder	BH3 domain	Bcl-2-binding	2VOF
Non-globular/Intrinsically	Helical binder	BH3 domain	Bcl-2-like protein 11	2WH6
Non-globular/Intrinsically	Helical binder	WH2 motif-like	Neural Wiskott-	3M3N
Non-globular/Intrinsically	Helical binder	WH3 motif-like	WAS/WASL-	2A41
Non-globular/Intrinsically	Helical binder	CCT motif-like	Protein TIFY 7	4RS9
Non-globular/Intrinsically	Etended binder	RBP-Jkappa-	Protein lin-12	2FO1
Non-globular/Intrinsically	Etended binder	Dynein intermediate	Cytoplasmic dynein	2P2T
Non-globular/Intrinsically	Etended binder	KASH domain-like	Nesprin-1	4DXR

Classification Criteria	Folds	Superfamilies	Families	PDB ID
Non-globular/Intrinsically	Etended binder	CB family C-terminal	Chromobo protein	3GS2
Non-globular/Intrinsically	Etended binder	Anophelin-like	Salivary anti-	4E05
Non-globular/Intrinsically	Etended binder	EspF protein repeat-	Secreted effector	2KXC

2.2 研究思路及步骤

2.2.1 数据采集

数据集中的所有目标蛋白将从 RCSB PDB 数据库中下载获得 PDB 格式的蛋白质三维结构文件，接着在基于 Python 的开源分子可视化系统软件 Pymol^[43-45] (<https://pymol.org/>) 上使用 save .fasta 命令转化为 FASTA 格式序列文件，用于后续的结构预测。

本项研究使用部署在本地实验室的 AlphaFold2 同时配合使用了搭载在 Google Colab 上的 ColabFold^[46] (<https://github.com/sokrypton/ColabFold>) 进行 AlphaFold2 预测数据集蛋白结构，从预测结果中选取排名最高位的模拟构型为研究对象，进行后续差异化比较。

同理，我们选择了由欧洲瑞士生物信息中心开发的蛋白质同源建模程序 SWISS-MODEL^[20] (<https://swissmodel.expasy.org/>) 作为传统同源建模工具代表进行待测数据集蛋白结构预测，并根据序列同源性 (Sequence Identity) 分成了 1.0-0.9、0.9-0.8、0.8-0.7、0.7-0.6、0.6-0.5、0.5-0.4、0.4-0.3 七个区间呈现结果，待服务器计算完成后，下载返回 PDB 格式文件。

2.2.2 蛋白质结构差异的量化

RMSD (Root-Mean-Square Deviation, 均方根偏差)^[35]常用作衡量两个或多个蛋白质结构之间相似性的定量指标。RMSD计算前需要把模拟蛋白构型和目标蛋白构型进行重叠，使两者重合或部分重合，

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2}$$

其中 δ_i 为对应原子间的距离， i 为原子数目。在本项研究中，计划使用Pymol进行RMSD计算的同时，直观感受模拟蛋白与目标蛋白的重叠程度。

TM-score^[37,38]是一个评估蛋白质结构拓扑学相似性的指标。TM-score的值为(0,1)，其中1表示两个结构之间完全匹配：

$$TM - score = Max \left[\frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1 + \left(\frac{d_i}{d_0} \right)^2} \right]$$

其中 L_N 是模拟蛋白构型的长度， L_T 是与目标蛋白构型对齐的残基长度， d_i 是第 i 对对齐的残基之间的距离，而 d_0 用于归一化匹配差异‘Max’表示最佳空间重叠后的最大值。按照PDB中结构的严格统计，得分低于0.17的结构对应于随机选择的不相关的蛋白质，而得分高于0.5的结构在SCOP^[40]/CATH^[41]中一般都是相同的折叠。在本项研究中，计划使用Zhang Lab的在线运行TM-align (<https://zhanggroup.org/TM-align/>) 的服务计算出TM-score。TM-align^[47]是一种用于序列独立的蛋白质结构比较的算法。对于两个未知等价的蛋白质结构，TM-align首先使用启发式动态编程迭代，根据结构相似性生成优化的残基到残基排列。在检测到对准基础上建立的两个结构最佳叠加后，最终输出衡量结构相似性的TM-score值。

2.2.3 两类工具表现差异的统计学研究

本项研究使用了Python语言，在Python 3.9.13版本环境下对已测数据集进行了统计分析，并用可视化数据呈现结果，涉及了Pandas函数包用于数据处理和分析，例如读取、筛选、分组和汇总数据；Matplotlib包用于绘制基本的图形，例如散点图、饼图等；Seaborn包用于绘制更高级的图形，例如蜂窝散点图等。详细步骤包括根据之前对蛋白的分类指定不同颜色画出散点图，横轴为蛋白长度，纵轴为预测结构差异量化指标，直观推断蛋白结构预测工具的整体水平，同时画出饼图显示每一类蛋白结构在各个预测工具中的准确性占比，进一步衡量两类工具的表现优劣。除此之外，还绘制了蜂窝散点搭配频率分布直方图观察同一工具里不同蛋白分类的准确度表现，依此判断工具的预测准确倾向性。

涉及的代码均被汇总到了Jupyter文件中以便后续研究中观察使用，您在这里可以下载源码文件 <https://github.com/Xie-Tiao/AF2-Vs-SM>。

3 结果

3.1 不同种类蛋白表现

汇总预备数据集（表1）中的Globular proteins（球蛋白）、Membrane proteins（膜蛋白）、Fibrous proteins（纤维蛋白）、Non-globular/Intrinsically unstructured proteins（非球形/本征无序蛋白）四类蛋白，在得到蛋白的AlphaFold2预测结果PDB文件后，通过线上TM-align^[47]平台 (<https://zhanggroup.org/TM-align/>) 与其在RCSB PDB数据库^[42]获得的原结构PDB文件进行

TM-score 计算后, 将结果 (AF-TM) 汇入不同种类蛋白结构预测情况表 (表 2)。同理, 在得到蛋白的 SWISS-MODEL 不同序列同源性 (Sequence Identity) 预测结果 PDB 文件后, 算出 TM-score, 根据划分的不同区间 (100%-90%、90%-80%、80%-70%、70%-60%、60%-50%、50%-40%、40%-30%) 各自汇入表 2。

表 2 不同种类蛋白结构预测情况表²

PDB ID	AF-TM	1.0-0.9	0.9-0.8	0.8-0.7	0.7-0.6	0.6-0.5	0.5-0.4	0.4-0.3
4W8Z	0.99373	0.98579	NaN	NaN	NaN	NaN	NaN	NaN
4W8V	0.79617	0.98924	NaN	NaN	NaN	NaN	NaN	NaN
4F3M	0.9341	0.99215	NaN	NaN	NaN	NaN	NaN	0.75291
2CG4	0.97439	0.99962	NaN	NaN	NaN	NaN	NaN	0.83466
2IIZ	0.97639	0.96386	NaN	NaN	NaN	NaN	0.78728	0.91986
2IFX	0.82656	NaN	0.86901	NaN	0.79059	0.83379	0.80676	NaN
4ABN	0.97883	NaN	0.9689	NaN	NaN	NaN	NaN	NaN
2VYI	0.96912	NaN	NaN	NaN	NaN	NaN	NaN	0.80715
5FSH	0.93409	0.98962	NaN	NaN	NaN	NaN	NaN	NaN
3QXL	0.99428	0.9954	NaN	NaN	NaN	NaN	NaN	0.91977
1UG3	0.9803	0.98641	NaN	NaN	NaN	NaN	NaN	NaN
1Z3X	0.98308	0.98394	NaN	NaN	NaN	NaN	0.87201	NaN
1U9K	0.97822	NaN	NaN	NaN	NaN	NaN	NaN	0.77307
2P39	0.90693	0.99978	NaN	NaN	NaN	NaN	NaN	0.78733
3HAL	0.98653	0.98232	0.93429	0.82214	NaN	NaN	NaN	NaN
1Q1U	0.98613	0.99983	NaN	0.96545	NaN	NaN	NaN	0.91599
1L4I	0.96626	0.98456	NaN	0.92503	0.9177	NaN	NaN	NaN
2CO6	0.94948	0.48312	NaN	NaN	NaN	0.46543	0.43543	0.45352
2Q1K	0.85529	NaN	0.99403	NaN	NaN	NaN	NaN	0.64667
4DNN	0.5286	NaN	0.5286	0.72466	NaN	NaN	0.67314	NaN
2YMJ	0.74298	0.6763	NaN	NaN	NaN	NaN	0.74298	0.66998

² AF-TM 指用 AlphaFold2 预测后的蛋白模型与原结构进行相似性评估所得 TM-score; 1.0-0.9 指用 SWISS-MODEL 预测后序列同源性位于 100%-90%区间的蛋白模型与原结构进行相似性评估所得 TM-score; 其余区间 0.9-0.8、0.8-0.7、0.7-0.6、0.6-0.5、0.5-0.4、0.4-0.3 同理; NaN 指该部分没有数据。

PDB ID	AF-TM	1.0-0.9	0.9-0.8	0.8-0.7	0.7-0.6	0.6-0.5	0.5-0.4	0.4-0.3
5FMN	0.93258	0.99921	NaN	NaN	NaN	NaN	NaN	0.78709
6RZY	0.93078	0.97097	NaN	NaN	NaN	0.83509	0.72939	0.84626
2LO1	0.7857	0.9855	NaN	0.73938	0.74913	0.66202	0.67747	NaN
2E74	0.2022	NaN	0.24807	0.25489	0.24844	0.24016	NaN	0.83365
2ZT9	0.12576	NaN	0.24753	0.24926	0.24797	0.24025	NaN	0.83545
1Q90	0.96902	0.98647	0.19644	0.84432	0.83127	NaN	NaN	0.83127
1RH5	0.95503	0.95188	NaN	NaN	NaN	NaN	NaN	0.70996
1RKL	0.46885	0.45848	NaN	NaN	NaN	NaN	NaN	NaN
1XME	0.99069	0.99178	0.12266	NaN	NaN	NaN	NaN	NaN
1C99	0.56505	0.5048	NaN	NaN	NaN	0.55183	NaN	0.53009
1H2S	0.97727	0.97727	NaN	NaN	NaN	NaN	0.90588	0.92754
2MPN	0.50125	0.99825	NaN	NaN	NaN	NaN	NaN	NaN
5AWW	0.89459	0.99991	NaN	NaN	NaN	NaN	0.63672	0.92869
2H3O	0.54927	0.45965	NaN	NaN	NaN	NaN	NaN	NaN
6CXH	0.95291	0.99975	NaN	0.17417	0.10238	0.12026	0.13648	NaN
2POR	0.97902	0.99982	NaN	NaN	NaN	NaN	NaN	NaN
1FEP	0.98116	0.98082	NaN	NaN	0.93739	0.93155	NaN	NaN
2HDF	0.88101	0.85017	NaN	NaN	NaN	NaN	0.80307	0.75273
2ERV	0.99778	0.99975	NaN	NaN	NaN	NaN	NaN	NaN
5G3A	0.95522	0.9536	NaN	NaN	NaN	NaN	NaN	0.13737
2X27	0.98205	0.98856	NaN	NaN	NaN	NaN	NaN	NaN
1COL	0.99594	0.98434	NaN	0.95073	0.73705	0.93544	NaN	NaN
1DDT	0.71744	0.96818	NaN	NaN	NaN	NaN	NaN	NaN
1DLC	0.99128	0.99964	NaN	NaN	NaN	NaN	NaN	NaN
1I5P	0.95586	0.99992	NaN	NaN	NaN	NaN	NaN	0.33396
1CIY	0.98092	0.99991	NaN	0.98393	NaN	NaN	NaN	NaN
1IKQ	0.83138	0.99579	NaN	NaN	NaN	NaN	NaN	NaN
6EJN	0.8558	NaN	NaN	NaN	NaN	NaN	0.63396	0.55569

PDB ID	AF-TM	1.0-0.9	0.9-0.8	0.8-0.7	0.7-0.6	0.6-0.5	0.5-0.4	0.4-0.3
2W83	0.98741	0.97902	NaN	0.83727	0.84416	NaN	NaN	NaN
2H7H	0.93075	0.99898	NaN	NaN	NaN	NaN	NaN	NaN
1LJ2	0.8344	0.9733	NaN	NaN	NaN	NaN	NaN	NaN
1UIX	0.84228	NaN	0.41233	NaN	NaN	0.71275	NaN	0.35696
1S1C	0.99185	NaN	0.91435	0.93838	NaN	NaN	NaN	NaN
1JAD	0.86451	NaN	0.9836	NaN	NaN	NaN	NaN	NaN
2LW9	0.56277	NaN	0.58441	NaN	NaN	NaN	NaN	NaN
3TER	0.73535	NaN	0.93407	NaN	NaN	NaN	NaN	0.64977
2VSG	0.93685	0.9993	NaN	NaN	NaN	NaN	NaN	NaN
2YMY	0.76836	NaN	0.76532	NaN	NaN	NaN	NaN	NaN
2V0X	0.88421	0.94574	NaN	NaN	NaN	NaN	NaN	NaN
1FE6	0.91405	0.90215	NaN	NaN	NaN	0.90288	NaN	NaN
1USE	0.89737	0.97327	NaN	NaN	NaN	NaN	NaN	NaN
4JPN	0.77402	0.99958	0.9716	0.97216	NaN	NaN	NaN	NaN
3V6I	0.90486	0.83327	NaN	NaN	NaN	NaN	NaN	NaN
1QEY	0.67491	0.96253	NaN	0.311	NaN	NaN	NaN	NaN
2H8N	0.91555	0.97786	0.97786	NaN	NaN	NaN	NaN	NaN
1XQ8	0.3375	0.21324	0.11359	NaN	NaN	NaN	NaN	NaN
3QC7	0.50504	0.99968	NaN	NaN	NaN	NaN	NaN	NaN
3S12	0.90525	0.98122	0.93429	0.97183	0.92102	NaN	NaN	NaN
3EYJ	0.94173	NaN	NaN	0.96964	0.97897	0.92353	0.62258	NaN
4JPR	0.86768	0.92913	NaN	NaN	NaN	NaN	NaN	0.80109
2EBO	0.95879	NaN	0.29494	0.41675	NaN	NaN	NaN	NaN
1HFE	0.33495	0.31874	NaN	NaN	NaN	NaN	0.5395	0.58218
6RMV	0.98917	0.99012	NaN	NaN	NaN	NaN	NaN	NaN
1DPJ	0.97422	0.99963	NaN	NaN	NaN	NaN	0.89986	0.934
1PBY	0.99067	0.99878	NaN	NaN	NaN	NaN	0.94931	NaN
2FO1	0.25212	0.98135	0.20518	NaN	0.86396	NaN	NaN	0.86923

PDB ID	AF-TM	1.0-0.9	0.9-0.8	0.8-0.7	0.7-0.6	0.6-0.5	0.5-0.4	0.4-0.3
1WMI	0.9233	0.22494	NaN	NaN	NaN	NaN	NaN	NaN
1ZOQ	0.93616	0.89447	0.13646	0.12683	NaN	NaN	NaN	NaN
2C52	0.69919	NaN	0.31661	0.2978	0.32278	0.30855	NaN	NaN
1KBH	0.56597	NaN	0.23991	0.33651	NaN	NaN	NaN	NaN
1WVK	0.29882	0.89439	NaN	NaN	NaN	NaN	NaN	NaN
1IK9	0.81141	NaN	0.73373	0.7642	0.67927	NaN	NaN	NaN
4PC0	0.97288	0.95685	NaN	NaN	NaN	NaN	NaN	NaN
5V02	0.68416	NaN	0.28386	NaN	NaN	NaN	NaN	NaN
2VOF	0.9324	0.90729	NaN	0.88645	0.91791	NaN	NaN	0.64876
2WH6	0.9432	0.98386	NaN	NaN	NaN	NaN	NaN	NaN
3M3N	0.96793	NaN	0.97203	NaN	NaN	NaN	NaN	NaN
2A41	0.97265	NaN	0.90166	0.91281	NaN	NaN	NaN	NaN
4RS9	0.1519	0.94324	NaN	NaN	NaN	NaN	NaN	NaN
2FO1	0.25212	0.98135	0.20518	NaN	0.86396	0.84843	NaN	NaN
2P2T	0.97907	NaN	0.94264	NaN	NaN	NaN	NaN	NaN
4DXR	0.95655	NaN	0.08063	0.04779	NaN	NaN	NaN	NaN
3GS2	0.8538	0.81049	NaN	NaN	NaN	NaN	NaN	NaN
4E05	0.40408	NaN	0.39758	0.83097	NaN	0.83911	NaN	0.32252
2KXC	0.83362	0.83793	NaN	NaN	NaN	NaN	NaN	0.74581

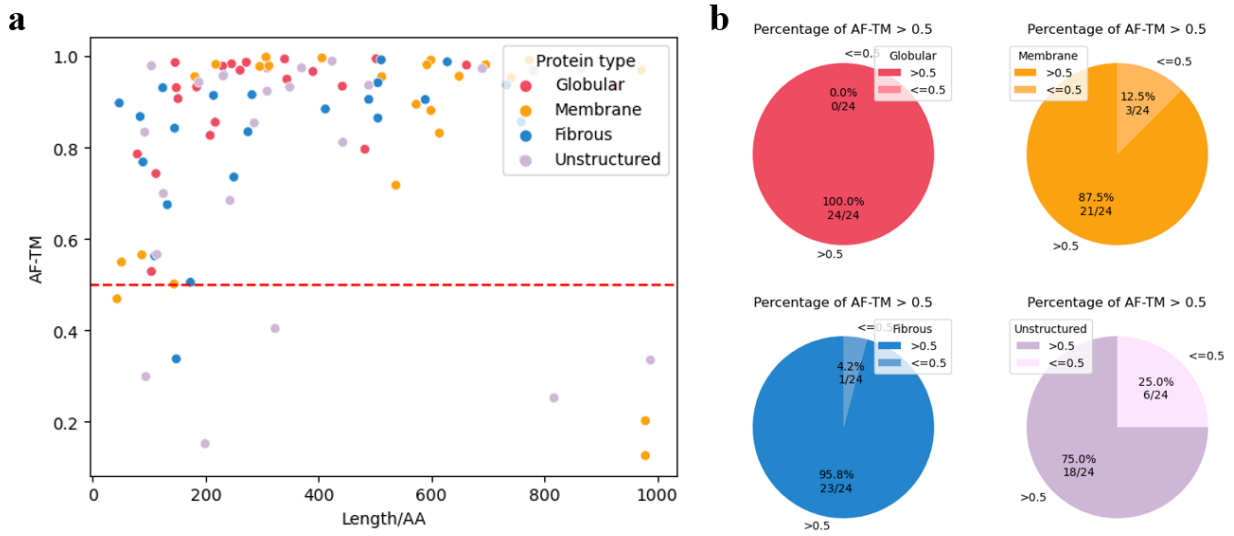


图 4 AlphaFold2 整体表现。图 a 展示了 AlphaFold2 预测不同 protein type 的蛋白的总体情况；图 b 统计分析了该数据集中模拟蛋白 TM-score 具体情况。

在获得 AlphaFold2 的预测结果数据后，我们首先需要从全局的角度审视 AlphaFold2 的表现。为此，我们可以考虑将蛋白质长度（Length/AA）用作横轴，AF-TM 用作纵轴，并以 $y=0.5$ 为分界线，该分界线代表 TM-score 在蛋白质结构预测准确性上的重要指标^[37, 38]。我们可以利用绘制散点图（图 4a）的方法，将 AlphaFold2 的预测结果标注在该图上，以此来评估其性能。除此之外，针对预备数据集分好的四类蛋白 Globular proteins（球蛋白）、Membrane proteins（膜蛋白）、Fibrous proteins（纤维蛋白）、Non-globular/Intrinsically unstructured proteins（非球形/本征无序蛋白）指定不同颜色，以方便我们观察和分析结果。

从图 4 中我们能看出，AlphaFold2 预测的 95 个样本中，绝大多数（86 个）结果均超过了 $y=0.5$ ，即 TM-score 得分高于 0.5，即 AlphaFold2 模拟的结构与样本蛋白原结构几乎一致^[37, 38]，意味着 AlphaFold2 预测的绝大多数蛋白均准确，总体表现相当优秀。仅观察图 4b 我们发现 AlphaFold2 在预测 Globular proteins（球蛋白，红色）方面表现最出色，在 Non-globular/Intrinsically unstructured proteins（非球形/本征无序蛋白，紫色）方面表现稍微逊色。由此我们可以认为，AlphaFold2 擅长预测 Globular proteins（球蛋白）一点，同时其他类型蛋白预测较不擅长，其中尤以 Non-globular/Intrinsically unstructured proteins（非球形/本征无序蛋白）最明显。

3.1.1 正向思考

通常情况下遇到的问题是，我们被给予一个未知的蛋白质，通过查询计算能够确定它的序

列同源性 (Sequence Identity) 可能非常高, 也可能稍低。我们计划使用 AlphaFold2 或传统同源建模工具 SWISS-MODEL 来预测该蛋白质的结构以接着探索其他性质。然而, 不同 Sequence Identity 值会对 SWISS-MODEL 预测的结果产生不同的影响, 因此我们需要探究 AlphaFold2 和不同 Sequence Identity 值下 SWISS-MODEL 的表现优劣, 以确定不同 Sequence Identity 值的情况下, 哪个工具的预测结果更准确。换句话说, 通过进行这项研究, 可以确定我们预测的蛋白质是否是合适的, 可以验证我们的数据分析结果是否具有说服力。

考虑到预备数据集中的样本中不是每个蛋白都涵盖所有序列同源性 (Sequence Identity) 区间, 这意味着部分区间数据不可避免会出现无效输入 (Not a Number, NaN) 情况 (表 2), 但在其他一个或多个区间又存在有效数据, 为了保证比赛的公平性, 我们决定将不同区间分开进行比较。因此, 在 SWISS-MODEL 预测出的每一列区间结果中, 我们剔除了无效数据 (NaN 所在行), 只提取包含有效结果的行, 并将它们标注在散点图上 (图 5a)。为了方便观察分析, 我们也重新绘制了有效行蛋白的 AlphaFold2 预测结果于同一幅散点图上 (图 5a)。

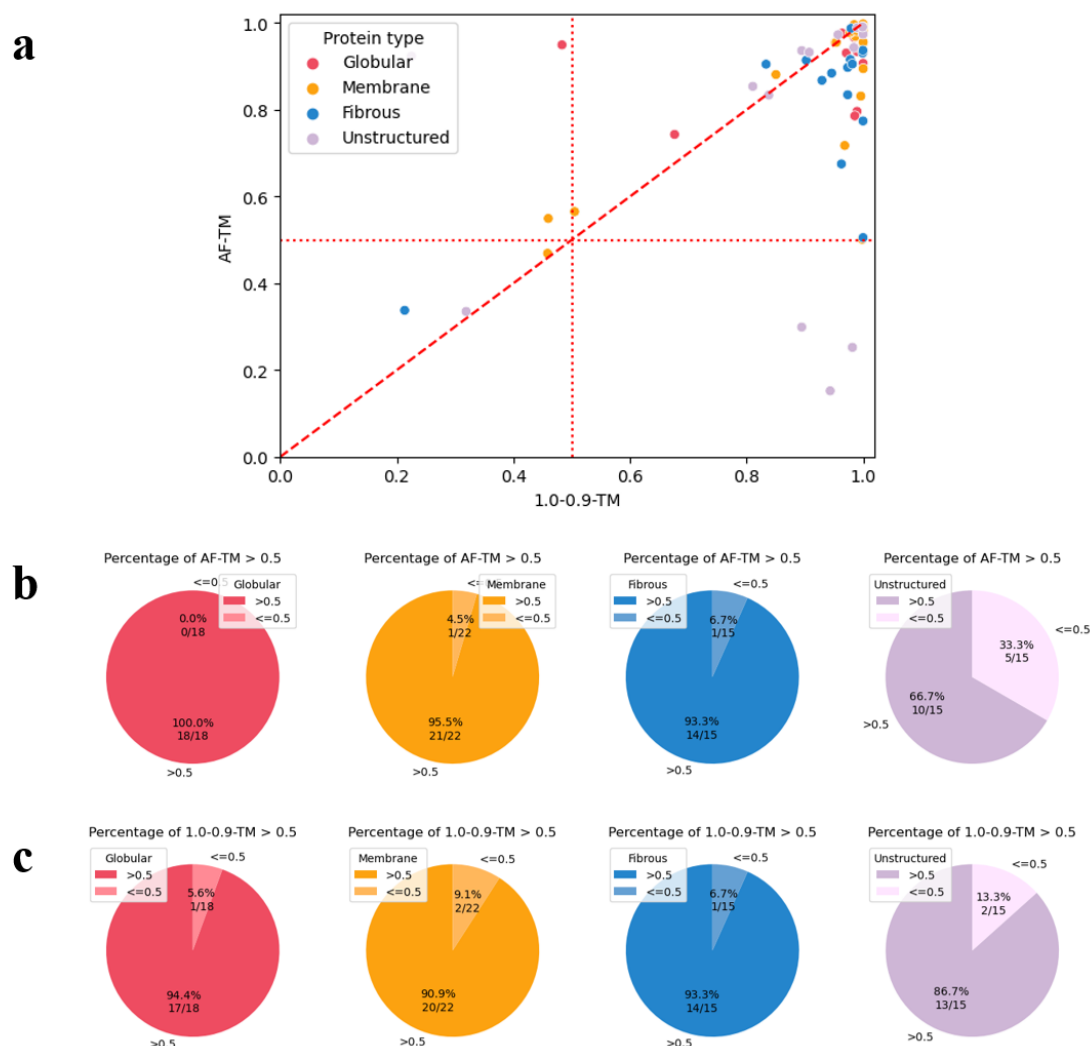


图 5 序列同源性位于(1.0-0.9)区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 100%-90%区间的不同 protein type 的蛋白的 TM-score 结果为横坐标, 以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图; 图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况; 图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

由图 5 分析可得, 当序列同源性 (Sequence Identity) 大于 90%且小于 100%的时候, 不论是 AlphaFold2 还是同源建模代表 SWISS-MODEL 表现都很不错, 势均力敌。然而在处理 Non-globular/Intrinsically unstructured proteins (非球形/本征无序蛋白) 时, SWISS-MODEL 比 AlphaFold2 表现更优秀, 取得了稍微更好的结果。

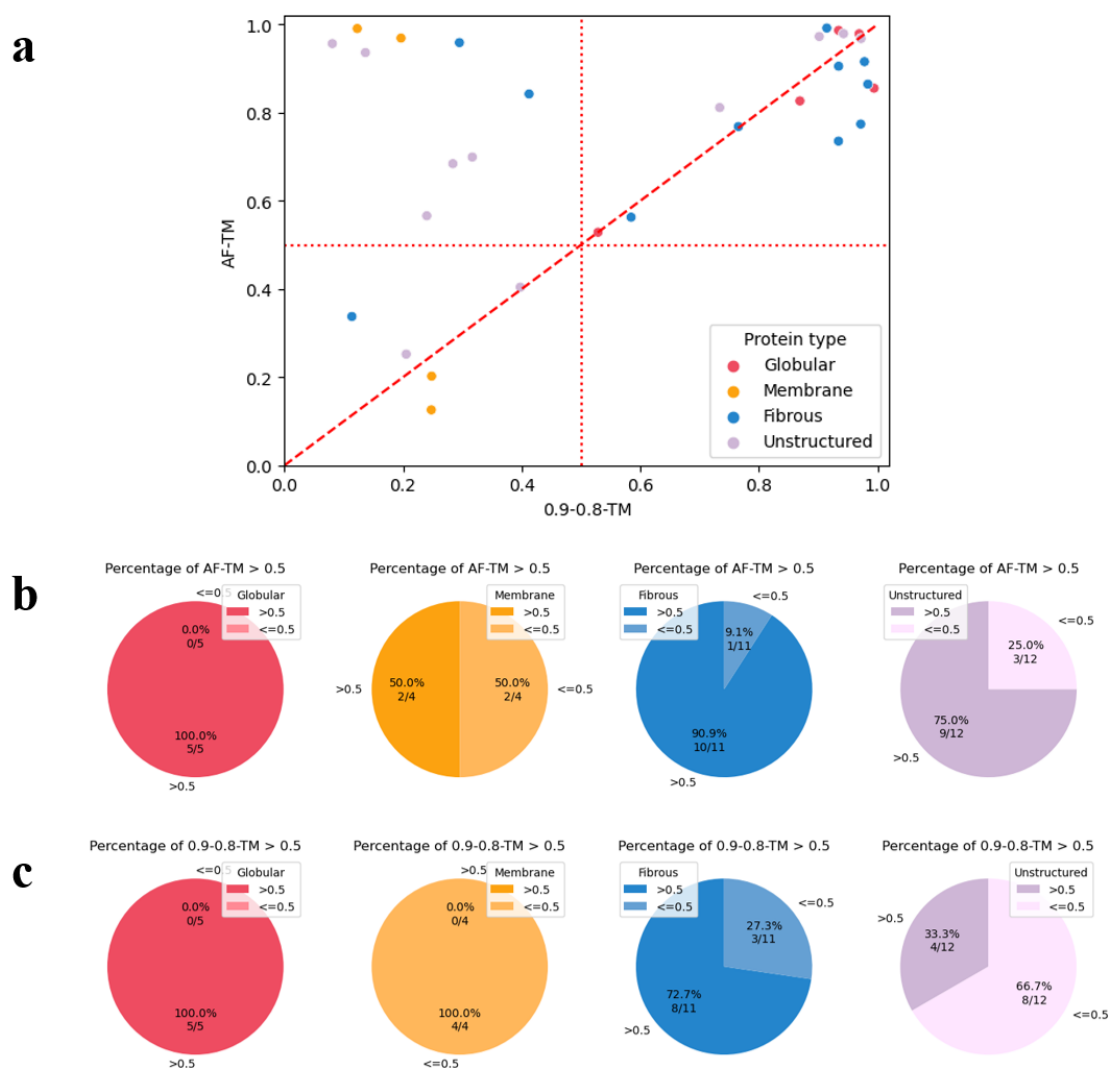


图 6 序列同源性位于(0.9-0.8)区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 90%-80%区间的不同 protein type 的蛋白的 TM-score 结果为横坐标, 以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图; 图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况; 图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

从图 6 能看出, 当序列同源性 (Sequence Identity) 大于 80%且小于 90%的时候, 不论是 AlphaFold2 还是同源建模代表 SWISS-MODEL 表现都很一般, 不分伯仲。且在处理 Non-globular/Intrinsically unstructured proteins (非球形/本征无序蛋白) 时, AlphaFold2 比 SWISS-MODEL 表现更优秀, 取得了稍微更好的结果。

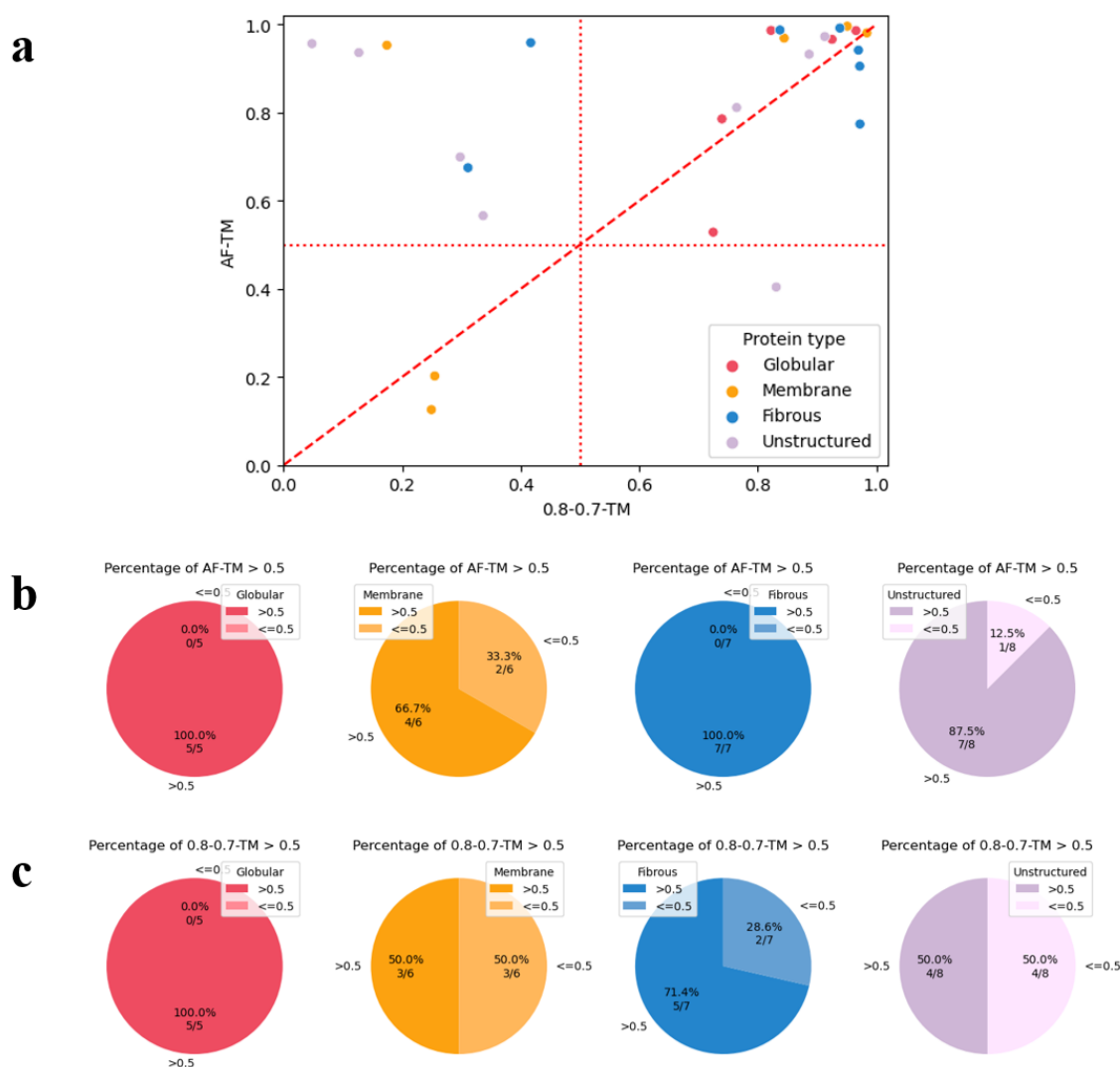


图 7 序列同源性位于 (0.8-0.7) 区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 80%-70% 区间的不同 protein type 的蛋白的 TM-score 结果为横坐标, 以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图; 图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况; 图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

在图 7 中, 当序列同源性 (Sequence Identity) 大于 70% 且小于 80% 的时候, AlphaFold2 在所有方向 (包括 Globular, Membrane 以及 Non-globular/Intrinsically unstructured proteins) 表现都优于 SWISS-MODEL, 这样的结果可能表明 AlphaFold2 在这个序列同源性范围内对蛋白质结构预测表现更佳。

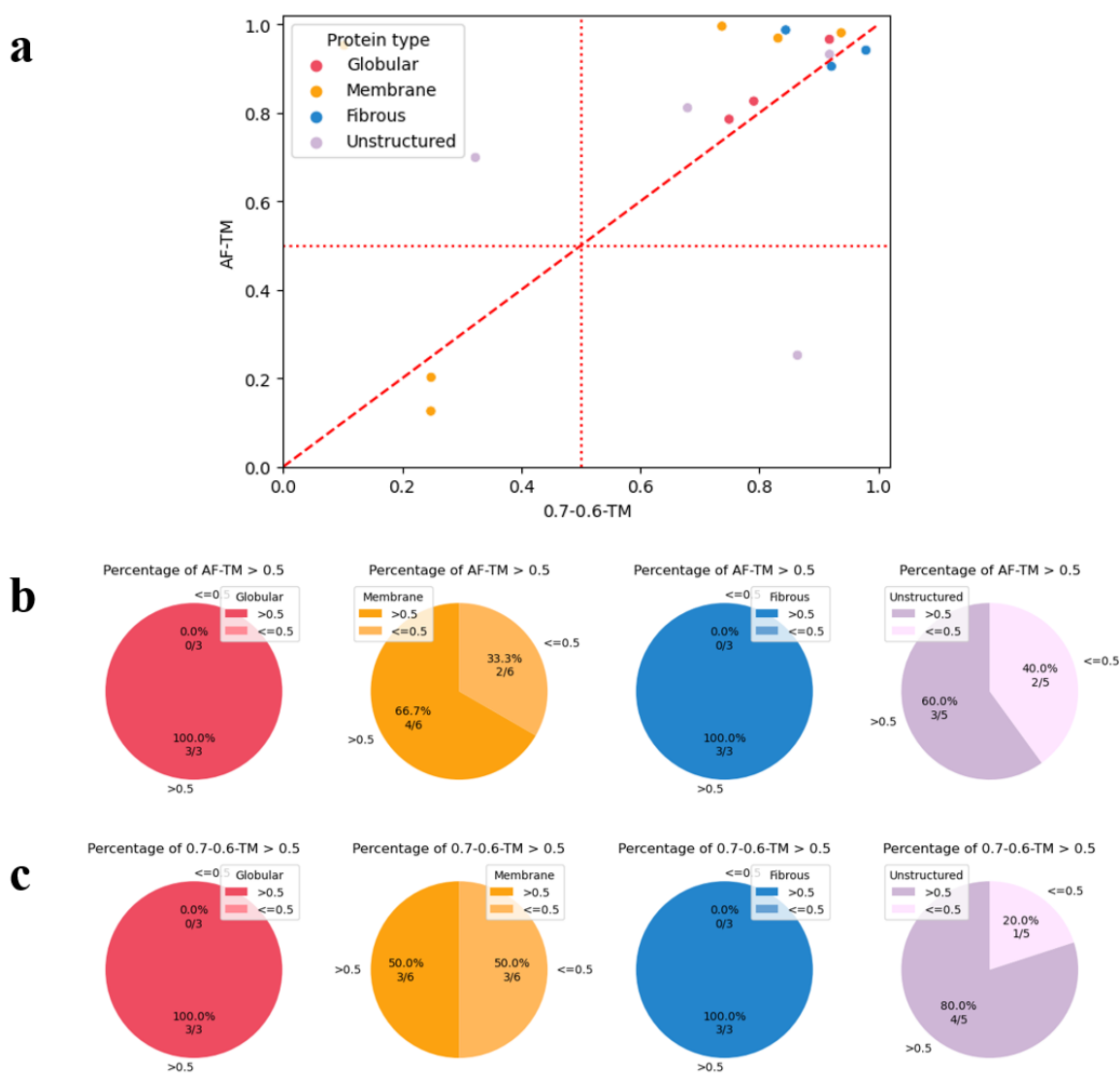


图 8 序列同源性位于 (0.7-0.6) 区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 70%-60% 区间的不同 protein type 的蛋白的 TM-score 结果为横坐标, 以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图; 图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况; 图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

在图 8 中, 当序列同源性 (Sequence Identity) 大于 60% 且小于 70% 的时候, 不论是 AlphaFold2 还是同源建模代表 SWISS-MODEL 表现都很一般, 旗鼓相当。且两者都对 Globular proteins (球蛋白)、Fibrous proteins (纤维蛋白) 具有倾向性。

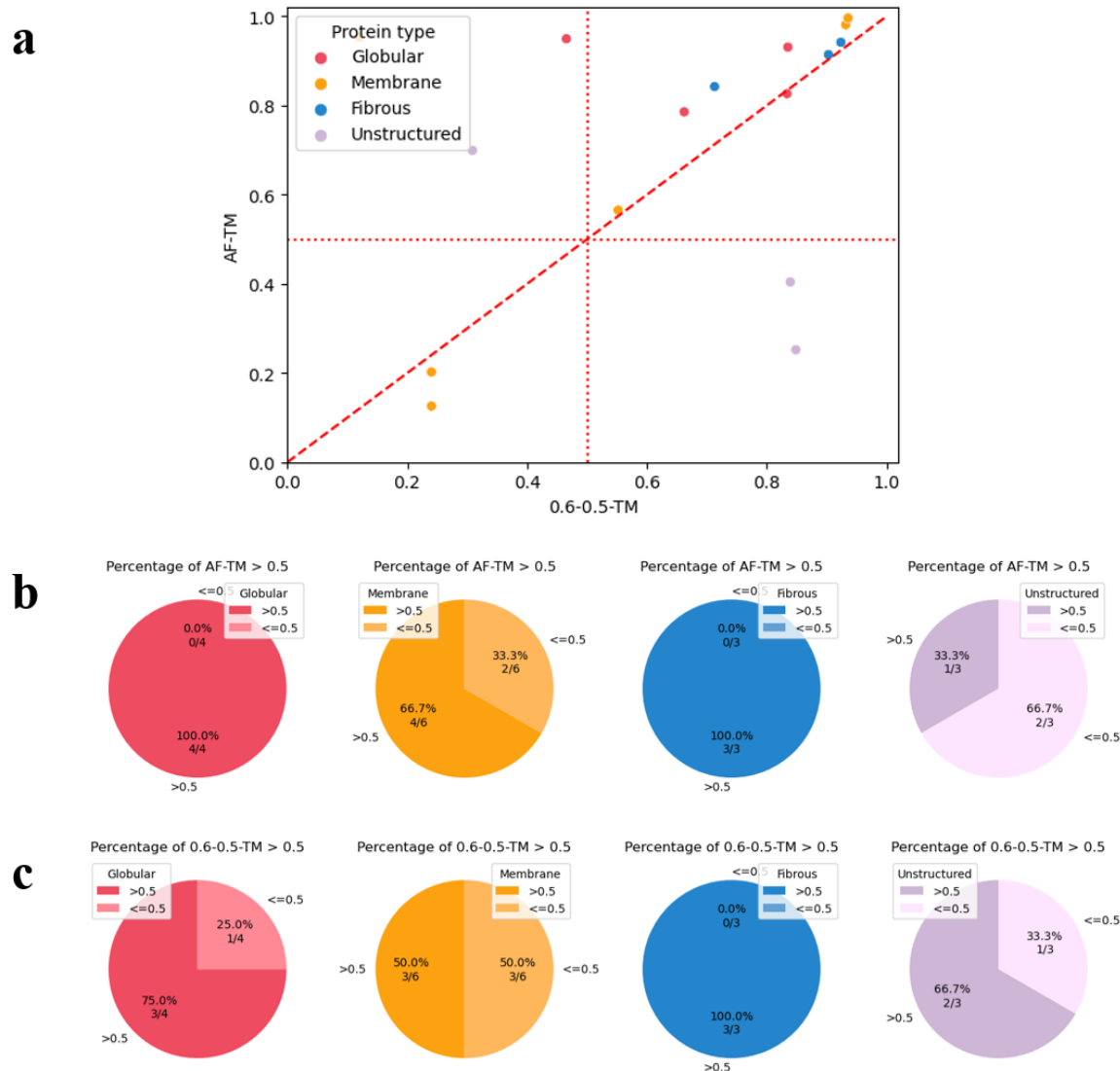


图 9 序列同源性位于 (0.6-0.5) 区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 60%-50% 区间的不同 protein type 的蛋白的 TM-score 结果为横坐标, 以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图; 图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况; 图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

由图 9 可知, 当序列同源性 (Sequence Identity) 大于 50% 且小于 60% 的时候, 不论是 AlphaFold2 还是同源建模代表 SWISS-MODEL 表现都很一般, 平分秋色。且两者都对 Globular proteins (球蛋白)、Fibrous proteins (纤维蛋白) 具有一定倾向性。

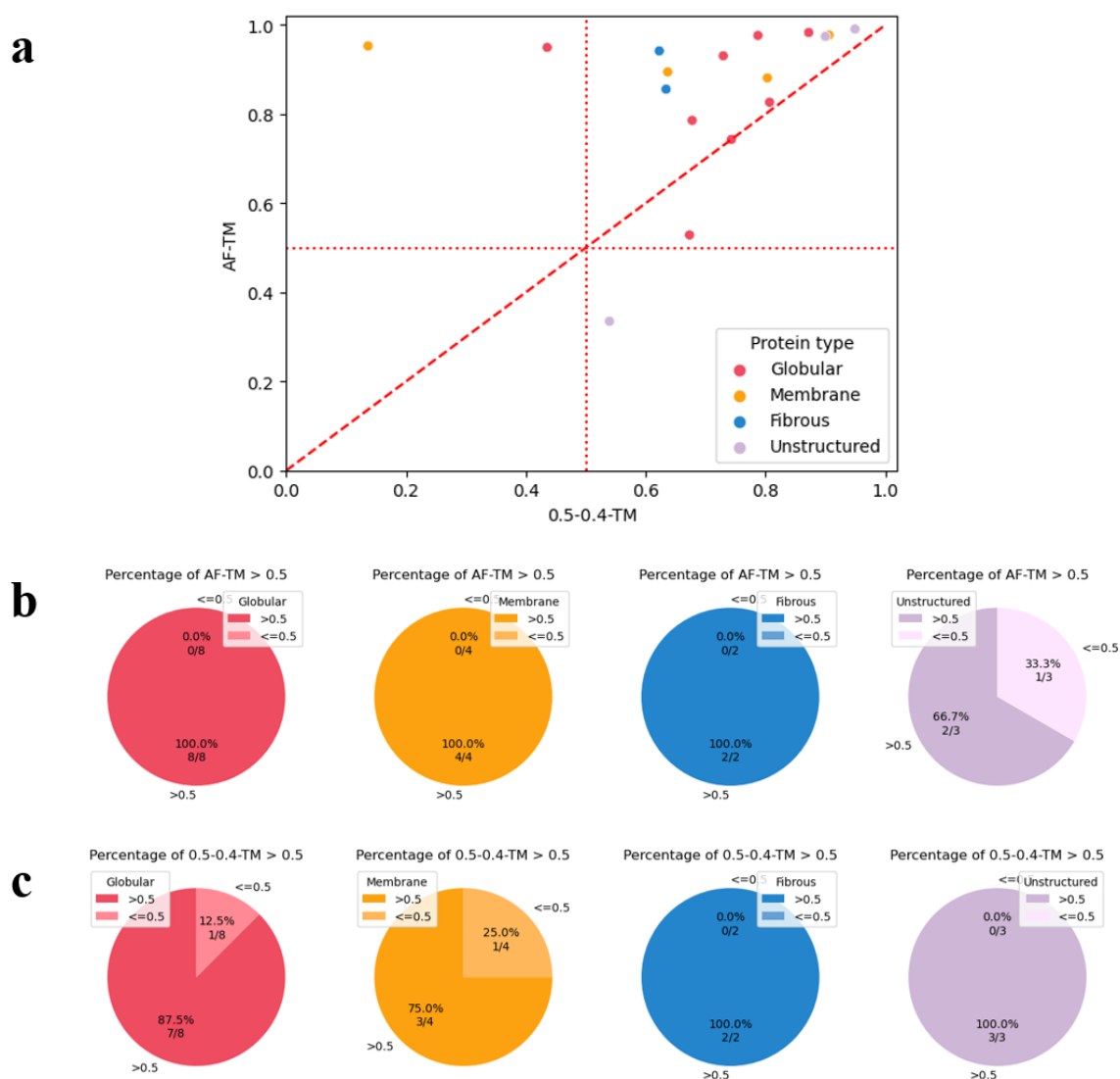


图 10 序列同源性位于 (0.5-0.4) 区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 50%-40% 区间的不同 protein type 的蛋白的 TM-score 结果为横坐标, 以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图; 图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况; 图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

从图 10 分析得知, 当序列同源性 (Sequence Identity) 大于 40% 且小于 50% 的时候, 不论是 AlphaFold2 还是同源建模代表 SWISS-MODEL 都对 Globular proteins (球蛋白)、Fibrous proteins (纤维蛋白) 表现出一定倾向性。此外, 我们还发现在预测 Membrane proteins (膜蛋白) 方面, AlphaFold2 也具有一定的倾向性, 比 SWISS-MODEL 更优, 而 SWISS-MODEL 在

预测 Non-globular/Intrinsically unstructured proteins (非球形/本征无序蛋白) 方面也有一定趋势。这可能说明在预测这些类型蛋白质的结构特征时, AlphaFold2 和 SWISS-MODEL 存在差异和特定的适用性。

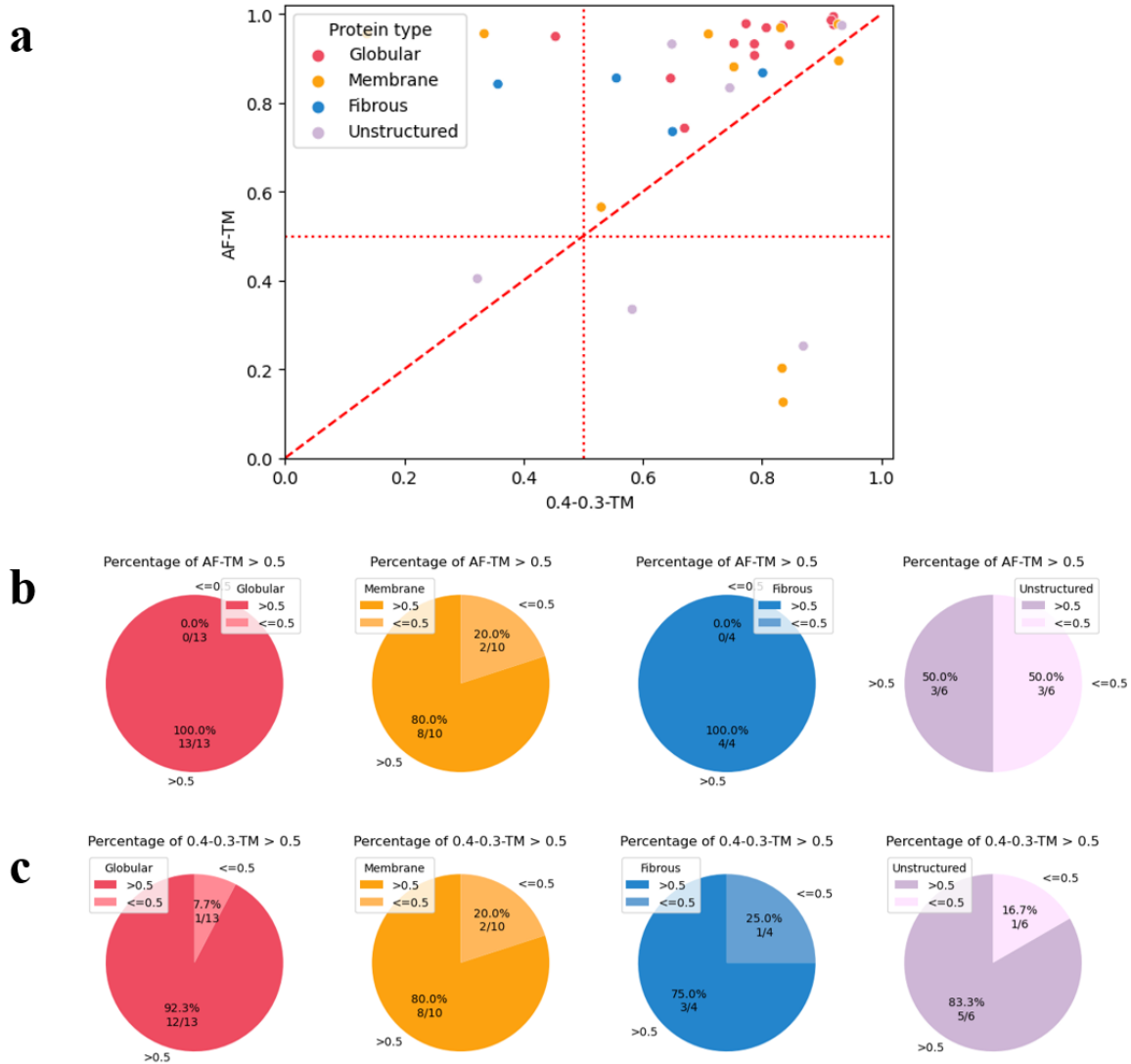


图 11 序列同源性位于 (0.4-0.3) 区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 40%-30% 区间的不同 protein type 的蛋白的 TM-score 结果为横坐标, 以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图; 图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况; 图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

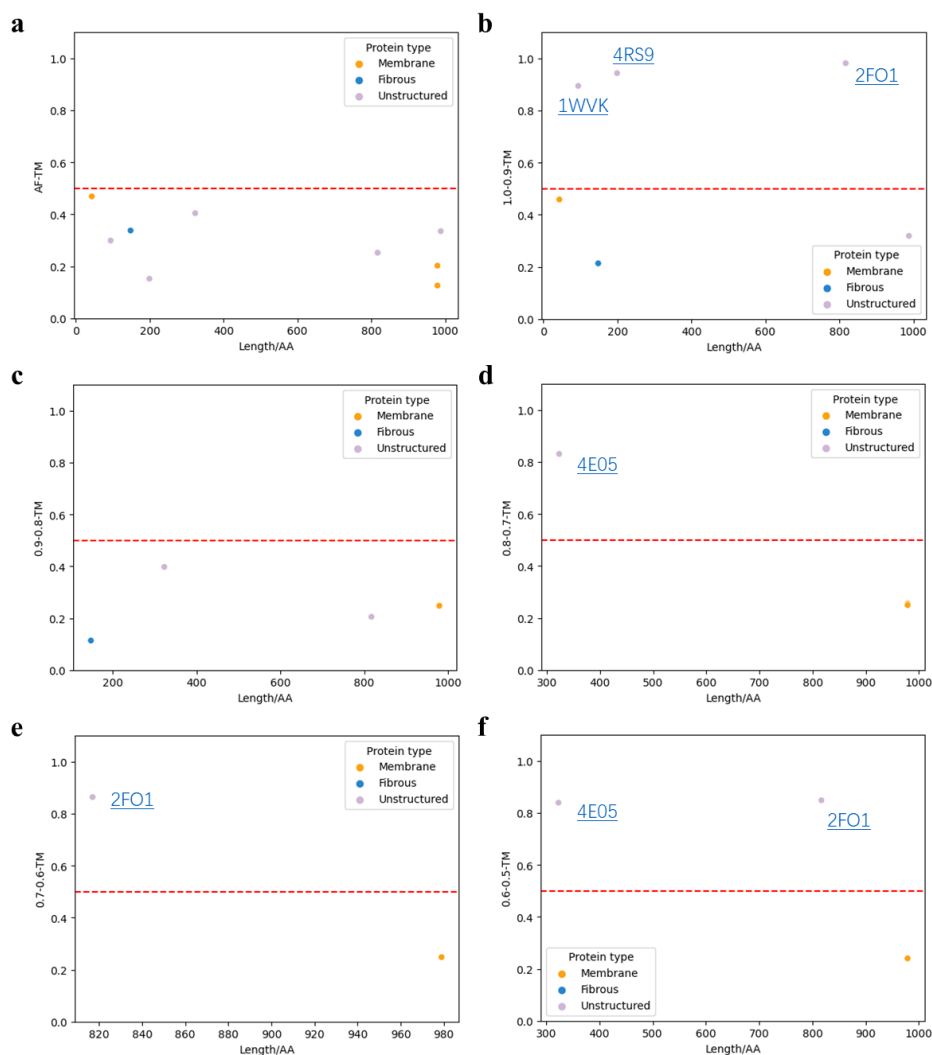
图 11 中, 在序列同源性 (Sequence Identity) 大于 30% 且小于 40% 的时候, 不论是 AlphaFold2

还是同源建模代表 SWISS-MODEL 表现都很不错，不分高下。然而在处理 Non-globular/*Intrinsically unstructured proteins*（非球形/本征无序蛋白）时，SWISS-MODEL 比 AlphaFold2 表现更优秀，取得了稍微更好的结果。

3.1.2 反向思考

通过对比图 5、9、10、11，我们可以得出一个明显的结论：在预测 Non-globular/*Intrinsically unstructured proteins*（非球形/本征无序蛋白）方面，AlphaFold2 相较 SWISS-MODEL 表现更一般。正因为观察到了这一明显的缺陷，我们打算深入挖掘 AlphaFold2 性能劣于 SWISS-MODEL 的原因。

为了达到这一目的，我们特意保留了 AlphaFold2 预测结果集合中 TM-score 小于 0.5 的 9 个样本，并剔除了其他有效数据行。接下来，我们根据不同序列同源性（Sequence Identity）区间重新比较了 SWISS-MODEL 性能，然后集中展示如下（图 12）。



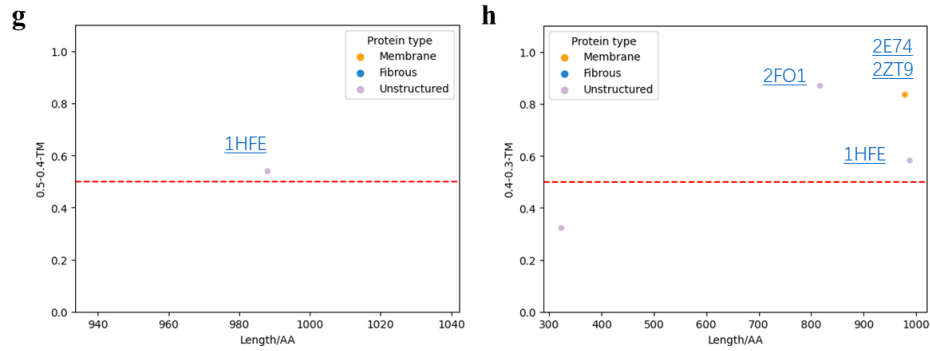


图 12 反向思考 AlphaFold2 对不同类别蛋白预测缺陷的整合图。图 a 是单独提取的 AlphaFold2 表现不足的样本散点图；图 b-h 是同一样本在不同序列同源性区间的 SWISS-MODEL 表现散点图。

在图 12 中，我们发现一些值得关注的蛋白，其预测结构模型的准确性表现引人深思。它们在 SWISS-MODEL 中表现优秀，但在 AlphaFold2 中表现一般。为了便于未来的研究和分析，我们特意标注了这些蛋白的名称以及其对应的 SWISS-MODEL 预测结果和 AlphaFold2 预测结果。这些蛋白的预测结果可能会对未来研究类似蛋白的结构和功能提供帮助和启示。通过这些特殊的蛋白范例，我们也能更好地了解 AlphaFold2 和 SWISS-MODEL 在蛋白结构预测方面的优劣，进一步提高蛋白质结构预测技术的发展水平。

3.1.3 AlphaFold2 倾向性探究

在得知 AlphaFold2 优秀的基础上，我们不禁想要进一步在不同种类蛋白内部之间比较，以确认 AlphaFold2 是否针对某一类预测会格外准确。基于图 4 的结果，我们将四类蛋白分别作为不同组别，生成了对应的蜂窝散点图，并提供了配套的直方图以方便观察和分析。这样有助于我们更深入地研究 AlphaFold2 在特定种类蛋白预测中的表现特点。

通过图 13 我们发现，图 13a 中最深的点最靠近 TM=1.0 的同时蜂窝分布也集中在 TM>0.5 的区域；其次是图 13c 在保证靠近 TM=1.0 的同时，整体分布也在往 TM>0.5 区域靠拢；然后是图 13b 有小部分 TM-score 相当低掩盖了其他相当一部分靠近满分这一优势；最后是 13d，虽然有分布在 TM>0.5 甚至 TM=1.0 的部分，但有过多 TM<0.5 的样本大大拉低了排名。由此我们简单得出结论，说明在不同种类蛋白这一赛道上，AlphaFold2 存在预测准确度的倾向性：Globular proteins（球蛋白）>Fibrous proteins（纤维蛋白）>Membrane proteins（膜蛋白）>Non-globular/Intrinsically unstructured proteins（非球形/本征无序蛋白）。即 AlphaFold2 预测 Globular proteins（球蛋白）时最准确，而预测 Non-globular/Intrinsically unstructured proteins（非球形/本征无序蛋白）相对不太准确。

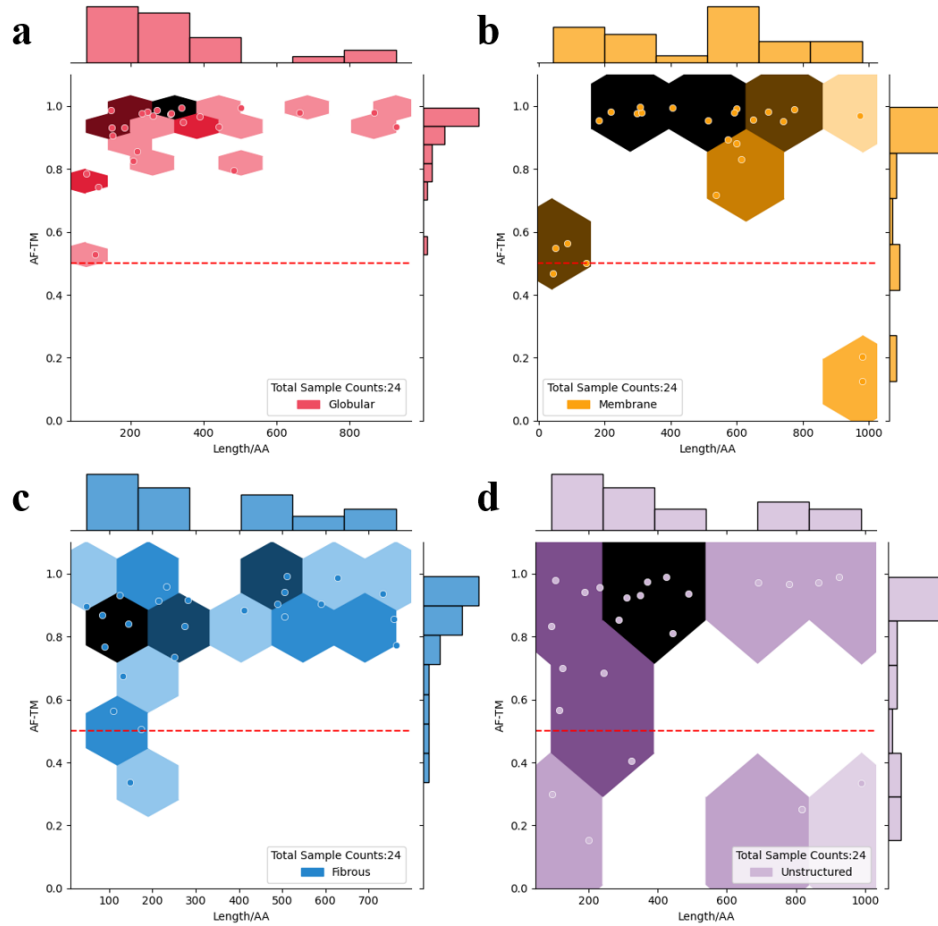


图 13 不同种类蛋白的 AlphaFold2 预测倾向性表现图。这四幅子图都以 AlphaFold2 获得的 TM-score 作为纵轴，以蛋白长度为横轴，绘制了蜂窝散点图。同时，在各自的轴向呈现了分布直方图。在蜂窝图上，不同深浅的颜色代表着样本量的大小，颜色越深的区域预示着此处分布的预测模拟蛋白数量更多。

3.2 不同结构分类蛋白表现

汇总预备数据集（表 1）中的 All alpha proteins（全 α 蛋白）、All beta proteins（全 β 蛋白）、Alpha and beta proteins (a+b)（ α 和 β 皆有且彼此分离的蛋白）三类蛋白，在得到蛋白的 AlphaFold2 预测结果 PDB 文件后，通过线上 TM-align^[47]平台 (<https://zhanggroup.org/TM-align/>) 与其在 RCSB PDB 数据库^[42]获得的原结构 PDB 文件进行 TM-score 计算后，将结果(AF-TM) 汇入不同结构分类蛋白结构预测情况表（表 3）。同理，在得到蛋白的 SWISS-MODEL 不同序列同源性（Sequence Identity）预测结果 PDB 文件后，算出 TM-score，根据划分的不同区间（100%-90%、90%-80%、80%-70%、70%-60%、60%-50%、50%-40%、40%-30%）各自汇入

表 3。

表 3 不同结构分类蛋白结构预测情况表³

PDB ID	AF-TM	1.0-0.9	0.9-0.8	0.8-0.7	0.7-0.6	0.6-0.5	0.5-0.4	0.4-0.3
6EJN	0.8558	NaN	NaN	NaN	NaN	NaN	0.63396	0.55569
2W83	0.98741	0.97902	NaN	0.83727	0.84416	NaN	NaN	NaN
2H7H	0.93075	0.99898	NaN	NaN	NaN	NaN	NaN	NaN
1LJ2	0.8344	0.9733	NaN	NaN	NaN	NaN	NaN	NaN
1UIX	0.84228	NaN	0.41233	NaN	NaN	0.71275	NaN	0.35696
1S1C	0.99185	NaN	0.91435	0.93838	NaN	NaN	NaN	NaN
1JAD	0.86451	NaN	0.9836	NaN	NaN	NaN	NaN	NaN
2LW9	0.56277	NaN	0.58441	NaN	NaN	NaN	NaN	NaN
3TER	0.73535	NaN	0.93407	NaN	NaN	NaN	NaN	0.64977
2VSG	0.93685	0.9993	NaN	NaN	NaN	NaN	NaN	NaN
2YMY	0.76836	NaN	0.76532	NaN	NaN	NaN	NaN	NaN
2V0X	0.88421	0.94574	NaN	NaN	NaN	NaN	NaN	NaN
4ABN	0.97883	NaN	0.9689	NaN	NaN	NaN	NaN	NaN
2VYI	0.96912	NaN	NaN	NaN	NaN	NaN	NaN	0.80715
5FSH	0.93409	0.98962	NaN	NaN	NaN	NaN	NaN	NaN
3QXL	0.99428	0.9954	NaN	NaN	NaN	NaN	NaN	0.91977
1UG3	0.9803	0.98641	NaN	NaN	NaN	NaN	NaN	NaN
1Z3X	0.98308	0.98394	NaN	NaN	NaN	NaN	0.87201	NaN
2Q1K	0.85529	NaN	0.99403	NaN	NaN	NaN	NaN	0.64667
4DNN	0.5286	NaN	0.5286	0.72466	NaN	NaN	0.67314	NaN
2YMJ	0.74298	0.6763	NaN	NaN	NaN	NaN	0.74298	0.66998
5FMN	0.93258	0.99921	NaN	NaN	NaN	NaN	NaN	0.78709
6RZY	0.93078	0.97097	NaN	NaN	NaN	0.83509	0.72939	0.84626

³ AF-TM 指用 AlphaFold2 预测后的蛋白模型与原结构进行相似性评估所得 TM-score; 1.0-0.9 指用 SWISS-MODEL 预测后序列同源性位于 100%-90%区间的蛋白模型与原结构进行相似性评估所得 TM-score; 其余区间 0.9-0.8、0.8-0.7、0.7-0.6、0.6-0.5、0.5-0.4、0.4-0.3 同理; NaN 指该部分没有数据。

PDB ID	AF-TM	1.0-0.9	0.9-0.8	0.8-0.7	0.7-0.6	0.6-0.5	0.5-0.4	0.4-0.3
2LO1	0.7857	0.9855	NaN	0.73938	0.74913	0.66202	0.67747	NaN
1OGM	0.99533	0.99997	NaN	NaN	NaN	NaN	0.88514	NaN
1CZF	0.99257	0.99988	NaN	0.97426	0.97687	0.98535	NaN	NaN
5C1C	0.9985	0.99454	NaN	NaN	NaN	NaN	NaN	NaN
3VN3	0.99043	NaN	0.98537	NaN	NaN	NaN	0.9714	0.38805
6BG8	0.99883	0.99575	NaN	NaN	NaN	NaN	NaN	0.80291
5UYT	0.98296	0.99995	NaN	NaN	NaN	NaN	0.16641	0.16119
3DB5	0.91566	0.94748	NaN	NaN	NaN	NaN	NaN	0.82644
3DAL	0.91392	0.92217	NaN	NaN	NaN	NaN	0.54638	0.56256
3EP0	0.78939	0.95983	NaN	NaN	NaN	NaN	0.74889	0.82067
3ECY	0.82786	0.97591	NaN	NaN	0.79127	0.7759	NaN	NaN
6MAI	0.99089	0.99987	NaN	NaN	0.95585	NaN	NaN	0.74665
1DUN	0.91806	0.97844	NaN	NaN	NaN	NaN	0.79589	0.7886
2OL5	0.95376	0.97095	NaN	NaN	NaN	NaN	NaN	NaN
5BNC	0.99216	0.99971	NaN	NaN	NaN	NaN	NaN	0.56994
4YBN	0.97472	0.99977	NaN	NaN	NaN	NaN	NaN	NaN
4I86	0.90802	0.95968	NaN	NaN	NaN	NaN	NaN	NaN
3KYF	0.84903	0.99981	NaN	NaN	NaN	NaN	NaN	NaN
5VX6	0.78967	0.75635	NaN	NaN	NaN	NaN	NaN	NaN
1U9K	0.97822	NaN	NaN	NaN	NaN	NaN	NaN	0.77307
2P39	0.90693	0.99978	NaN	NaN	NaN	NaN	NaN	0.78733
3HAL	0.98653	0.98232	0.93429	0.82214	NaN	NaN	NaN	NaN
1Q1U	0.98613	0.99983	NaN	0.96545	NaN	NaN	NaN	0.91599
1L4I	0.96626	0.98456	NaN	0.92503	0.9177	NaN	NaN	NaN
2CO6	0.94948	0.48312	NaN	NaN	NaN	0.46543	0.43543	0.45352
4W8Z	0.99373	0.98579	NaN	NaN	NaN	NaN	NaN	NaN
4W8V	0.79617	0.98924	NaN	NaN	NaN	NaN	NaN	NaN
4F3M	0.9341	0.99215	NaN	NaN	NaN	NaN	NaN	0.75291

PDB ID	AF-TM	1.0-0.9	0.9-0.8	0.8-0.7	0.7-0.6	0.6-0.5	0.5-0.4	0.4-0.3
2CG4	0.97439	0.99962	NaN	NaN	NaN	NaN	NaN	0.83466
2IIZ	0.97639	0.96386	NaN	NaN	NaN	NaN	0.78728	0.91986
2IFX	0.82656	NaN	0.86901	NaN	0.79059	0.83379	0.80676	NaN
3TOE	0.98627	0.99942	NaN	NaN	0.94623	0.90833	NaN	NaN
2YVW	0.99859	0.99987	NaN	NaN	NaN	NaN	0.96651	NaN
3R38	0.94215	0.97647	NaN	0.92903	0.92562	0.90745	0.91135	NaN
1MSZ	0.81695	0.95104	NaN	NaN	NaN	NaN	NaN	NaN
1JO0	0.9749	0.99972	NaN	0.90963	NaN	NaN	NaN	NaN
3ZIG	0.93796	0.97251	NaN	NaN	NaN	NaN	NaN	NaN
1MP9	0.95981	0.96872	NaN	NaN	NaN	NaN	0.8216	0.82575
1E09	0.91003	NaN	0.86276	0.8796	0.85138	0.86129	NaN	NaN
1X53	0.92372	0.90272	NaN	NaN	NaN	NaN	NaN	0.7912
2NS9	0.95499	0.99968	NaN	NaN	NaN	NaN	NaN	NaN
5I8J	0.91588	0.98394	NaN	NaN	NaN	NaN	NaN	NaN
2R5X	0.94599	0.94567	NaN	NaN	NaN	NaN	NaN	NaN
1N5H	0.83228	0.8825	NaN	NaN	NaN	NaN	NaN	NaN
4AEQ	0.6829	0.99944	NaN	NaN	NaN	NaN	NaN	NaN
1ROA	0.94954	0.99203	NaN	NaN	NaN	0.53952	0.8319	0.45336
1ZO2	0.99128	0.99972	NaN	NaN	NaN	0.87793	0.88509	NaN
2CC3	0.9643	0.99932	NaN	NaN	NaN	NaN	NaN	0.76118
2GEX	0.80321	0.93153	NaN	NaN	NaN	NaN	NaN	NaN

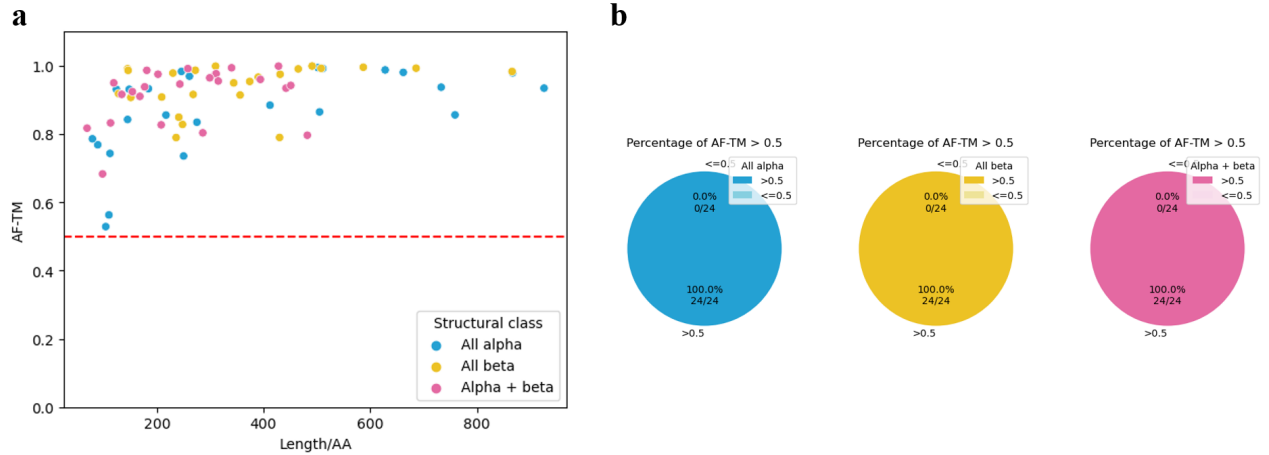


图 14 AlphaFold2 整体表现。图 a 展示了 AlphaFold2 预测不同 structural class 的蛋白的总体情况；图 b 统计分析了该数据集中模拟蛋白 TM-score 具体情况。

在获得 AlphaFold2 的预测结果数据后，参照 3.1 的思考逻辑，仍然先从全局的角度审视 AlphaFold2 的表现。为此，我们使用蛋白质长度（Length/AA）用作横轴，AF-TM 用作纵轴，并以 $y=0.5$ 为分界线，该分界线代表 TM-score 在蛋白质结构预测准确性上的重要指标^[37, 38]。我们可以利用绘制散点图（图 13a）的方法，将 AlphaFold2 的预测结果标注在该图上，以此来评估其性能。除此之外，为方便观察和分析结果，我们还将预备数据集中分好的三类蛋白 All alpha proteins（全 α 蛋白）、All beta proteins（全 β 蛋白）、Alpha and beta proteins (a+b)（ α 和 β 皆有且彼此分离的蛋白）分别用不同颜色进行标记。

从图 14 中我们能看出，AlphaFold2 预测的 72 个蛋白中，全数样本超过了 $y=0.5$ ，即 TM-score 得分高于 0.5，即 AlphaFold2 模拟的结构与样本蛋白原结构几乎一致^[37, 38]，意味着 AlphaFold2 预测的全部蛋白均准确，总体表现非常优秀。仅看图 14a 我们发现 AlphaFold2 在预测 All alpha proteins（全 α 蛋白）方面表现稍逊于另外两类蛋白。或许我们可以初步认为，相较于 All alpha proteins（全 α 蛋白），AlphaFold2 更擅长预测 All beta proteins（全 β 蛋白）和 Alpha and beta proteins (a+b)（ α 和 β 皆有且彼此分离的蛋白）一点。

3.2.1 正向思考

在研究蛋白质结构和性质时，常常需要对未知的蛋白质进行预测。为了预测这些蛋白质的结构，我们通常会使用 AlphaFold2 或传统的同源建模工具 SWISS-MODEL 来进行预测。在进行预测之前，我们需要先查询计算以确定蛋白质的序列同源性（Sequence Identity）情况，根据这个指标的不同，可能会对 SWISS-MODEL 预测的结果产生不同的影响。因此，我们需要探究 AlphaFold2 和 SWISS-MODEL 在不同 Sequence Identity 值下的表现，以确定不同 Sequence

Identity 值的情况下，哪个工具的预测结果更准确。简而言之，进行这项研究的目的是为了确定预测的蛋白质是否合适，并验证我们的数据分析结果是否可靠。

考虑到预备数据集中的样本没有涵盖所有的序列同源性（Sequence Identity）区间，可能会导致部分区间出现无效输入（Not a Number, NaN）情况（表 3）。然而，在其他一个或多个区间又存在有效数据，为了保证比赛的公平性，我们决定将不同区间分开进行比较。因此，在 SWISS-MODEL 预测出的每一列区间结果中，我们移除了无效数据（NaN 所在行），只提取包含有效结果的行，并将它们标注在散点图上（图 15a）。同时，我们还重新绘制了包含有效行蛋白的 AlphaFold2 预测结果于同一张散点图上（图 15a），以方便观察和分析结果。

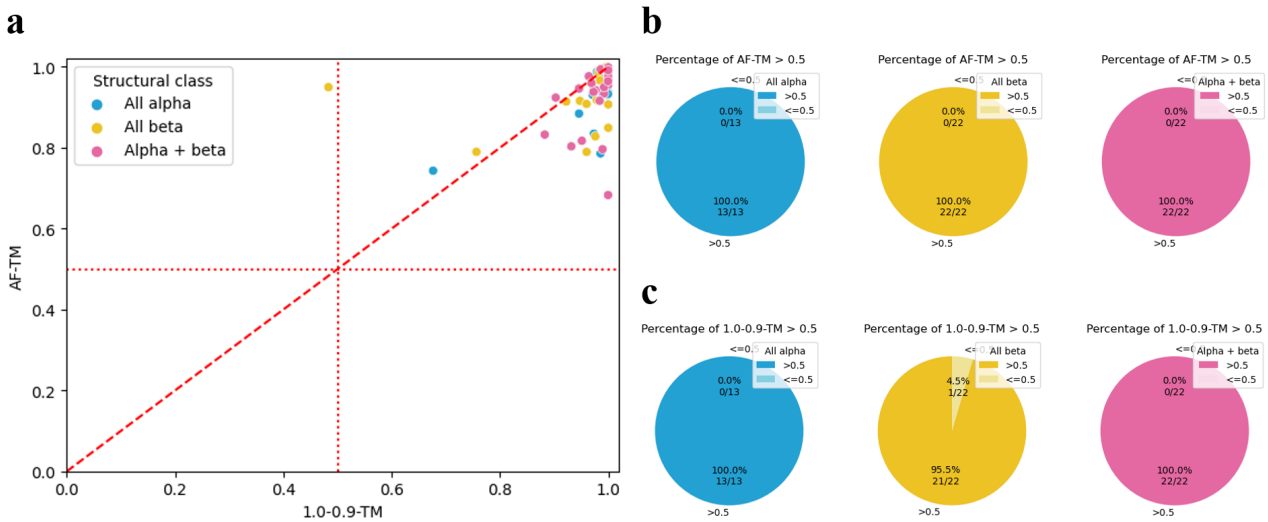


图 15 序列同源性位于(1.0-0.9)区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性（Sequence Identity）位于 100%-90%区间的不同 structural class 蛋白的 TM-score 结果为横坐标,以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图;图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况;图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

在图 15 中，当序列同源性（Sequence Identity）大于 90%且小于 100%的时候，不论是 AlphaFold2 还是同源建模代表 SWISS-MODEL 表现都很不错。虽然 SWISS-MODEL 在 All beta proteins（全 β 蛋白）表现稍有不足之处，但我们也不能忽视其整体结果都比 AlphaFold2 更接近 TM=1.0 这一特点。

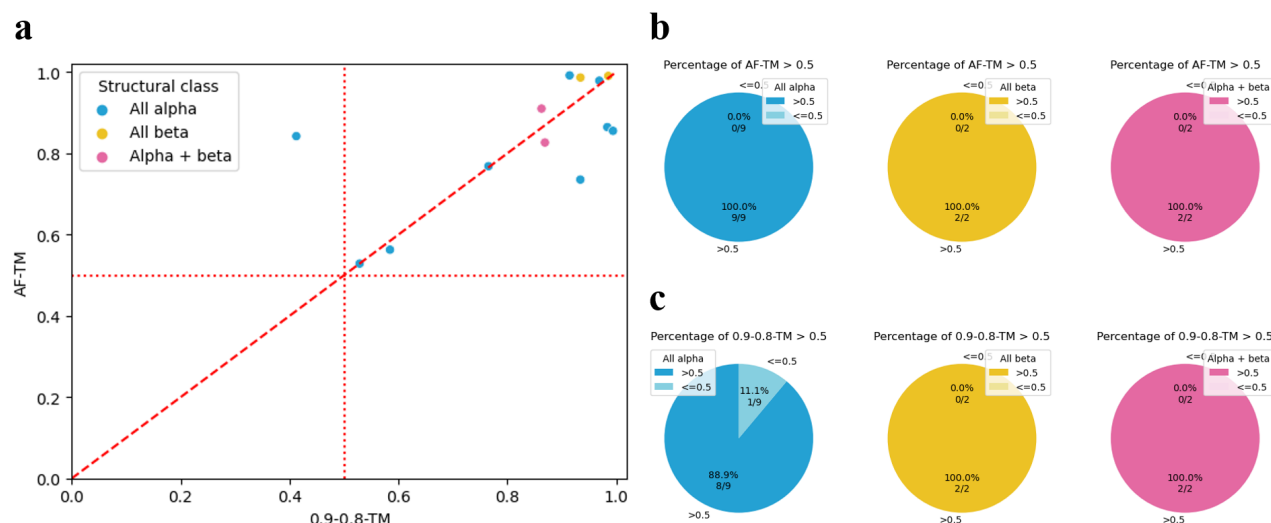


图 16 序列同源性位于 (0.9-0.8) 区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 90%-80% 区间的不同 structural class 蛋白的 TM-score 结果为横坐标, 以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图; 图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况; 图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

观察图 16 时, 我们能轻易看出, 当序列同源性 (Sequence Identity) 大于 80% 且小于 90% 的时候, SWISS-MODEL 在 All alpha proteins (全 α 蛋白) 上的表现是不如 AlphaFold2 稳定的。

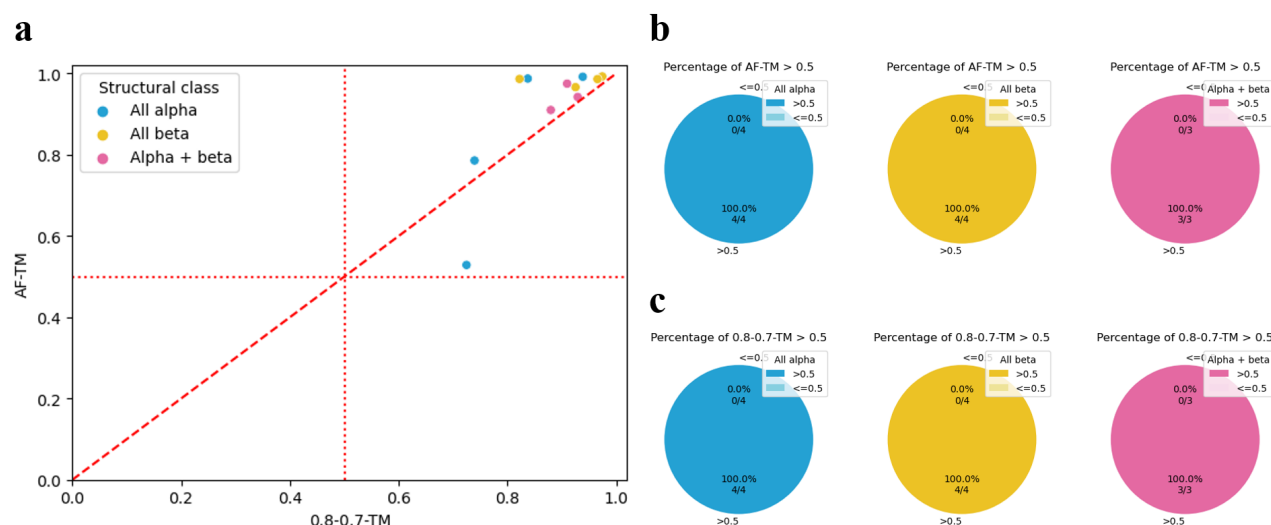


图 17 序列同源性位于 (0.8-0.7) 区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 80%-70% 区间的不同 structural class 蛋白的 TM-score 结果为横坐标，以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图；图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况；图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

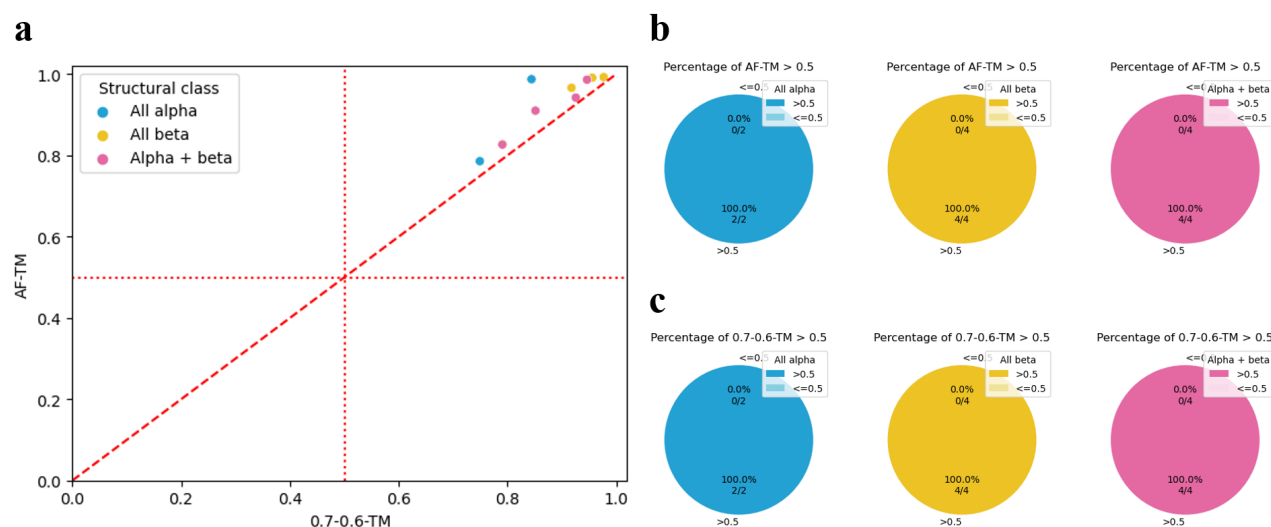


图 18 序列同源性位于 (0.7-0.6) 区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 70%-60% 区间的不同 structural class 蛋白的 TM-score 结果为横坐标，以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图；图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况；图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

在图 17 中，当序列同源性（Sequence Identity）大于 70%且小于 80%的时候，不论是 AlphaFold2 还是同源建模代表 SWISS-MODEL 表现都很出色。

同样的，从图 18 中分析可得，在序列同源性（Sequence Identity）大于 60%且小于 70%的时候，不论是 AlphaFold2 还是同源建模代表 SWISS-MODEL 表现都很出色，且 AlphaFold2 更优秀一点。

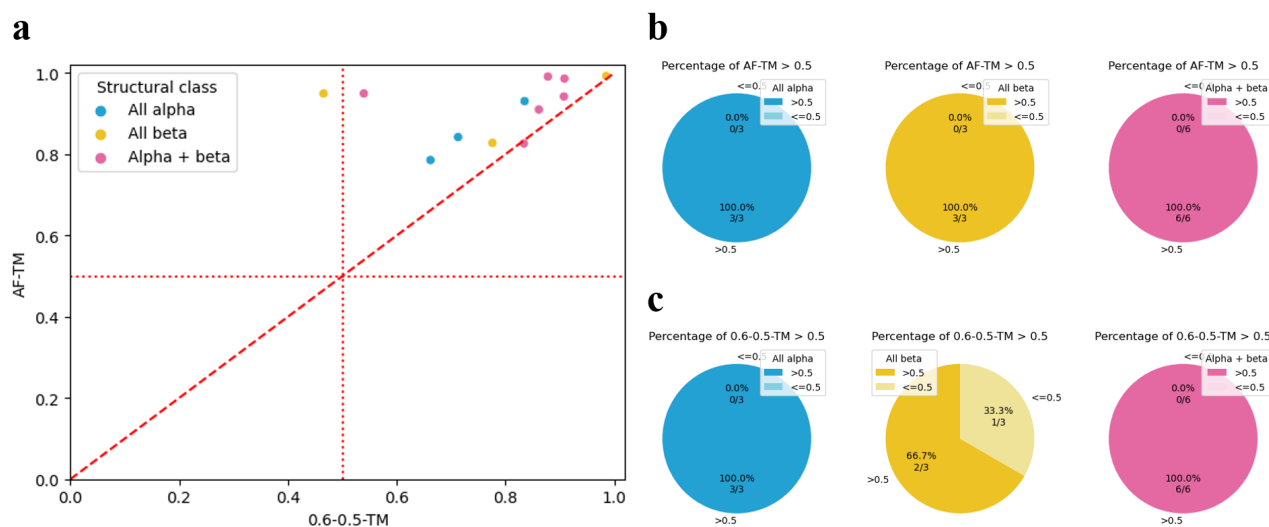


图 19 序列同源性位于(0.6-0.5)区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性(Sequence Identity)位于 60%-50%区间的不同 structural class 蛋白的 TM-score 结果为横坐标，以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图；图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况；图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

观察图 19 时，发现当序列同源性(Sequence Identity)大于 50%且小于 60%的时候，SWISS-MODEL 在 All beta proteins（全 β 蛋白）方面表现逊于 AlphaFold2。

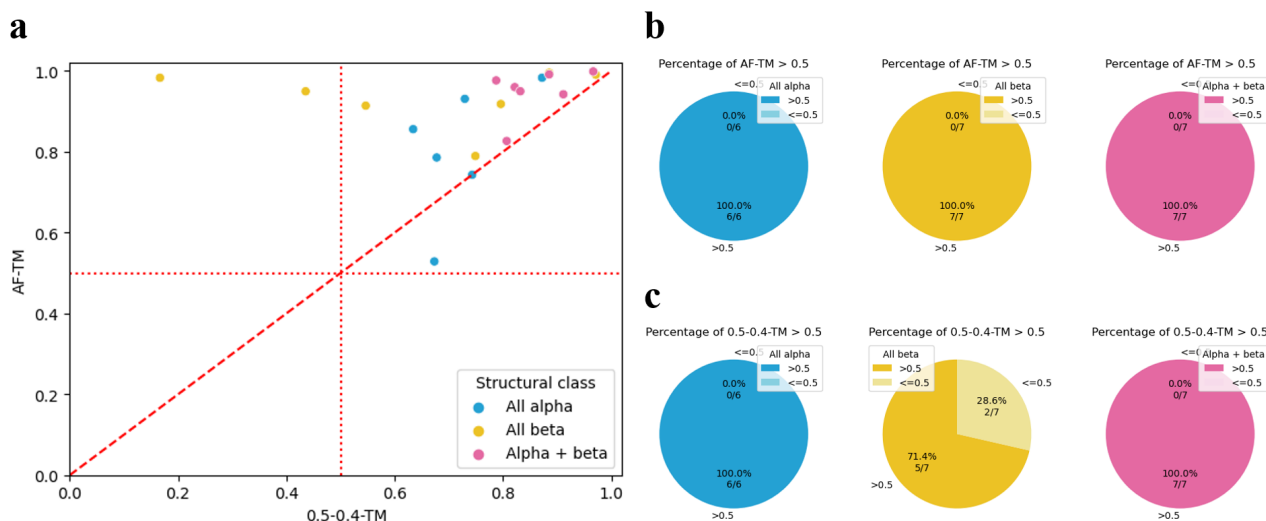


图 20 序列同源性位于 (0.5-0.4) 区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 50%-40% 区间的不同 structural class 蛋白的 TM-score 结果为横坐标, 以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图; 图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况; 图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

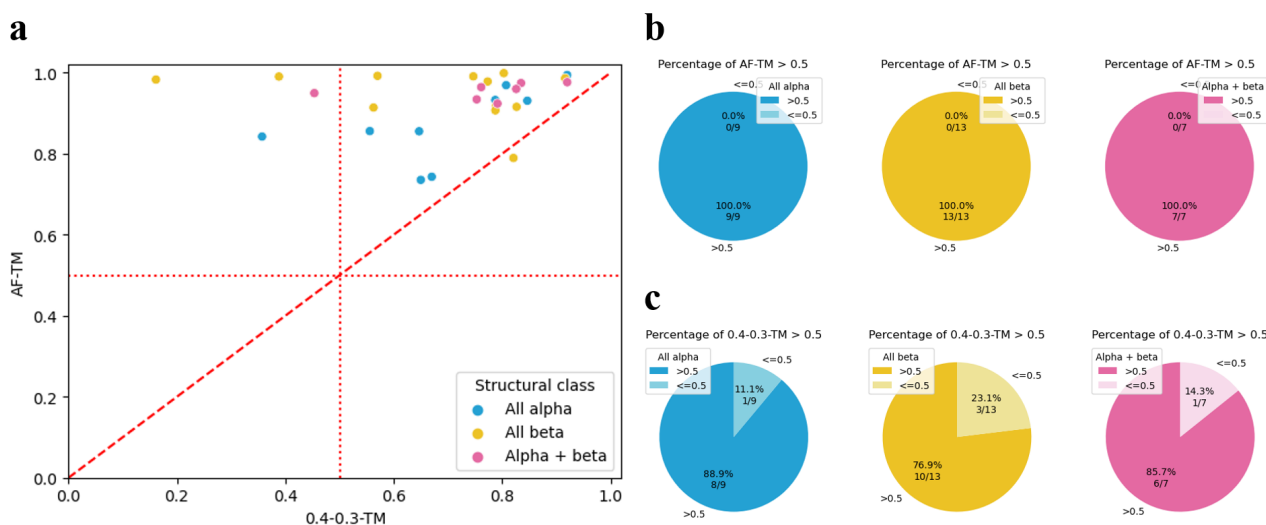


图 21 序列同源性位于 (0.4-0.3) 区间的蛋白结构预测准确度比较。图 a 展示了以 SWISS-MODEL 预测序列同源性 (Sequence Identity) 位于 40%-30% 区间的不同 structural class 蛋白的 TM-score 结果为横坐标, 以 AlphaFold2 预测同样样本所得 TM-score 为纵坐标呈现的散点图; 图 b 统计了该数据集中 AlphaFold2 模拟蛋白 TM-score 具体情况; 图 c 统计了同一数据集中 SWISS-MODEL 模拟蛋白 TM-score 具体情况。

图 20 中, 当序列同源性 (Sequence Identity) 大于 40% 且小于 50% 的时候, SWISS-MODEL

在 All beta proteins（全 β 蛋白）方面表现稍逊于 AlphaFold2。

观察图 21 时，我们发现序列同源性（Sequence Identity）大于 30%且小于 40%的时候，SWISS-MODEL 在全方面表现都稍逊于 AlphaFold2。

3.2.2 AlphaFold2 倾向性探究

AlphaFold2 在不同结构分类（Structural Class）这一赛道上表现太过卓越，以至于 SWISS-MODEL 与其几乎没有可比性。因此，我们需要重新审视 AlphaFold2 的内部表现，并进一步探究其在 All alpha proteins（全 α 蛋白）、All beta proteins（全 β 蛋白）、Alpha and beta proteins (a+b)（ α 和 β 皆有且彼此分离的蛋白）这三个相互独立的类别中是否存在预测准确度的倾向性。

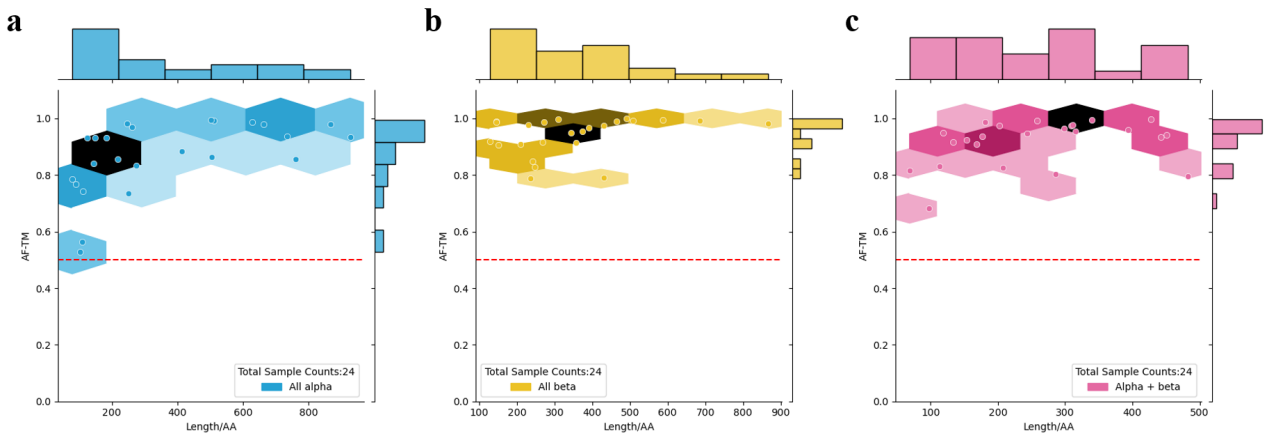


图 22 不同结构分类蛋白的 AlphaFold2 预测倾向性表现图。这三幅子图都以 AlphaFold2 获得的 TM-score 作为纵轴，以蛋白长度为横轴，绘制了蜂窝散点图。同时，在各自的轴向呈现了分布直方图。在蜂窝图上，不同深浅的颜色代表着样本量的大小，颜色越深的区域预示着此处分布的预测模拟蛋白数量更多。

具体来讲，在图 22 中，图 22c 颜色最深的点最靠近 $TM=1.0$ ，其次是图 22b，最后是图 22a。由此我们可以得出以下的结论：在不同的结构分类赛道上，AlphaFold2 的预测准确度存在一定的倾向性，即 Alpha and beta proteins (a+b)（ α 和 β 皆有且彼此分离的蛋白）> All beta proteins（全 β 蛋白）> All alpha proteins（全 α 蛋白）。这一结果表明，AlphaFold2 在预测 Alpha and beta proteins (a+b)（ α 和 β 皆有且彼此分离的蛋白）时表现最佳，而在预测 All alpha proteins（全 α 蛋白）时相对表现不太准确。这一简单的分析，能够为我们更好地理解 AlphaFold2 在不同结构分类中的预测准确性提供参考和启示。

3.3 不同物种来源蛋白表现

在之前的研究中，AlphaFold2 的出色表现一度让我们感到惊叹和赞叹，由此激发出我们对其更深入探究的兴趣。因此，我们引入了新的任务赛道，探索 AlphaFold2 在不同物种来源的蛋白中的表现情况。为了实现这一目标，我们对预备数据集（表 1）中的所有蛋白添加了物种（SPECIES）标签，并将其划分为 Eukaryotic（真核）、Prokaryotic（原核）和 Virus（病毒）三类蛋白。接着，我们通过获得蛋白的 AlphaFold2 预测结果 PDB 文件，在线上 TM-align^[47]平台（<https://zhanggroup.org/TM-align/>）与其在 RCSB PDB 数据库^[42]获得的原结构 PDB 文件进行 TM-score 计算后，将结果（AF-TM）汇入不同物种来源蛋白结构预测情况表（表 4）。通过这样的研究，我们希望能够进一步探讨 AlphaFold2 在不同物种来源蛋白的预测中是否存在特定的表现趋势，以及其在生物学实践中的应用潜力。

表 4 不同物种来源蛋白结构预测情况表⁴

PDB ID	Length/AA	SPECIES	AF-TM	PDB ID	Length/AA	SPECIES	AF-TM
6EJN	759	Eukaryotic	0.8558	1N5H	113	Eukaryotic	0.83228
2W83	628	Eukaryotic	0.98741	4AEQ	98	Prokaryotic	0.6829
2H7H	124	Virus	0.93075	1ROA	119	Eukaryotic	0.94954
1LJ2	275	Virus	0.8344	1ZO2	258	Eukaryotic	0.99128
1UIX	145	Eukaryotic	0.84228	2CC3	299	Prokaryotic	0.9643
1S1C	511	Eukaryotic	0.99185	2GEX	286	Prokaryotic	0.80321
1JAD	505	Eukaryotic	0.86451	2E74	979	Prokaryotic	0.2022
2LW9	110	Eukaryotic	0.56277	2ZT9	979	Prokaryotic	0.12576
3TER	250	Eukaryotic	0.73535	1Q90	972	Eukaryotic	0.96902
2VSG	733	Eukaryotic	0.93685	1RH5	512	Prokaryotic	0.95503
2YMY	89	Eukaryotic	0.76836	1RKL	43	Eukaryotic	0.46885
2V0X	412	Eukaryotic	0.88421	1XME	773	Prokaryotic	0.99069
4ABN	867	Eukaryotic	0.97883	1C99	87	Prokaryotic	0.56505
2VYI	261	Eukaryotic	0.96912	1H2S	296	Prokaryotic	0.97727
5FSH	926	Prokaryotic	0.93409	2MPN	144	Prokaryotic	0.50125

⁴ AF-TM 指用 AlphaFold2 预测后的蛋白模型与原结构进行相似性评估所得 TM-score。

PDB ID	Length/AA	SPECIES	AF-TM	PDB ID	Length/AA	SPECIES	AF-TM
3QXL	502	Eukaryotic	0.99428	5AWW	573	Prokaryotic	0.89459
1UG3	662	Eukaryotic	0.9803	2H3O	51	Prokaryotic	0.54927
1Z3X	246	Prokaryotic	0.98308	6CXH	741	Prokaryotic	0.95291
2Q1K	217	Prokaryotic	0.85529	2POR	312	Prokaryotic	0.97902
4DNN	104	Eukaryotic	0.5286	1FEP	696	Prokaryotic	0.98116
2YMJ	112	Eukaryotic	0.74298	2HDF	599	Prokaryotic	0.88101
5FMN	184	Prokaryotic	0.93258	2ERV	307	Prokaryotic	0.99778
6RZY	148	Eukaryotic	0.93078	5G3A	181	Prokaryotic	0.95522
2LO1	79	Eukaryotic	0.7857	2X27	218	Prokaryotic	0.98205
1OGM	587	Eukaryotic	0.99533	1COL	406	Prokaryotic	0.99594
1CZF	686	Eukaryotic	0.99257	1DDT	537	Virus	0.71744
5C1C	310	Eukaryotic	0.9985	1DLC	599	Prokaryotic	0.99128
3VN3	465	Eukaryotic	0.99043	1I5P	649	Prokaryotic	0.95586
6BG8	491	Prokaryotic	0.99883	1CIY	592	Prokaryotic	0.98092
5UYT	866	Prokaryotic	0.98296	1IKQ	614	Prokaryotic	0.83138
3DB5	268	Eukaryotic	0.91566	1FE6	214	Prokaryotic	0.91405
3DAL	356	Eukaryotic	0.91392	1USE	47	Eukaryotic	0.89737
3EP0	236	Eukaryotic	0.78939	4JPN	763	Virus	0.77402
3ECY	248	Eukaryotic	0.82786	3V6I	589	Prokaryotic	0.90486
6MAI	145	Prokaryotic	0.99089	1QEY	132	Virus	0.67491
1DUN	128	Virus	0.91806	2H8N	282	Eukaryotic	0.91555
2OL5	374	Prokaryotic	0.95376	1XQ8	148	Eukaryotic	0.3375
5BNC	508	Prokaryotic	0.99216	3QC7	173	Virus	0.50504
4YBN	431	Prokaryotic	0.97472	3S12	489	Virus	0.90525
4I86	209	Prokaryotic	0.90802	3EYJ	505	Virus	0.94173
3KYF	241	Prokaryotic	0.84903	4JPR	84	Virus	0.86768
5VX6	430	Prokaryotic	0.78967	2EBO	232	Virus	0.95879
1U9K	230	Eukaryotic	0.97822	1HFE	988	Prokaryotic	0.33495

PDB ID	Length/AA	SPECIES	AF-TM	PDB ID	Length/AA	SPECIES	AF-TM
2P39	151	Eukaryotic	0.90693	6RMV	424	Eukaryotic	0.98917
3HAL	272	Eukaryotic	0.98653	1DPJ	370	Eukaryotic	0.97422
1Q1U	146	Eukaryotic	0.98613	1PBY	924	Prokaryotic	0.99067
1L4I	390	Prokaryotic	0.96626	2FO1	817	Eukaryotic	0.25212
2CO6	344	Prokaryotic	0.94948	1WMI	309	Prokaryotic	0.9233
4W8Z	340	Prokaryotic	0.99373	1ZOQ	489	Eukaryotic	0.93616
4W8V	482	Prokaryotic	0.79617	2C52	125	Eukaryotic	0.69919
4F3M	442	Prokaryotic	0.9341	1KBH	114	Eukaryotic	0.56597
2CG4	309	Prokaryotic	0.97439	1WVK	94	Eukaryotic	0.29882
2IIZ	311	Prokaryotic	0.97639	1IK9	443	Eukaryotic	0.81141
2IFX	208	Prokaryotic	0.82656	4PC0	865	Eukaryotic	0.97288
3TOE	181	Prokaryotic	0.98627	5V02	243	Eukaryotic	0.68416
2YVW	428	Prokaryotic	0.99859	2VOF	349	Eukaryotic	0.9324
3R38	451	Prokaryotic	0.94215	2WH6	188	Eukaryotic	0.9432
1MSZ	69	Eukaryotic	0.81695	3M3N	780	Eukaryotic	0.96793
1JO0	202	Prokaryotic	0.9749	2A41	690	Eukaryotic	0.97265
3ZIG	177	Prokaryotic	0.93796	4RS9	199	Eukaryotic	0.1519
1MP9	394	Prokaryotic	0.95981	2FO1	817	Eukaryotic	0.25212
1E09	168	Eukaryotic	0.91003	2P2T	104	Eukaryotic	0.97907
1X53	154	Eukaryotic	0.92372	4DXR	231	Eukaryotic	0.95655
2NS9	315	Prokaryotic	0.95499	3GS2	286	Eukaryotic	0.8538
5I8J	134	Prokaryotic	0.91588	4E05	323	Eukaryotic	0.40408
2R5X	243	Prokaryotic	0.94599	2KXC	92	Eukaryotic	0.83362

从图 23 中我们观察到，图 23b 颜色最深的蜂窝区域明显比图 23a 中的更接近 $TM=1.0$ ，除此之外，图 23b 中 $TM<0.5$ 的区域也比图 23a 的少。这足以说明，AlphaFold2 在预测不同物种来源时具备一定倾向性，即 Prokaryotic proteins（原核来源蛋白）>Eukaryotic proteins（真核来源蛋白），换句话说，AlphaFold2 在预测 Prokaryotic proteins 时相较 Eukaryotic proteins 更准确。需要注意的是，对于 Virus proteins（病毒来源蛋白）的预测表现，由于样本量较少（仅

11 个), 我们无法进行充分的评估和分析。这些发现和结论对于未来 AlphaFold2 性能的进一步完善, 以及针对特定物种或蛋白进行定制化的蛋白质结构预测工作具有参考和借鉴意义。在未来的研究中, 我们或许可以结合不同物种来源的蛋白的预测表现特点, 继续优化 AlphaFold2 等蛋白质结构预测工具, 提高其在特定物种来源蛋白的预测准确性和实用性。

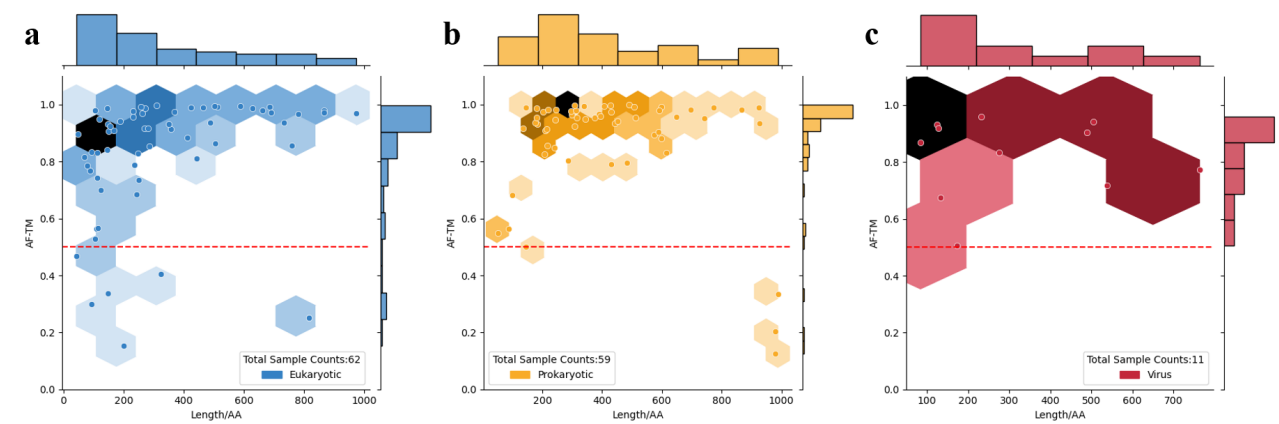


图 23 不同物种来源蛋白的 AlphaFold2 预测倾向性表现图。这三幅子图都以 AlphaFold2 获得的 TM-score 作为纵轴, 以蛋白长度为横轴, 绘制了蜂窝散点图。同时, 在各自的轴向呈现了分布直方图。在蜂窝图上, 不同深浅的颜色代表着样本量的大小, 颜色越深的区域预示着此处分布的预测模拟蛋白数量更多。

3.4 结果概览

为了更好地呈现本次毕设研究的结论, 我们对所有数据进行了综合评估, 并将整个 AlphaFold2 与传统同源建模 SWISS-MODEL 的蛋白结构预测准确性对比研究情况整合成一张蛋白结构预测工具选择偏好表 (表 5), 其中有详细涉及这两个工具的优劣势和适用范围:

表 5 蛋白结构预测工具选择偏好表⁵

蛋白结构预测工具选择偏好	AlphaFold2	SWISS-MODEL
Globular proteins	√	
Membrane proteins	√	
Fibrous proteins	√	
Non-globular/ Intrinsically unstructured proteins	√*仅限 SI 值 ∈ (0.9,0.7)	√*仅限 SI 值 ∈ (1.0,0.9) ∪ (0.7,0.3)
All Alpha proteins	√	

⁵ SI 值 ∈ (0.9,0.7)意味着序列同源性 (Sequence Identity) 在 90%-70%之间, 其他同理。

All Beta proteins	√
Alpha and beta proteins (a+b)	√

除此之外，我们还针对 AlphaFold2 在特定种类蛋白质的预测表现偏好性（表 6）进行了汇总，具体如下：

表 6 AlphaFold2 蛋白种类预测准确度偏好表

AlphaFold2 预测倾向	1st	2nd	3rd	4th
By protein type	Globular proteins	Fibrous proteins	Membrane proteins	Non-globular/Intrinsically unstructured proteins
By structural class	Alpha and beta proteins (a+b)	All beta proteins	All alpha proteins	
By species	Prokaryotic proteins	Eukaryotic proteins		

4 讨论

随着科技的不断进步，越来越多的蛋白质结构预测工具涌现，其中 AlphaFold2 和 SWISS-MODEL 已经成为著名的代表。这些工具的预测效果存在差异，因此如何选择适合的工具成为一个需要探索的问题。对于不同的蛋白质序列和研究目的，不同的预测工具可能会有不同的优劣，例如有些工具适合快速预测主要由 α -螺旋构成的蛋白质结构，或者其他更先进的工具则需要更多的时间和资源，但能够提供更为准确的结果。对于那些缺乏模板的未知蛋白质，人们需要根据预测工具在类似情况下的综合表现来判断预测结果的可信度和可靠性。鉴于 AlphaFold2 在最近的竞赛中的出色表现，AlphaFold2 的预测结果准确度相比同源建模工具有了明显提升，它的应用潜力备受关注，这使得 AlphaFold2 成为研究人员在特定情况下的首选工具，例如在预测缺乏模板的蛋白质结构时。然而，尽管 AlphaFold2 性能明显优于 SWISS-MODEL 等同源建模工具，但仍存在一定的局限性和误差。因此，为评估和选择不同预测工具的优劣，我们需要综合考虑它们的特点和机制，构建出一套完善的评估系统，为蛋白质结构预测提供更准确和可靠的技术支持。

本次毕业设计的研究聚焦于综合评估 AlphaFold2 和 SWISS-MODEL 的表现能力,考察了蛋白质种类、结构特征和物种来源多方面因素对两种预测工具的性能影响。据此我们构建出了一套比较完整的评估体系(表 5)。通过这个系统,可以帮助研究者在处理未知的蛋白质结构时更有针对性地选择合适的预测工具,从而提高实验数据的可信度和有效性。除此之外,我们还评估了 AlphaFold2 的工作潜力,深入挖掘了其预测质量的优势和局限(表 6)。这些研究结果对于进一步提升蛋白质结构预测准确性、优化预测工具以及促进蛋白质相关研究具有重要意义。通过细致的研究评估和深入探究工具特性和表现,我们可以更好地应用这些工具来解决科研中的实际问题。

需要特别指出的是,虽然本项研究未记录两个工具预测蛋白所花费的时间,但实际经验表明,对于越长的蛋白质序列,AlphaFold2 所需时间也就越多,比如对于长度为 500 个氨基酸的蛋白质而言,SWISS-MODEL 可能只需 5 分钟就能完成预测,但 AlphaFold2 则要花费 30 分钟甚至更长时间。因此,如果需要处理成百上千个长度大于 500 氨基酸的蛋白质数据集,其所需的预测时间将成倍增加。所以,在赶时间或进度的情况下,我们建议谨慎使用 AlphaFold2,换用同源建模的方法时可以引用本研究的成果,以提升预测结果的说服力。

另外有一点值得说明的是,本研究中两个地方的处理方法可能存在一些争议:其一是在预备数据集中存在不确定氨基酸项('?')的部分蛋白序列,'?'的存在会影响到两个工具的运行。为解决这一问题,我们移除了这些不确定氨基酸项,并使用剩余的完整序列进行预测,这样做的原因是不确定氨基酸项在整个序列中所占比重很小,微乎其微,我们认为其不会对蛋白整体构型造成过大影响;其二是使用了两个 AlphaFold2 运行,一个是部署在本地实验的开源软件,一个是部署在 Google Colab 的开源社区 ColabFold^[46](<https://github.com/sokrypton/ColabFold>)。这一决策是基于我们追求研究进度的产生的。两个软件都同样基于 DeepMind 开源 AlphaFold2 的基本架构,并使用相似的算法和数据处理过程来完成蛋白质结构预测,因此推测不会对蛋白结构预测造成影响。

尽管本次毕业设计中存在一点不足之处,但我们仍然希望这项成果能够对未来的相关研究提供足够的支持和参考。我们相信,通过逐步完善和更新,这项研究的结果和结论将会更加准确和可靠。我们也希望,这项研究的成果能够成为一个有用的指导,帮助更多的研究者更加准确地预测和分析蛋白质结构,从而推动蛋白质相关领域的研究和应用。

参考文献

- [1] KRYSHTAFOVYCH A, SCHWEDE T, TOPF M, et al. Critical assessment of methods of protein structure prediction (CASP)-Round XIV [J]. *Proteins*, 2021, 89(12): 1607-17.
- [2] MARTI-RENO M A, MADHUSUDHAN M S, FISER A, et al. Reliability of Assessment of Protein Structure Prediction Methods [J]. *Structure*, 2002, 10(3): 435-40.
- [3] ALTSCHUL S F, MADDEN T L, SCHÄFFER A A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs [J]. *Nucleic Acids Research*, 1997, 25(17): 3389-402.
- [4] KARPLUS K, KARCHIN R, BARRETT C, et al. What is the value added by human intervention in protein structure prediction? [J]. *Proteins*, 2001, Suppl 5(86-91).
- [5] JONES D T, TAYLOR W R, THORNTON J M. A new approach to protein fold recognition [J]. *Nature*, 1992, 358(6381): 86-9.
- [6] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold [J]. *Nature*, 2021, 596(7873): 583-9.
- [7] DILL K A, MACCALLUM J L. The protein-folding problem, 50 years on [J]. *Science*, 2012, 338(6110): 1042-6.
- [8] SINGH A, KAUSHIK R, JAYARAM B. Quality Assessment of Protein Tertiary Structures: Past, Present, and Future [M]//SHANKER A. *Bioinformatics: Sequences, Structures, Phylogeny*. Singapore; Springer Singapore. 2018: 271-88.
- [9] ALQURAISHI M. AlphaFold at CASP13 [J]. *Bioinformatics*, 2019, 35(22): 4862-5.
- [10] WHEELER T J, EDDY S R. nhmmer: DNA homology search with profile HMMs [J]. *Bioinformatics*, 2013, 29(19): 2487-9.
- [11] THE UNIPROT C. UniProt: the Universal Protein Knowledgebase in 2023 [J]. *Nucleic Acids Research*, 2023, 51(D1): D523-D31.
- [12] RICHARDSON L, ALLEN B, BALDI G, et al. MGnify: the microbiome sequence data analysis resource in 2023 [J]. *Nucleic Acids Research*, 2023, 51(D1): D753-D9.
- [13] STEINEGGER M, MEIER M, MIRDITA M, et al. HH-suite3 for fast remote homology detection and deep protein annotation [J]. *BMC Bioinformatics*, 2019, 20(1): 473.
- [14] SENIOR A W, EVANS R, JUMPER J, et al. Improved protein structure prediction using

- potentials from deep learning [J]. *Nature*, 2020, 577(7792): 706-10.
- [15] VARADI M, ANYANGO S, DESHPANDE M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models [J]. *Nucleic Acids Research*, 2022, 50(D1): D439-D44.
- [16] CALLAWAY E. 'The entire protein universe': AI predicts shape of nearly every known protein [J]. *Nature*, 2022, 608(7921): 15-6.
- [17] BAGDONAS H, FOGARTY C A, FADDA E, et al. The case for post-predictional modifications in the AlphaFold Protein Structure Database [J]. *Nature Structural & Molecular Biology*, 2021, 28(11): 869-70.
- [18] AN H J, FROEHLICH J W, LEBRILLA C B. Determination of glycosylation sites and site-specific heterogeneity in glycoproteins [J]. *Curr Opin Chem Biol*, 2009, 13(4): 421-6.
- [19] HE X-H, YOU C-Z, JIANG H-L, et al. AlphaFold2 versus experimental structures: evaluation on G protein-coupled receptors [J]. *Acta Pharmacologica Sinica*, 2023, 44(1): 1-7.
- [20] WATERHOUSE A, BERTONI M, BIENERT S, et al. SWISS-MODEL: homology modelling of protein structures and complexes [J]. *Nucleic Acids Research*, 2018, 46(W1): W296-W303.
- [21] BIENERT S, WATERHOUSE A, DE BEER T A, et al. The SWISS-MODEL Repository-new features and functionality [J]. *Nucleic Acids Res*, 2017, 45(D1): D313-D9.
- [22] CAMACHO C, COULOURIS G, AVAGYAN V, et al. BLAST+: architecture and applications [J]. *BMC Bioinformatics*, 2009, 10(421).
- [23] REMMERT M, BIEGERT A, HAUSER A, et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment [J]. *Nat Methods*, 2011, 9(2): 173-5.
- [24] BIASINI M, BIENERT S, WATERHOUSE A, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information [J]. *Nucleic Acids Res*, 2014, 42(Web Server issue): W252-8.
- [25] BERTONI M, KIEFER F, BIASINI M, et al. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology [J]. *Sci Rep*, 2017, 7(1): 10480.
- [26] BIASINI M, SCHMIDT T, BIENERT S, et al. OpenStructure: an integrated software framework for computational structural biology [J]. *Acta Crystallogr D Biol Crystallogr*, 2013, 69(Pt 5): 701-9.
- [27] BENKERT P, BIASINI M, SCHWEDE T. Toward the estimation of the absolute quality of

- individual protein structure models [J]. *Bioinformatics*, 2011, 27(3): 343-50.
- [28] GUEx N, PEITSCH M C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling [J]. *Electrophoresis*, 1997, 18(15): 2714-23.
- [29] SHAPOVALOV M V, DUNBRACK R L, JR. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions [J]. *Structure*, 2011, 19(6): 844-58.
- [30] MIYANO S, MESIROV J, KASIF S, et al. Research in Computational Molecular Biology: 9th Annual International Conference, RECOMB 2005, Cambridge, MA, USA, May 14-18, 2005, Proceedings (Lecture ... Science / Lecture Notes in Bioinformatics) [M]. Springer-Verlag, 2005.
- [31] KRIVOV G G, SHAPOVALOV M V, DUNBRACK R L, JR. Improved prediction of protein side-chain conformations with SCWRL4 [J]. *Proteins*, 2009, 77(4): 778-95.
- [32] EASTMAN P, SWAILS J, CHODERA J D, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics [J]. *PLoS Comput Biol*, 2017, 13(7): e1005659.
- [33] MACKERELL A D, JR., FEIG M, BROOKS C L, 3RD. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations [J]. *J Comput Chem*, 2004, 25(11): 1400-15.
- [34] BORDOLI L, KIEFER F, ARNOLD K, et al. Protein structure homology modeling using SWISS-MODEL workspace [J]. *Nature Protocols*, 2009, 4(1): 1-13.
- [35] KIRCHMAIR J, MARKT P, DISTINTO S, et al. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes? [J]. *Journal of Computer-Aided Molecular Design*, 2008, 22(3): 213-28.
- [36] REVA B A, FINKELSTEIN A V, SKOLNICK J. What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å? [J]. *Folding and Design*, 1998, 3(2): 141-7.
- [37] XU J, ZHANG Y. How significant is a protein structure similarity with TM-score = 0.5? [J]. *Bioinformatics*, 2010, 26(7): 889-95.
- [38] ZHANG Y, SKOLNICK J. Scoring function for automated assessment of protein structure template quality [J]. *Proteins*, 2004, 57(4): 702-10.
- [39] ZEMLA A. LGA: A method for finding 3D similarities in protein structures [J]. *Nucleic Acids*

- Res, 2003, 31(13): 3370-4.
- [40] ANDREEVA A, HOWORTH D, CHOTHIA C, et al. SCOP2 prototype: a new approach to protein structure mining [J]. Nucleic Acids Res, 2014, 42(Database issue): D310-4.
- [41] SILLITOE I, BORDIN N, DAWSON N, et al. CATH: increased structural coverage of functional space [J]. Nucleic Acids Res, 2021, 49(D1): D266-D73.
- [42] ROSE P W, BI C, BLUHM W F, et al. The RCSB Protein Data Bank: new resources for research and education [J]. Nucleic Acids Research, 2013, 41(D1): D475-D82.
- [43] SCHRODINGER, LLC. The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version 1.8 [M]. 2015.
- [44] SCHRODINGER, LLC. The JyMOL Molecular Graphics Development Component, Version 1.8 [M]. 2015.
- [45] SCHRODINGER, LLC. The PyMOL Molecular Graphics System, Version 1.8 [M]. 2015.
- [46] MILOT M, KONSTANTIN S, YOSHITAKA M, et al. ColabFold - Making protein folding accessible to all [J]. bioRxiv, 2022, 2021.08.15.456425.
- [47] ZHANG Y, SKOLNICK J. TM-align: a protein structure alignment algorithm based on the TM-score [J]. Nucleic Acids Res, 2005, 33(7): 2302-9.

致 谢

青山遮不住，毕竟东流去。时光只能匆匆流逝，而人生旅程却因为有您们默默的陪伴和支持，变得更加丰富多彩。在本科即将结束之际，我通过这篇致谢语，向您们表达我最衷心的感谢和谢意。

首先，我想向我的指导老师曹洋老师表达最深切的感谢。您开启了我学术研究的大门，引领我探索领域的奥秘，教导我选择合适的方法和标准，洞悉知识的本质。曹老师夜以继日的精神、认真负责的态度，深深影响着我，让我更加深刻地理解学科的基础和重点，并在思考和方法上为我提供了宝贵的指导和思路，使我在研究过程中不断迸发灵感和思考，更成为了我学术探索路上的良师益友。

其次，我也要衷心感谢学校领导、教师、同学们的关注和支持。正是在您们的培养和关怀下，我才能在这个学术领域中度过了四年难忘的时光。在此，我要感谢您们为我提供的学习和实践机会，您的激励和鼓励让我在充满挑战的道路上始终保持着激情和动力，在学习和成长过程中得到了有力的支撑和帮助。

我还想特别感谢我的同学们和室友，在你们的陪伴下，我度过了欢声笑语、患难与共的大学时光。感谢你们与我共享了每一次饕餮盛宴的浪费，你们的友情和信任，让我感受到了人性的温暖与美好，让我对未来充满了希望和信心。

最后，我还要感谢家人和亲友们对我的付出和支持，是你们的爱和关心，让我在这个学术领域中能够克服困难、奋发向上。尽管距离有时让我们不能真正地在一起，但你们的心意和祝福一直陪伴着我，让我牢记着最初的梦想和目标，让我更加坚定地前行。

此时我即将告别本科生活，回首曾经的岁月，心中充满感慨和不舍。然而，面对即将到来的新人生，我充满着期待和信心。再次感谢您们对我的支持和鼓励，让我在这条充满挑战但又充满希望的道路上不断前行。愿我们不忘初心，再续同行之缘！

声 明

本人声明所呈交的学位论文是本人在导师指导下进行的
研究工作及取得的研究成果。据我所知,除了文中特别加以标注和
致谢的地方外,论文中不包含其他人已经发表或撰写过的研究成
果,也不包含为获得四川大学或其他教育机构的学位或证书而使
用过的材料。与我一同工作的同志对本研究所做的任何贡献均已
在论文中作了明确的说明并表示谢意。

本学位论文成果是本人在四川大学读书期间在导师指导下
取得的,论文成果归四川大学所有,特此声明。

作者签名: 李金汕

导师签名: 唐洋

2023 年 5 月 26 日

学位论文使用授权书

本学位论文作者完全了解四川大学有关保留、使用学位论
文的规定,同意学校保留并向国家有关部门或相关机构送交论文
的原件、复印件和电子版,允许论文被查阅和借阅。本人授权川
大学将本学位论文的全部或部分内容编入有关数据库进行信息
技术服务,可以采用影印、缩印或扫描等复制手段保存、汇编学
位论文,并用于学术活动。

(涉密学位论文在解密后适用于本授权书)

作者签名: 李金汕

导师签名: 唐洋

2023 年 5 月 26 日