专题: 生物分子模拟中的机器学习

融合结构知识的蛋白质预训练模型进展*

汤天一[1)] 熊翊名[1)] 张睿格[1)] 张建[1)2)†]
李文飞[1)2)] 王骏[1)2)] 王炜[1)2)‡]

1) (南京大学物理学院, 南京 210093)
2) (南京大学脑科学研究院, 南京 210093)

(2024 年 6 月 7 日收到; 2024 年 7 月 12 日收到修改稿)

自然语言和图像处理领域引发的人工智能革命给蛋白质计算领域带来了新的思路和研究范式. 其中一个重大的进展是从海量蛋白质序列通过自监督学习得到预训练的蛋白质语言模型. 这类预训练模型编码了蛋白质的序列、进化、结构乃至功能等多种信息, 可方便地迁移至多种下游任务, 并展现了强大的泛化能力. 在此基础上, 人们正进一步发展融合更多种类数据的多模态预训练模型. 考虑到蛋白质结构是决定其功能的主要因素, 融合了结构信息的蛋白质预训练模型可更好地支持下游多种任务, 本文对这一方向的研究工作进行了介绍和总结. 此外, 还简介了融合先验知识的蛋白质预训练模型、RNA 语言模型、蛋白质设计等方面的工作, 讨论了这些领域目前的现状、困难及可能的解决方案.

关键词: 蛋白质基础模型, 蛋白质多模态模型, 蛋白质结构, 机器学习

PACS: 87.10.Vg, 87.16.A–, 87.14.E–, 87.15.A– **DOI:** 10.7498/aps.73.20240811

1 引 言

随着 2019 年 AlphaFold 以及后来的 AlphaFold2 在蛋白质结构预测领域取得巨大成功[1,2], 深度学习在各个科学研究领域攻城略地, 颠覆了诸多传统的研究方法, 催生出一批令人兴奋的成果. 2023 年初, OpenAI 公司推出了 ChatGPT[3–6], 更是在全球范围掀起了一股人工智能热潮. ChatGPT 背后的技术支持来自于语言大模型, 它通过海量的数据训练极大规模的模型. 事实上, 在 ChatGPT 之前, 科学界和工业界已经开始重点关注语言大模型, 包括 Google 的 Bert 和 T5、DeepMind 的 Gogher、阿里的八卦炉、清华的 GLM、华为的盘古、百度的文心、浪潮的源 1.0 等[7–10], 不一一列举. 其中阿里的八卦炉是第一个参数量达到了 10^{14} 规模的模型[8], 这和人脑中的突触数量处于同一数量级. 在此类大模型的支持下, 人们可能不再需要为特定任务搭建特定的数据集和模型, 如翻译、情感分析、阅读理解等, 而是直接训练一个超大的通用模型, 其他任务只需要在此模型基础上微调即可. 这颠覆了传统的工作模式, 且为通用人工智能 (artificial general intelligence, AGI) 提供了一条可能的道路. 更重要的是, 随着规模的增大, 这类大模型会突然在某个方面展现出乎意料的智能, 类似物理复杂系统的“涌现”效应, 催生意外的能力[11]. 这也已经在 Chat GPT 中被观察到. 自 ChatGPT 推出以来, 大模型进化之路正飞速向多模态前进, 以融合从文本到图像、语音、视频等多种模态的更海量的数据, 典型模型如 Flamingo[12], GPT-4[13], PaLM-E[14], LLaMA[15],

* 科技部科技创新项目 (批准号: 2030-2021ZD0201300) 和国家自然科学基金 (批准号: 11934008) 资助的课题.
† 通信作者. E-mail: jzhang@nju.edu.cn
‡ 通信作者. E-mail: wangwei@nju.edu.cn

Gemini[16], X-LLM[17], VideoChat[18] 等. 这里, 多模态模型指一种能够处理来自不同模态 (如图像、语音、文本等) 的多种信息的机器学习模型. 多模态技术可以将这些不同形式的数据整合起来, 实现对数据更加全面和准确的分析与理解. 在生物计算领域, 多模态模型指在序列信息之外, 还将如结构、功能、动力学等其他模态信息融入模型.

深度学习在自然语言处理 (natural language processing, NLP) 大模型技术方向的突破给了其他领域的工作者极大启发. 在蛋白质计算领域, 上述技术被移植过来用于从海量蛋白质序列信息学习其内在数据分布. 通过设计合适的深度学习网络和进行相应的训练, 网络把输入数据映射到其对应的特征表示空间 (representation space), 或称潜在空间 (latent space), 或称嵌入 (embedding), 得到数据的表示 (representation) 或称编码 (encoding). 一般认为此表示编码了蛋白质的序列、结构、进化、乃至功能等信息, 可加速下游多种任务的开发. 这类从海量序列数据出发, 并借鉴 NLP 技术进行预训练得到的模型通常被称为蛋白质语言模型 (protein language model, PLM), 它是一种蛋白质基础模型 (protein foundation model) 或蛋白质预训练模型 (pre-trained model, PTM).

基于蛋白质基础模型或预训练模型的方案相对于传统建模方法有诸多显著优势. 首先, 预训练模型从海量数据进行学习, 能自动挖掘和捕捉其中的深层次特征, 从而更好地编码蛋白质的序列、进化、结构、功能等多种信息, 在预测蛋白质结构和功能方面常表现出更高的准确性. 其次, 预训练模型通常采用自监督学习方式 (self-supervised learning), 不依赖特定的标签 (labels) 或标注 (annotation) 数据, 使模型在数据稀疏或标签不足的情况下仍然能够进行有效的学习, 降低了学习成本, 加速了开发过程. 再次, 海量数据和大型算力赋予了预训练模型强大的泛化能力, 通常无需对每个下游任务进行额外训练. 只需用预训练模型的特征表示作为输入, 通过采用少量样本微调 (fine-tuning)、零样本学习 (zero-shot learning)、或提示学习 (prompt learning) 等方式即可迅速开发出适应下游任务的模型. 这同时也有利于解决特定下游任务标签数据稀少的问题.

2019 年以来, 各种蛋白质预训练模型如雨后春笋般发展起来. 知名的工作如 BB-model[19], Seq Vec[20], UniRep[21], ESM 系列[22–26], Progen[27,28], PM LM[29], ProtTrans[30], xTrimoPGLM[31], Evo[32] 等. 在这些预训练模型的加持下, 人们测试了大量的下游任务并展示了预训练模型的强大. 这些任务包括但不限于二级三级结构预测、折叠类型分类、蛋白质相互作用、蛋白-药物相互作用、配体亲和性预测、蛋白质功能预测、细胞内定位、突变功能预测、适应性预测等. 由于此类工作数量众多, 不一一列举, 详细的介绍可参考相关综述[33–40].

近三年以来, 几乎与自然语言处理领域齐头并进, 蛋白质预训练模型也由单纯从序列进行学习, 进化到同时学习序列、结构、功能、动力学信息等多种模态数据, 涌现出了一批多模态预训练模型. 如图 1 所示. 假如把蛋白质一级序列信息类比于人类语言的话, 那么三维结构就可类比为图片, 而三

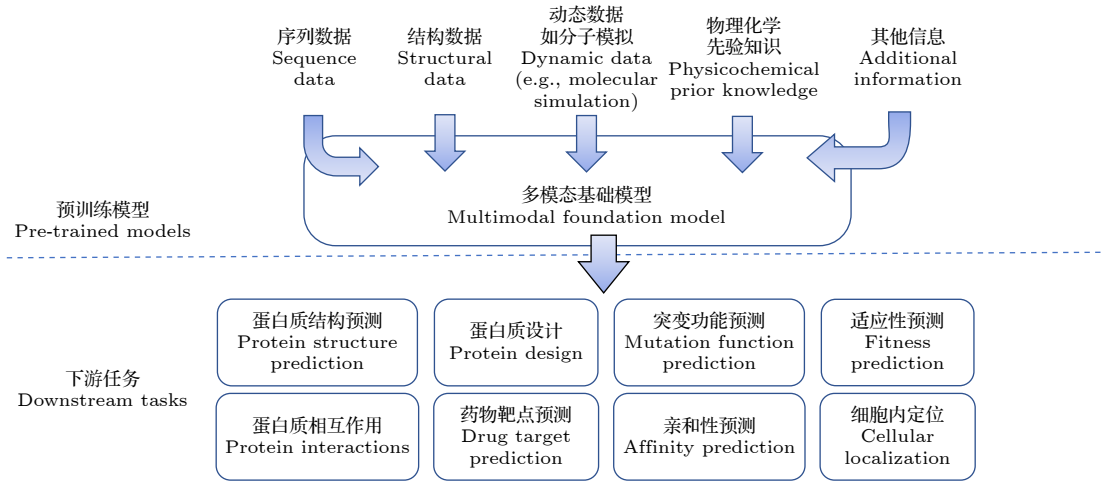

图 1 蛋白质多模态基础 (预训练) 模型及其应用 (只示意性列出若干下游任务)

Fig. 1. Protein multi-modal foundation (pre-trained) models and the downstream tasks.

维动态结构则可被类比为视频. 将更多模态的数据和知识融合在一个大模型内, 可显著地提高模型的智能, 这已在自然语言处理领域有清楚的展现[41].

除序列信息之外, 在蛋白质预训练模型中融合结构信息尤其重要, 这是因为存在序列类似而结构全然不同的蛋白质, 同时也存在序列相似度极低但空间结构相似的例子. 另外, 由于 AlphaFold 系列的革命性突破, 包括 AlphaFold 预测的结构在内, 可用蛋白质三级结构已到数亿量级, 也为训练对应模型提供了大量数据[2].

基于上述原因, 本文重点关注融合了结构信息的蛋白质多模态预训练模型. 此外, 如 AlphaFold2 中的 EvoFormer 模块等为特殊目的而优化的网络[2], 虽并非为通用目的而设计的预训练模型, 也是本文关注的对象. 这是考虑到此类模型具有高度优化的编码器, 性能优异, 且容易从网络中分离出优化的蛋白质特征表示, 对一些特定下游任务, 亦可作为预训练模型使用. 特别需要声明的是, 由于作者学识所限, 并且由于这个领域发展极为迅速, 可能会遗漏部分优秀的工作, 在此致歉.

2 融合了结构信息的通用蛋白质预训练模型

Bepler 和 Berger[19,42] 开创性地提出把蛋白质语言模型通过迁移学习用于下游各种任务的方案. 在他们的方案中, 在通过自监督学习从大量蛋白质序列中学习其语义表示的基础上, 进一步利用监督学习, 把蛋白质三维结构信息也进行编码, 获取同时编码了蛋白质序列和结构信息的表示, 用于支持下游任务. 具体来说, 他们采用多任务 (multi-task) 方式训练一个双向三层 LSTM 网络, 任务包括基于自监督学习的掩码语言建模 (masked language modeling, MLM) 和基于监督学习的残基间接触预测与结构相似性预测. 其中结构相似性根据 SCOP 数据库中的分类结果定义. 所使用的训练集包括来自 UniRef 数据库的 76M 条蛋白质序列和来自 SCOP 数据库的 28K 个蛋白质结构. 网络把输入蛋白质序列映射到一个低维语义空间, 得到一个和输入序列同长度的表示 (MT-LSTM), 并可通过如池化操作 (pooling) 得到对整个蛋白质的表示. 上述表示编码了蛋白质进化、结构和功能的信息, 可用于多种下游任务. 他们首先测试了基于此表示的模型在区分蛋白质类别 (class)、折叠类型 (fold)、家族 (family) 方面的能力, 这可以通过简单地比较蛋白质在表示空间的矢量得到. 结果表明, 基于多任务 MT-LSTM 训练得到的表示模型优于只基于序列训练得到的表示模型 (DLM-LSTM), 也优于传统的序列比对或结构比对模型. 在预测蛋白质跨膜区域的任务上, 与其他模型相比, MT-LSTM 同样具有更优的表现. 通过结合 MT-LSTM 表示和高斯过程回归方法, 他们还在预测序列突变表现型的任务上取得了领先的结果. 关于模型的简要信息和模型对比见表 1.

Guo 等[43] 发展了一个通过自监督方案直接从蛋白质三维结构进行学习的预训练模型. 这一方案没有使用大量序列数据. 训练所用结构数据来自 PDB 数据库, 经过处理后得到约 7 万个蛋白质三维结构. Guo 等在蛋白质 C^{α} 原子三维结构坐标上添加高斯噪声, 把扰动后的残基距离矩阵输入网络. 网络的训练目标是估计扰动后的距离矩阵的梯度. 通过这种自监督学习方式, 网络可以获取蛋白质结构的三个层级的表示: 残基层次、残基对层次和蛋白质层次. 通过对两个下游任务进行测试, 包括蛋白质结构质量评估和蛋白-蛋白互作用位点预测, 他们发现与不使用预训练模型和使用基于纯序列的预训练模型相比, 这一新方案具有明显的优势. 另外还指出, 虽然蛋白质结构数量显著小于序列数量, 但由于结构包含更多的信息, 基于结构进行的预训练模型很有效.

自监督学习的另一个常用方案是对比学习. Hermosilla 和 Ropinski [44] 发展了 New IEConv 模型用于从蛋白质三维结构中学习其表示. 具体地, 他们将蛋白质的三维结构转化为一张图, 从中随机采样两个子片段, 经编码器映射到表示空间后得到两个矢量, 然后计算两个矢量之间的余弦距离. 网络训练的目标是最小化来自同一个蛋白质的两个片段在表示空间的距离, 同时最大化来自不同蛋白质的片段在表示空间的距离. 网络的训练使用了来自 PDB 数据库的所有长度大于 25 残基的蛋白质结构. 经过实验, 他们发现训练中所采用的子片段的最优长度为蛋白质总长度的 40%—60%. 他们在多个下游任务测试了这一蛋白质结构表示的有效性, 包括基于 SCOP 分类的蛋白质结构相似性、折叠类型分类、蛋白质功能预测、酶催化反映类型预测、蛋白质-配体亲和性预测. 与不经预训练的模

表 1　多模态蛋白质预训练模型

Table 1.　Multimodal protein pre-trained models.

模型名	时间	模型	数据模态	预训练方法	训练集	参数量	算力要求	下游任务	文献
融合了结构信息的通用蛋白质预训练模型									
Bepler &Berger	2019	Bi-LSTM	Sequence, structure	MLM for sequences, supervised learning for 3D structures	76M sequences, 28K structures	—	1X 32G-V100, 13 to 51 days	Fold classification transmembrane region prediction	[19,42]
Guo model	2022	CNN	Structure	Self-supervised pre-training on noised pair-distance	73K structures	—	—	QA, PPI	[43]
New IEConv	2022	GCN	Sequence, structure	Contrastive learning between randomly sampled 3D substructures	476K chains	30M	—	protein function prediction, protein fold classification, structural similarity prediction, protein-ligand binding affinity prediction	[44]
GearNet	2023	ESM-1b, GearNet	Sequence, structure	PLM, contrastive learning	805K structures from AlphaFoldDB	—	4X A100	Fold classification, EC, GO	
STEPS	2023	BERT, GCN	Sequence, structure	PLM, supervised learning from 3D structures	40K structures	—	—	Membrane protein classification, cellular location prediction, EC	
UNI-MOL	2023	Transformer	Sequence, structure	Atom 3D position denoise, masked atom type prediction	209M molecule conformations, 3.2M protein pockets structure	—	8X 32G-V100, 3 days	molecular property prediction, molecular conformation generation, pocket property prediction, protein-ligand binding pose prediction	
SaProt	2023	BERT	Sequence, structure	Convert structures to structure-aware vocabulary, MLM	40M sequences and structures from PDB/AlphaFoldDB	650M	64X 80G-A100, 3 months	Thermostability, HumanPPI, Metal Ion Binding, EC, GO, DeepLoc, contact prediction	[51]
融合了结构信息的非通用蛋白质预训练模型									
Evoformer	2021	Evoformer	Sequence, structure	MLM, Supervised learning	BPD+Uniclust30, PDB	—	128TPU-v3, 11 days	Structure prediction	[2]
DeepFRI	2021	LSTM+GCN	Sequence, structure	PLM(pretrained, frozen), supervised learning for 3D structures	10M sequences for pre-training	—	—	GO, EC, PPI interaction sites	[47]
LM-GVP	2022	Transformer +GVP	Sequence, structure	PLM(changeable), supervised learning for 3D structures	—	—	8X 32G-V100	fluorescence, protease stability, GO, mutational effects	[48]
ProNet	2023	GCN	Sequence, structure	Supervised learning	—	—	—	Fold classification, reaction classification, binding affinity, PI	
HoloProt	2022	MPN	Sequence, structure surface	Supervised learning	—	1.8M	1X 1080Ti, 1 day	Ligand binding affinity, EC	[56]

表 1（续） 多模态蛋白质预训练模型

Table 1 (continued). Multimodal protein pre-trained models.

模型名	时间	模型	数据模态	预训练方法	训练集	参数量	算力要求	下游任务	文献
编码动态三维结构信息的预训练模型									
ProtMD	2022	E(3)-Equivariant Graph Matching Network	Sequence, structure trajectory	Self-supervised learning, atom-level prompt-based denoising generative task, conformation-level snapshot ordering task	62.8K snapshots from MD for 64 protein-ligand pairs	5.2M	4X V100	Binding affinity prediction, binary classification of ligand efficacy	[58]
融合了知识的蛋白质预训练模型									
OntoProtein	2022	ProtBert, Gu-model	Sequence, knowledge	MLM, contrastive learning	ProteinKG25 with 5M knowledge triples	—	V100	TAPE, PPI, Protein function prediction	[60]
KeAP	2023	ProtBert, Gu-model	Sequence, knowledge	MLM	ProteinKG25	—	—	TAPE, PPI, Protein function prediction	[62]
ProtST	2023	ProtBert, ESM-1b, ESM-2, PubMedBert	Sequence, knowledge	MLM, Multimodal Representation Alignment, Multimodal Mask Prediction	ProtDescribe with 553K sequence-property pairs	—	4X V100	Protein localization prediction, Fitness landscape prediction, Protein function annotation	[63]
RNA语言模型									
RNA-FM	2022.8	BERT	Sequence	MLM	RNAcentral, 23.7M ncRNA sequences	—	8X A100 80G, 1 month	SS prediction, 3D contact/distance map, 3D reconstruction, evolution study, RNA-protein interaction, MRL prediction	[78]
RNABert	2022	BERT	Sequence	MLM	RNAcentral (762K) & Rfam 14.3 dataset	—	V100	structural alignment, clustering	[86]
SpliceBERT	2023	BERT	Sequence	MLM	Pre-mRNA of 72 vertebrates, 2M sequences, 64B nucleotides	19.4M	8X V100, 1 week	multi-species splice site prediction, human branch point prediction	[79]
RNA-MSM	2023	MSA-transformer	Sequence	MLM	4069 RNA families from Rfam 14.7	—	8X V100 32G	SS prediction, solvent accessibility prediction	[83]
Uni-RNA	2023	BERT	Sequence	MLM	RNAcentral & nt & GWH (1billion sequences)	25—400M	128X A100	SS prediction, 3D structure prediction, MRL, Isoform percentage prediction on 3' UTR, splice site prediction, classification of ncRNA functional families, modification site prediction	[84]
RNAErnie	2024	ERNIE	Sequence, motif information	MLM at base/subsequence/motif level masking	RNAcentral, 23M ncRNA sequences	105M	4X V100 32G, 250 hours	sequence classification, RNA–RNA interaction, SS prediction	[85]

*PLM, protein language model; MLM, masked language model; GCN, graph convolutional network; GVP, geometric vector perceptrons; EC, enzyme commission number prediction; GO, gene ontology term prediction; PPI, protein-protein interaction; TAPE, the tasks assessing protein embeddings database; QA, quality assessment of structures; SS, secondary structure; MRL, mean ribosome load prediction in mRNA.

型、基于监督学习的模型和基于纯序列预训练的模型如 EMS-1b 等相比, 均具有更好的性能.

GearNet 借鉴 SimCLR 的多视角对比学习方案以编码蛋白质结构信息[45,46]. 模型把蛋白质结构转化为一张图, 从中随机抽取两个子图, 使用不同的加噪方案以获取不同视角 (view), 然后通过网络计算它们相应的表示. 网络优化的目标是根据两个子图是来自于同一蛋白或不同蛋白, 分别增加或减少两个视角在表示空间的相似度. 预训练使用来自 PDB 数据库和 Alphafold2 预测的约 805K 个蛋白质结构. 文献中在 4 个下游任务测试了 GearNet 表示的有效性, 包括酶 EC 编号预测、基因 Ontology(GO) 条目预测、折叠类型分类、酶催化反应类型预测. 通过和基于序列的预训练得到的蛋白质表示 (ProtTrans, ESM-1b)、基于序列和结构结合的表示 (DeepFRI[47], LM-GVP[48]), 以及基于结构的表示 (NewIEConv)[44] 进行对比实验, 发现 GearNet 在 8 个测试集的 7 个中给出了最好的结果. 另外, 考虑到 GearNet 用较少数量的蛋白质结构 (805K) 进行训练, 性能却优于基于大量序列预训练的编码器 (ESM-1b: 250M 序列, ProtTrans: 2.1B 序列), 证明结构比序列中蕴含了更多的信息, 能导致更好的表示. 另外, GearNet 还测试了使用 PDB 数据库和使用 Alphafold2 预测结构数据库进行预训练的差异, 结果显示不同的数据库选择对模型性能影响很小, 模型有很好的健壮性.

Chen 等[49] 发展了 STEPS 方法, 以融合从序列和结构得到的两个特征表示. 对于序列, 使用蛋白质语言模型得到其表示 h^{s}. 对于结构, 将其转化为一张图 G, 计算其隐含层表示 h_{G}. 为优化 h^{s} 和 h_{G} 之间的关系, Chen 等设计了两个自监督学习代理任务. 第一个为残基间距离预测, 第二个为残基二面角掩码预测 (对特定蛋白, 遮蔽其中 15% 的二面角信息). 注意在自监督学习过程中, 语言模型的输出 h^{s} 被冻结保持不变. 预训练使用了包含 Alphafold 预测结构在内的约 4 万蛋白质三维结构. 他们在三个下游任务对模型进行了微调和测试, 包括判定是否膜蛋白、蛋白质细胞内定位、酶催化反应分类. 与蛋白语言模型 (基于 BERT)、DeepFRI 等相比, STEPS 均具有较大优势. 另外, 消融实验证明, 蛋白质结构中的残基对距离信息、对获取更好的表示具有决定性的贡献.

UNI-MOL 是一个为蛋白质和小分子结合而设计的预训练模型[50]. Zhou 等[50] 收集了 209M 小分子构象以及 3.2M 个蛋白质结合口袋的三维模型, 在原子层面上, 设计了两个代理任务来对模型进行预训练. 第一个任务为给原子位置加入噪声, 然后训练网络预测其正确的位置. 第二个任务为遮蔽原子类型, 训练网络对其类型进行预测. 并在多个下游任务对预训练模型的性能进行了测试, 包括分子属性预测、分子构象生成任务、蛋白质结合口袋性质、配体结合构象预测. 发现 UNI-MOL在大部分任务中优于其他模型. 尤其是当下游任务只有很少的标签数据情况下, 如蛋白质结合口袋性质预测, 相比其他模型更是有显著的提高. 他们将其归因为预训练模型编码了蛋白质的三维结构信息.

Su 等[51,52] 提出了一个统一处理序列与结构信息的方案 SaProt, 其创新之处在于把蛋白质三级结构通过 Foldseek 工具编码成与原序列等长的含有结构信息的 token 序列. Foldseek 的输入是指定氨基酸临近区域的三维 (3D) 构象, 它通过一个离散化变分自编码器 (VQ-VAE) 网络, 把构象转化为 20 个离散矢量中的一个, 称为 3D token. 相比于通过图来表示蛋白质结构, 这一方案的优势在于把蛋白质序列和三维结构都转化为一个语句, 可无缝地使用 NLP 领域的各种大模型架构. SaProt 模型采用了 ESM[22,24] 的训练框架, 即掩码语言模型, 在一个包含 40M 蛋白质序列和结构的数据集上对网络进行预训练, 得到一个大规模的具有通用性的蛋白质表示 SaProt. 在 10 个下游任务的测试表明, 此表示方案具有优异的性能和广泛的适用性.

3 融合了结构信息的非通用蛋白质模型

如引言所述, EvoFormer 等为特殊目的而优化的网络模型, 虽非通用预训练模型, 亦在本文讨论之列. 另外, 由于此类模型众多, 只选其中的一部分予以介绍.

Evoformer 是著名的 AlphaFold2 的编码器部分, 即去掉后部生成模块之后余下的部分[2]. 它接受 MSA 和残基对信息作为输入、输出对应的表示. 由于 Evoformer 同时使用大量蛋白质序列和结构进行监督训练, 它输出的表示融合了序列和结构的信息. Hu 等[53] 在结构预测、功能预测、适应度 (fitness) 预测三类共 7 个任务上, 详细测试了 Evoformer 的表征能力, 并与 ESM-1b 和 MSA-Transformer 进

行了对比. 他们发现: 1) 经 AlphaFold2 训练的 Evoformer 参数是通用的, 可被用于各种结构和功能预测任务. 2) AlphaFold 在结构预测任务 (包括二级结构预测和接触图预测) 和小蛋白稳定性预测中, 比 ESM-1b 和 MSA-Transformer 具有更优的性能. 但在蛋白质功能预测任务上不如后两者, 在零样本适应度预测上表现不好. 3) Evoformer 对输入 MSA 信息的依赖很强, 另外, 如使用从 EMS-1b 转化来的 MSA 信息替代原 Evoformer 的输入, 几乎没有性能损失.

另外, 我们注意到在刚刚发布的 AlphaFold-3 中[54], Evoformer 被一个更简单的 Pairformer 代替, 它简化了对 MSA 信息处理的过程. Pairformer 只对单个氨基酸表示 (single representation) 和氨基酸对表示 (pair representation) 进行处理, MSA 表示不再传递给下游模块. 虽然 AlphaFold3 具有更强大的性能, 尤其是在复合体结构预测上, 但 Pairformer 模块本身对蛋白质信息的表征能力尚未被系统地测试.

DeepFRI 是一个两阶段蛋白质功能预测模型 [47]. 第一阶段在一个大小为 10M 的蛋白质序列数据集上训练一个基于 LSTM 的语言模型, 从中抽取残基分辨率的序列特征. 具体训练方法借鉴了 Bepler 和 Berger[19] 采用的掩码语言建模 (MLM) 方法. 在第二阶段, 上述序列特征和残基接触图以及用于表示三维结构的图网络向量一起, 被输入到下游的图卷积层, 得到一个融合了序列和结构信息的表示层, 再经两个全连接层后, 输出蛋白质的功能信息. 网络第二阶段利用具体任务对应的标签数据进行监督学习, 并冻结第一阶段获得的语言模型参数. DeepFRI 这一融合了序列和结构信息的模型具有良好的抗噪声特性, 即使在模型中以预测的蛋白质结构代替实验结构, 预测准确度也只有可忽略的下降.

LM-GVP 模型结合了蛋白质语言模型 (PLM) 和一个对三维空间平移和旋转具有不变性的网络模块 (geometric vector perceptrons, GVP), 在序列数据、结构数据以及若干下游数据集上进行训练, 用于预测蛋白质特性 [48]. 其中 PLM 模块基于 Transformer 架构并且是预训练的, 它的输出向量与由蛋白质结构转化来的图向量结合, 被输入给下游 GVP 模块. 与 DeepFRI 模型不同, LM-GVP 在使用下游数据集进行监督学习时, 允许梯度回传至 PLM 模块, 因此模型给出的特征表示融合了蛋白质的结构信息. 然而, 由于这些结构信息不是经由下游任务无关的方式融入, 因此 LM-GVP 给出的不是一个通用表示, 可能只在特定任务具有良好性能.

ProNet 在三个不同的层级学习蛋白质的三维结构, 包括氨基酸级 (C^{α} 原子)、主链级 (主链原子) 和全原子级 [55]. 这种分级方案的优势是: 1) 用不同层级的表示适配不同的下游任务, 如蛋白质功能预测只需要氨基酸层次的表示即可, 而亲和能预测可能需要原子级的表示. 2) 训练和推理的速度大大增加. Wang 等 [55] 在蛋白质折叠类型分类、酶反应类型分类、配体亲和能预测共三个任务上测试了这一模型, 结果显示比其他同类模型具有持平或略优的性能, 并且运算速度最高有 6 倍的提升.

HoloProt 从多个尺度对蛋白质结构进行表征, 包括序列、二级结构、三级和四级结构, 以及蛋白质表面形貌 [56]. 其中前四个层次被统称为结构信息, 并被转化为一张图. 图的节点为残基, 空间距离小于某个阈值的两个残基之间用一条边相连. 对于蛋白质表面形貌, 也在三角剖分后被转化为一张图. 与 MaSIF 模型类似 [57], 每个图节点的特征包括氨基酸标识、电荷、疏水性和局域曲率等信息, 如果两个节点同属于一个三角形, 则它们之间有一条边相连. 此外, 为了在两张图之间传递信息, HoloProt 模型还在分属于两张图的节点之间引入了边 (如果这两个节点属于同一个残基). HoloProt 模型在两个下游任务, 包括配体结合亲和性预测和酶催化反应分类任务上进行了训练和测试, 发现与之前的多个基于序列的模型和基于结构的模型相比, 多尺度的 HoloProt 模型具有更优的性能. 此外, 消融实验指出, 对于配体结合亲和性预测, 只基于蛋白质表面形貌的模型已经工作得很好. 对于酶催化反应分类任务, 只考虑蛋白质表面形貌会导致预测性能大幅下降, 因此对于这一任务, 结构信息非常重要.

4 编码动态三维结构信息的预训练模型

蛋白质结构的动态性对其生物功能至关重要, 尤其是可变构蛋白和天然无序蛋白. 在蛋白质相互作用和蛋白质-药物相互作用中, 结合口袋的构象

动力学对亲和性有重要影响. 然而大部分蛋白质表示模型仅从静态结构进行学习, 未考虑蛋白质的动态性. 可以预期, 如能在预训练模型中融入蛋白质的动态信息, 将有力地促进诸如蛋白-蛋白相互作用、蛋白-药物相互作用等下游任务的进行.

基于类似考虑, Wu 等[58] 发展了 ProtMD 方法, 从蛋白质-配体相互作用的动态结构中学习其特征表示. 他们首先对 64 个蛋白质-配体复合体进行分子动力学模拟, 得到共约 62.8 K 构象. 在自监督学习过程中, 第一个代理任务采用基于提示的去噪生成, 从 t 时刻加了噪声的蛋白质结构预测 $t+i$ 时刻的结构, 并与模拟的结果进行对比来计算损失函数. 这一任务被用于学习原子级别的、局域的时空相关信息. 第二个代理任务为构象重排序任务, 即把模拟中的若干构象顺序打乱, 迫使网络学习其正确顺序. 此任务被用于学习构象级别的时间域的上下文关系.

Wu 等[58] 对两个下游任务采用线性探测 (linear-probing) 和微调 (fine-tuning) 两种模式测试了模型性能, 并与之前的基于监督学习的多种基线模型进行了对比. 这些基线模型包含四类: 基于序列的、基于表面形状的、基于结构的以及多尺度方法. 与基线模型相比, 在基于 PDBbind 数据集的配体亲和性预测任务中, 线性探测模式的 ProtMD 具有良好的性能, 而经过任务微调的 ProtMD版本性能更优, 具有最小的误差 (RMSE) 和最高的相关系数. 在配体效力预测中 (预测一个配体分子的结合是否能激活蛋白质的功能), 经过微调的 ProtMD 模型预测准确度高于所有基线模型.

Wu 等[58] 还研究了预训练数据集的大小对模型性能的影响. 发现线性探测模式显著地依赖于样本量大小, 当蛋白质-复合体数目超过 50 对时, 模型性能达到最高. 与之相比, 微调模式对预训练样本量依赖程度较低, 较小样本量即可得到好的效果. 他们最后选用了 64 对蛋白质-配体复合物进行预训练, 所得模型对于一个大小为 3K 的测试集依然表现良好, 说明复合物三维结构中蕴含了足够多的相关信息, 从中学习的模型具有优异的泛化性能.

到目前为止, 从分子动力学轨迹学习蛋白质动态性质的预训练模型仍然很少, 大规模的为通用目的而设计的预训练模型还未见报道. 一个于 2010 年开始建立的大型分子模拟轨迹数据库对此类任务可能有帮助[59].

5 融合了知识的蛋白质预训练模型

蛋白质多模态模型另一个重要发展方向是在语言模型的基础上融合基于描述的知识. OntoProtein是第一个把蛋白质功能知识 (gene ontology) 融合到蛋白质表示中的多模态预训练模型[60]. Zhang 等[60] 整理了一个大型的蛋白质知识数据库 (ProteinKG25), 包含约 5M 数据条目, 其形式为三元组 (蛋白质-关系-属性). OntoProtein 的蛋白质编码器采用预训练的 ProtBert[30], 知识编码器使用微软开发的一个针对生物医学语言开发的预训练模型 PubMedBERT[61]. 序列输入和知识三元组分别被两个编码器编码, 并映射到同一个表示空间. 对于序列数据, 预训练采用代理任务为遮蔽率为 15% 的 MLM 方案, 损失函数为真实值与预测值之间的交叉熵. 而对于三元组形式的功能数据, 则利用对比学习技术设计和计算损失函数. 模型同时优化上述两个损失函数, 以获取融合了序列和知识的蛋白质表示. 他们在 TAPE 数据集的三类任务、蛋白-蛋白互作用和蛋白质功能预测等多方面测试了模型性能. 相比之前在大型语料数据集上训练的蛋白质语言模型, OntoProtein 性能稍有提高. Zhang 等[60] 将其归结为目前的功能知识条目偏少, 只能覆盖少部分蛋白质空间.

与 OntoProtein 模型类似, KeAP 致力于在一个更精细的令牌层次对蛋白质和知识进行融合[62]. 具体来说, 对于一个输入的知识三元组 (蛋白质-关系-属性), 蛋白质序列被遮蔽一部分 (约 20%) 后被一个 BERT 型编码器编码, 关系和属性则通过另一个自然语言编码器 PubMedBERT 得到其表示, 然后利用跨模态注意力机制先后从关系数据和属性数据查询与预测被遮蔽氨基酸相关的信息, 并对其进行预测. 与 OntoProtein 相比, KeAP 简化了代理任务, 只使用了 MLM 技术对网络进行预训练. 通过使用和 OntoProtein 类似的微调技术, KeAP 在残基接触预测, 同源探测、稳定性预测、蛋白相互作用、亲和能预测、语义相似性推理等多个下游任务对预训练模型进行了测试, 发现相比于 EMS-1b, ProtBert, OntoProtein 等模型, 预测准确度有显著的提高.

ProtST 也是一个融合蛋白质序列信息与功能信息的多模态预训练模型[63]. 它使用预训练的蛋

白质语言模型 (包括 ProtBert, ESM-1b, ESM-2) 来初始化序列编码器, 用 PubMedBERT 对功能知识进行编码并在后续训练中保持网络权重不变. 模型使用三个代理任务进行预训练, 目的是把序列的表示和知识的表示在语义空间进行对齐. 第一个任务为单模态 MLM 任务, 随机遮蔽 15% 的残基并利用上下文预测这一遮蔽信息. 第二个为多模态对齐任务, 在蛋白质序列和文本描述之间进行对比学习, 以拉近成对的信息在表示空间的距离. 第三个为多模态掩码预测任务. 这一任务随机的遮蔽 15% 的蛋白质序列以及 15% 的文本, 经过一个具有自注意力和交叉注意力的融合网络后, 输出对遮蔽信息的预测. 预训练所用知识数据库 ProtDescribe 包含约 553K 蛋白质序列-属性对. Xu 等[63] 在三类下游任务, 包括蛋白质定位预测、蛋白质突变适应度预测、蛋白质功能预测上对 ProtST 模型进行了微调和测试, 发现它显著优于 CNN 等基线模型, 也优于 ProtBert, OntoProtein, ESM-1b, ESM-2 等模型. 此外, 在亚细胞定位预测、反应类型预测、文本到蛋白搜索几个零样本实验中, 模型也表现出了较好的泛化能力.

6 RNA 预训练模型

RNA 结构和功能预测问题和蛋白质相关问题具有很高的相似性, 相当一部分针对蛋白质发展的计算方法稍加修改即可用于 RNA 领域. 人们很早就开始使用机器学习方法进行 RNA 结构预测, 如 SPOT-RNA 系列[64,65], 3DRNA 系列[66–68], FebRNA[69–71], RNA3DCNN[72], UFold[73], DeepFoldRNA[74], RoseTTAFoldNA[75] 等. 这方面工作完整的介绍可参考最新的综述文献 [76, 77]. 本文只针对 RNA 预训练模型进行介绍.

和蛋白质相比, 针对 RNA 的预训练模型还相对较少. 这可能是因为相对于 20 字符编码的蛋白质序列, 核酸序列是一种四字符编码语言, 相同长度的序列信息量远小于前者, 且 RNA 序列保守性也相对较低. 另外, 相对于 DNA, RNA 具有较多的修饰及高级结构, 更为复杂.

RNA-FM 是一个为通用目的而设计的大型 RNA 预训练语言模型[78]. 模型采用 BERT 架构, 在一个超过 2 千万非编码 RNA 序列数据集上通过自监督学习进行预训练, 训练过程中 15% 的核苷酸被随机遮盖并被模型预测. Chen 等[78] 在多个任务上对这一预训练模型得到的 RNA 表示进行了测试. 在二级结构预测任务上, RNA-FM 相较于之前的如 LinearFold 和 SPOT-RNA 有大幅的提高. 在三维 contact map 预测任务上, 基于 RNA-FM 的 ResNet 模型大幅度领先于一个基于 100 个子模型的集成学习方案. 把 RNA-FM 预测的二级结构和 3dRNA 相结合, 可用于预测 RNA 三级结构. 这一方案在 RNApuzzle 测试集上的平均 RMSD 为 4Å. 基于 RNA-FM 预训练模型, 他们还预测了 SARS-CoV-2 基因主要调控区域的二级结构, 并研究了这一病毒的演化路径, 所得结果均与 ground truth 高度符合. RNA-FM 还被用于协助预测 RNA-蛋白质相互作用. Chen 等[78] 把 RNA-FM 预测的二级结构代替 icSHAPE 实验结构, 使用 PrismNet 预测了海拉细胞中的 RNA-蛋白质相互作用, 发现预测结果全部优于基线模型. 和使用实验结果作为输入的 PrismNet 相比, 使用 RNA-FM 预测值作为输入的模型在 7 种蛋白情况下更优 (共 17 种). 最后, 虽然 RNA-FM 使用非编码 RNA 序列进行训练得到, 它在 mRNA 的 5' 非翻译区核糖体载量预测任务上也展现了良好的性能.

SpliceBERT 是一个在 pre-mRNA 序列数据集上训练的 RNA 语言模型, 主要用于预测 RNA 剪切位点[79]. 训练数据集包含来自 72 种脊椎动物的约 200 万 pre-mRNA, 碱基数目达到了 650 亿. 训练采用 BERT 架构, 单个核苷酸对应一个令牌, 随机遮盖 15% 的核苷酸并对其进行预测, 以强迫网络学习不同位点间的相互关系. 在多物种剪切位点预测和人类分支点预测两个下游任务进行的微调和测试表明, 基于 SpliceBERT 的模型优于传统的基线模型、DNABERT 和只在人类数据上预训练的 SpliceBERT-human. 这显示了在多物种数据上进行预训练的有效性. 与 SpliceBert 类似, 针对 mRNA 发展的语言模型还有 CodonBERT, UTR-LM, 3UTR-BERT 等[80–82], 篇幅关系不能一一详述.

考虑到 RNA 的序列保守性低于蛋白质, Zhang 等[83] 发展了 RNAcmap 方法, 它可提供比 Rfam 数据集更多的同源序列. 在此基础上, 他们采用 MSA Transformer 结构和 BERT 目标函数训练得到了一个 RNA 语言模型 RNA-MSM, 在其输出的二维注意力图和一维嵌入中编码了序列和结构信息. 针对下游任务微调后, 模型在二维碱基对概率预

测和一维溶液可及表面预测任务上, 优于目前的SOTA 方法如 SPOT-RNA2 和 RNA snap2, 也优于基于之前的语言模型 RNA-FM.

Uni-RNA 是一个利用约 10 亿条 RNA 序列进行大规模训练的 RNA 语言模型, 充分挖掘了 RNA 序列的潜在信息[84]. 预训练采用经过效率优化的 BERT 模型. 与 RNA-FM, SPOT-RNA 等方法相比, 基于 Uni-RNA 微调的模型在 RNA 二级结构预测、contact map 预测、mRNA 5'UTR 核糖体载量预测, 3'UTR 亚型占比预测、ncRNA 功能聚类, 剪切位点预测, RNA 修饰位点预测七个任务中均取得了优秀的结果.

RNAErnie 也是一个 RNA 语言模型[85]. 它使用了来自 RNAcentral 数据库的约 2 千万序列进行训练. 训练使用支持连续学习的 Ernie Transformer 架构. 与之前语言模型不同, RNAErnie 进一步把 RNA 片段 (motif) 信息作为先验引入模型. 具体来说, 在自监督预训练阶段, 除在碱基水平的随机遮盖、4—8 碱基长度的子序列随机遮盖之外, 模型还加入了一个片段水平的随机掩码任务, 并将 RNA 类型, 如 miRNA, mRNA, lnRNA 等, 以一个停止词的方式加入到序列尾部, 鼓励模型把不同类型的序列映射到 latent 空间的不同位置, 以更好地支持下游类型引导的微调任务. 在多个下游任务, 包括序列分类、RNA-RNA 互作用预测、和 RNA 二级结构预测, RNAErnie 的性能均大幅优于传统的方法以及之前的语言模型如 RNABert[86], RNA-FM[78] 等.

到目前为止, 据我们所知, RNA 预训练模型均基于序列数据, 尚未见到整合结构信息的模型. 只有 RNAErnie 通过遮盖 RNA 片段序列, 部分地引入了结构信息. 这可能是由于实验解出的 RNA 结构数量远少于蛋白质, 且虽有很多优秀的结构预测模型[65,68,75,87,88], 但尚未见到如 AlphaFold 的革命性突破, 这显示了 RNA 结构预测的难度, 同时说明这是一个大有可为的领域.

7 蛋白质预训练模型与蛋白质设计

蛋白质设计是蛋白质计算领域的一个重要方向. 这方面已经有大量优秀的工作[89–96] 和综述性报告[97–103].

和结构相关的蛋白质设计中, ProteinMPNN 是一个典型的从结构到序列的 Inverse-folding 模型. 它包括编码器和解码器两部分, 其中编码器学习一个和序列无关的蛋白质结构表示, 解码器则通过自回归的方式预测相应的序列[104,105]. ESM-IF1 模型也采用了类似的架构[106].

Baker 组[92,93] 发展了 hallucination 方法. 它首先在序列空间进行蒙特卡罗采样, 并使用 trRosetta 预测结构. 他们还使用类似的框架发展了 Protein Generator, 但把蒙特卡罗采样替换为序列空间的去噪扩散概率模型 (DDPM)[107,108]. 这类模型的特色是把序列空间的优化采样算法和成熟的结构预测模块相结合.

Baker 组还发展了 RFdiffusion 模型进行蛋白质从头设计 (de novo design). 这一模型使用去噪扩散概率模型直接在三维空间从初始噪声生成蛋白质结构, 并利用 ProteinMPNN 设计相匹配的蛋白质序列[109]. RFdiffusion 支持无条件和条件生成, 可进行蛋白质单体、高阶对称寡聚体、功能片段框架、结合蛋白设计等多种任务. 由于直接在结构空间进行去噪扩散生成, 模型生成的结构具有更好的多样性.

与 RFdiffusion 不同, ProteinSGM 在残基间 6 维坐标空间进行去噪扩散以生成结构. 它采用了一个基于随机微分方程的评分生成模型框架, 实现了一个连续的噪声注入和移除策略[110], 并使用 Rosetta 对主链结构进行能量最小化[111]. 这一方案还通过条件生成支持准确和模块化的设计, 可获得和天然蛋白相近的新型蛋白质结构.

上述去噪扩散模型倾向于生成刚性的蛋白质结构, 含有较多的螺旋和较短的 loop 区, 而较少生成对蛋白质功能更重要的柔性和动态结构. PVDQ (protein vector quantization and diffusion) 针对这一问题进行了改进[112]. 这一模型把蛋白质主链结构映射到潜在空间, 并使用一个离散自编码器学习对应的离散表示. 这些离散的表示构成一个代码本 (code book). 通过这种方式, 一个蛋白质主链被映射为一个离散表示序列. PVDQ 在这一潜在空间通过去噪扩散模型进行结构生成, 这一设计允许更高效的采样效率和更平滑的数据分布. 去噪扩散生成的离散表示序列被一个解码器翻译为三维结构, 另一个辅助解码器被用来生成对应的氨基酸序列. 与之前直接在结构空间进行去噪生成的模型不同, PVDQ 模型展现出了更强的生成 β 片和长 loop

区的能力, 这些结构具有较小的刚性和更好的动态性. PVQD 模型也支持条件概率生成.

蛋白质设计模型通常仅利用具有实验或预测结构的序列进行训练, 无法利用海量的结构未知的序列. 本文介绍的融合蛋白质结构的预训练模型可用于解决这一问题. 正如 LM-DESIGN 工作所指出的, 融合结构信息的语言模型是一个蛋白质设计器[113]. 这一模型把结构编码器 (如 GNN) 的输出和语言模型 (如 ESM 系列) 相结合, 利用语言模型的生成能力进行序列解码, 并通过反复迭代的方法对序列进行优化. 又如 MIF-ST 模型把一个预训练的蛋白质语言模型 (CARP-640M) 和一个表征蛋白质结构的图网络结合起来, 并使用 MLM 方案进行预训练[114]. 这些模型在核心架构上和前文介绍的 STEPS[49] 和 LM-GVP[48] 等模型非常类似, 显示了蛋白质预训练模型和蛋白质设计等不同任务在架构设计上逐渐合流的趋势[102].

8 讨 论

深度学习技术的成功, 在多个科学领域催生了新的思路和研究范式. 其中最具有代表性的是 AlphaFold 系列. 2023 年来, 以 ChatGPT 为代表的自然语言大模型取得了空前的成功, 并且快速朝着多模态大模型发展, 以融合更多的数据, 训练更大的模型. 在这个领域, 模型的大小和算力是推动性能提升的主要力量[115]. 自然语言处理领域的若干关键技术, 如模型预训练、自监督学习范式、被迅速借鉴到生物学领域. 自 2019 年开始, 尤其是近三年来, 人们发展了多种蛋白质预训练模型并应用于各种下游任务. 这一新的研究范式, 不仅可以充分利用海量无标注数据以提供强大且通用的表征能力, 为多种下游任务提供统一的框架并便于快速部署, 且特别有利于某些缺乏标注数据的下游任务, 可在相当程度上解决某些领域标注数据严重不足的问题.

到目前为止, 针对蛋白质序列的预训练模型已基本成熟. 考虑到蛋白质结构承载了更大的信息量, 且蛋白质功能主要和结构相关, 越来越多的工作开始关注如何把结构信息更好地融入蛋白质的表示空间. 从学科趋势上看, 蛋白质预训练模型明显地朝着更多模态发展, 以融合空间结构、物理化学知识、功能数据、甚至动态结构等信息, 以期多种数据的交叉融合能够催生出更强大的模型. 本文对这一方向的进展进行了回顾和总结.

本文所介绍的模型各有其优缺点和特色. Evoformer 和 LM-GVP 代表了同时融合序列和结构信息的、为特定目的而设计的蛋白质模型, 其中 Evoformer 针对蛋白质结构预测、而 LM-GVP 针对功能预测. 虽然它们都是为特定任务而设计, 但它们给出的特征表示均具有一定的通用性. 尤其是 Evoformer, 已被实验证实可泛化到比如功能预测任务, 虽然性能上相比 ESM 系列略差. 在为通用目的而设计的蛋白质预训练模型中, BB-model[19,42] 具有开创性且富有特色, 这一模型利用多任务学习框架同时在序列和结构上对模型进行训练, 且在下游任务只使用序列进行推理 (经过微调). 相比较而言, Guo-model[43], New IEConv[44], GearNet[46] 等模型, 在下游任务进行推理时必须提供蛋白质结构, 虽然这提高了模型的准确度, 但也同时限制了其应用范围. 从训练方法看, 大部分预训练模型借鉴了自然语言处理中的 MLM 方法或图像处理中的 Masked AutoEncoder(MAE) 方法[116], 也有部分采用对比学习方案[117], 不同训练方案在分子结构领域的有效性目前尚无定论. 另外, SaProt 提出了一个创新性的训练方案, 它把蛋白质三级结构信息通过 Foldseek 工具编码成与氨基酸序列等长的 token 序列, 和氨基酸对应的 token 结合, 将输入序列和结构转化成一个语句. 这样做的好处是可以无缝地使用 NSP 领域成熟的语言模型, 且可用于处理大规模数据[51,52]. 从数据模态角度看, 主流模型重点关注如何融合序列和结构信息, 如 BB-model[19,42], Guo-model[43], GearNet, STEPS, SaProt 等, 而 HoloProt 则引入了蛋白质表面形貌信息, ProtMD 模型从分子模拟数据中进行学习以建模蛋白质的动态特征, OntoProtein 和 ProtST 等模型则侧重于融合序列和功能信息. 另外, 生物计算领域还有相当数量的工作致力于集成 DNA、RNA、功能等多来源、多模态数据, 如 xTrimo[31]、Evo[32] 等. 一个全面的总结见文献 [35].

多模态模型的预训练需要大量配对数据, 如匹配蛋白质序列和功能描述. 然而配对数据通常很稀缺, 导致多模态模型训练困难. Biobridge 方案尝试解决这一问题[118]. 它不试图训练一个多模态模型, 而是使用知识图谱训练一个对齐模型, 把多个单模态的表示空间进行对齐, 把它们连接起来以解决多

模态任务. 这一方案同时解决了多模态模型计算量过大的问题, 是一个有益的探索. 最后, 通过采用主动学习 (active learning) 的方式, 有目的地选择配对数据样本, 亦可降低对数据量的要求.

和海量蛋白质序列相比, 三维结构数据的数量偏少可能并不是一个严重的问题. 这是考虑到与序列信息相比, 空间结构信息可能更相似于自然语言, 具有较高的信息密度. 首先, 和自然语言中的句子不同, 单一蛋白质序列并没有明显的语义特征, 难以找出相当于词的单位以及它们之间的相互关系. 反观空间结构, 由于共价键的刚性, 原子团具有明显的化学意义且种类并不太多, 可被视为基本结构单元, 并对应于自然语言中的词. 原子团之间的相互作用相当于句子中词的相互作用. 此外, 由于物理化学上的限制, 原子团之间的堆积模式可能并不太多, 无需海量实验结构即可覆盖大部分可能的相空间. 当然, 由于 AlphaFold 系列的成功, 目前可用的蛋白质三维结构被大大扩充了. 通过对目前融合了结构的多模态模型进行分析 (见表 1), 我们预测, 与蛋白质语言模型对序列数量的需求相比, 融合结构信息的多模态模型可能并不需要海量的三维结构.

将先验知识引入预训练模型也是一个重要研究方向. 这不仅可以利用现有的知识, 且可以丰富训练数据、增强模型泛化能力、提高模型的可解释性等. 在蛋白质计算领域, 目前常见工作是把功能相关的描述以自然语言编码器编码, 或以知识图谱形式通过对比学习融入模型. 然而, 先验知识的形式多种多样, 如逻辑规则、知识图谱、数学物理方程、人类反馈等[119]. 对于生物大分子结构来说, 如何把诸如长程静电相互作用等物理化学知识直接引入预训练模型, 是一个值得探索的方向. 这对于 RNA 结构尤其重要, 因为长程的静电相互作用是其结构稳定性的决定性因素之一[120–122]. 而目前常见的预训练模型中, 无论是遮蔽重建还是对比学习方案, 均局限于短程相互作用.

训练大模型通常需要庞大的算力. 如 ESM 系列, xTrimo 系列, 均需要大量的 GPU 进行训练. 然而, 纵观本文提到的多模态模型, 大部分并不需要十分强大的算力. 这一方面是因为多模态如结构数据、蛋白质功能数据并不十分庞大, 另一方面是因为这些模型利用了已预训练的单模态模型. 如 ProtST 模型分别利用 ESM 系列和 PubMedBERT 编码蛋白质序列信息和功能描述, 并冻结 PubMedBERT 的模型权重, 通过对比学习把蛋白质序列的表示和功能的表示进行对齐, 极大地降低了训练所需算力. 另外, 对于训练多模态模型, Biobridge 方案也可降低对算力的需求.

虽然蛋白质预训练模型领域已经取得了很多进展, 但仍面临诸多挑战. 最显著的问题是缺乏统一 benchmark, 难以判断各模型优劣. 另外, 由于使用大量数据训练模型, 测试数据的信息泄露到训练集中也是常见问题. 蛋白质结构的动态性也是目前大部分模型未考虑的问题. 然而蛋白质这一特性对其生物功能至关重要, 尤其是对于可变构蛋白和天然无序蛋白, 以及蛋白质-药物的非刚性结合, 蛋白质-RNA 相互作用等问题. 虽然 ProtMD 方法从 64 个蛋白质-配体复合体的分子动力学模拟轨迹中学习了结合界面的动态特性, 但由于训练数据集偏小 (62.8 K 构象), 模型的通用性和泛化能力尚未可知.

总之, 近三年来, 融合了蛋白质结构信息的预训练模型, 以及融合了更多模态信息的预训练模型如雨后春笋般出现. 这是一个令人兴奋的、新兴的交叉学科. 然而, 由于其多学科交叉特性、可用数据及算力的限制, 这一领域还处于发展早期, 仍面临诸多困难和挑战, 有大量工作可做. 本文希望能为刚进入这一领域的研究者提供一些指引和帮助.

参考文献

[1] Senior A W, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson A W, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones D T, Silver D, Kavukcuoglu K, Hassabis D 2020 *Nature* **577** 706

[2] 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 A, Ballard A J, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman 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

[3] Radford A, Narasimhan K, Salimans T, Sutskever I 2018 *Improving Language Understanding by Generative Pre-Training* [2024-6-9]

[4] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I 2019 *Language Models are Unsupervised Multitask Learners* [2024-6-9]

[5] Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen

M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodeis D 2020 arXiv: 2005.14165[cs.CV]
[6] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Low R 2022 arXiv: 2203.02155[cs.CV]
[7] Devlin J, Chang M W, Lee K, Toutanova K 2018 arXiv: 1810.04805[cs.CV]
[8] Ma Z, He J, Qiu J, Cao H, Wang Y, Sun Z, Zheng L, Wang H, Tang S, Zheng T, Lin J, Feng G, Huang Z, Gao J, Zeng A, Zhang J, Zhong R, Shi T, Liu S, Zheng W, Tang J, Yang H, Liu X, Zhai J, Chen W 2022 *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* Seoul, Republic of Korea, April 2–6, 2022 p192
[9] Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, Qiu J, Yao Y, Zhang A, Zhang L, Han W, Huang M, Jin Q, Lan Y, Liu Y, Liu Z, Lu Z, Qiu X, Song R, Tang J, Wen J R, Yuan J, Zhao W X, Zhu J 2021 arXiv: 2106.07139[AI]
[10] Yuan S, Zhao H, Zhao S, et al. 2022 arXiv: 2203.14101 [cs.LG]
[11] Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, Chi E H, Hashimoto T, Vinyals O, Liang P, Dean J, Fedus W 2022 arXiv: 2206.07682[cs.CV]
[12] Alayrac J B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R, Rutherford E, Cabi S, Han T, Gong Z, Samangooei S, Monteiro M, Menick J, Borgeaud S, Brock A, Nematzadeh A, Sharifzadeh S, Binkowski M, Barreira R, Vinyals O, Zisserman A, Simonyan K 2022 arXiv: 2204.14198[cs.CV]
[13] OpenAI, Achiam J, Adler S, et al. 2024 arXiv: 2303.08774 [cs.CV]
[14] Driess D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu T, Huang W, Chebotar Y, Sermanet P, Duckworth D, Levine S, Vanhoucke V, Hausman K, Toussaint M, Greff K, Zeng A, Mordatch I, Florence P 2023 arXiv: 2303.03378[cs.LG]
[15] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G 2023 arXiv: 2302.13971[cs.CV]
[16] Gemini Team Google, Anil R, Borgeaud S, et al. 2024 arXiv: 2312.11805[cs.CV]
[17] Chen F, Han M, Zhao H, Zhang Q, Shi J, Xu S, Xu B 2023 arXiv: 2305.04160[cs.CV]
[18] Li K, He Y, Wang Y, Li Y, Wang W, Luo P, Wang Y, Wang L, Qiao Y 2023 arXiv: 2305.06355[cs.CV]
[19] Bepler T, Berger B 2019 arXiv: 1902.08661[cs.LG]
[20] Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B 2019 bioRxiv: 614313[Bioinformatics]
[21] Alley E C, Khimulya G, Biswas S, Alquraishi M, Church G M 2019 *Nat. Methods* **16** 1315
[22] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick C L, Ma J, Fergus R 2021 *Proc. Natl. Acad. Sci.* **118** e2016239118
[23] Rao R, Liu J, Verkuil R, et al. 2021 bioRxiv: 2021.02.12.430858 [Synthetic Biology]
[24] Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A 2021 *Advances in Neural Information Processing Systems* **34** 29287
[25] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A 2023 *Science* **379** 1123
[26] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Santos Costa A d, Fazel-Zarandi M, Sercu T, Candido S, Rives A 2022 bioRxiv: 2022.07.20.500902[Synthetic Biology]
[27] Madani A, McCann B, Naik N, Keskar N S, Anand N, Eguchi R R, Huang P S, Socher R 2020 arXiv: 2004.03497[q-bio.QM]
[28] Madani A, Krause B, Greene E R, Subramanian S, Mohr B P, Holton J M, Olmos J L, Xiong C, Sun Z Z, Socher R, Fraser J S, Naik N 2023 *Nat. Biotechnol.* **41** 1099
[29] He L, Zhang S, Wu L, Xia H, Ju F, Zhang H, Liu S, Xia Y, Zhu J, Deng P, Shao B, Qin T, Liu T Y 2021 arXiv: 2110.15527[cs.CV]
[30] Elnaggar A, Heinzinger M, Dallago C, Rihawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, Bhowmik D, Rost B 2021 arXiv: 2007.06225[cs.LG]
[31] Chen B, Cheng X, Li P, Geng Y, Gong J, Li S, Bei Z, Tan X, Wang B, Zeng X, Liu C, Zeng A, Dong Y, Tang J, Song L 2024 arXiv: 2401.06199[q-bio.QM]
[32] Nguyen E, Poli M, Durrant M G, Thomas A W, Kang B, Sullivan J, Ng M Y, Lewis A, Patel A, Lou A, Ermon S, Baccus S A, Hernandez-Boussard T, Ré C, Hsu P D, Hie B L 2024 bioRxiv: 2024.02.27.582234[Synthetic Biology]
[33] Gao W, Mahajan S P, Sulam J, Gray J J 2020 *Patterns* **1** 100142
[34] Unsal S, Atas H, Albayrak M, Turhan K, Acar A C, Doğan T 2022 *Nature Machine Intelligence* **4** 227
[35] Zhang Q, Ding K, Lyv T, Wang X, Yin Q, Zhang Y, Yu J, Wang Y, Li X, Xiang Z, Feng K, Zhuang X, Wang Z, Qin M, Zhang M, Zhang J, Cui J, Huang T, Yan P, Xu R, Chen H, Li X, Fan X, Xing H, Chen H 2024 arXiv: 2401.14656[cs.CV]
[36] Guan X Y, Huang H Y, Peng H Q, Liu Y H, Li W F, Wang W 2023 *Acta Phys. Sin.* **72** 248708 (in Chinese) [管星悦, 黄恒焱, 彭华祺, 刘彦航, 李文飞, 王炜 2023 物理学报 **72** 248708]
[37] Chen G L, Zhang Z Y 2023 *Acta Phys. Sin.* **72** 248705 (in Chinese) [陈光临, 张志勇 2023 物理学报 **72** 248705]
[38] Zhang J H 2024 *Acta Phys. Sin.* **73** 069301 (in Chinese) [张嘉晖 2024 物理学报 **73** 069301]
[39] Zeng C, Jian Y, Vosoughi S, Zeng C, Zhao Y 2023 *Nat. Commun.* **14** 1060
[40] Zeng C, Zhao Y 2023 *Scientia Sinica Physica, Mechanica & Astronomica* **53** 290018
[41] Huh M, Cheung B, Wang T, Isola P 2024 arXiv: 2405.07987 [cs.LG]
[42] Bepler T, Berger B 2021 *Cell Systems* **12** 654
[43] Guo Y, Wu J, Ma H, Huang J 2022 *Proceedings of the AAAI Conference on Artificial Intelligence* **36** 6801
[44] Hermosilla P, Ropinski T 2022 arXiv: 2205.15675[q-bio.BM]
[45] Zhang Z, Xu M, Jamasb A, Chenthamarakshan V, Lozano A, Das P, Tang J 2022 arXiv: 2203.06125[cs.LG]
[46] Zhang Z, Xu M, Lozano A, Chenthamarakshan V, Das P, Tang J 2023 arXiv: 2303.06275[q-bio.QM]
[47] Gligorijević V, Renfrew P D, Kosciolek T, Leman J K, Berenberg D, Vatanen T, Chandler C, Taylor B C, Fisk I M, Vlamakis H, Xavier R J, Knight R, Cho K, Bonneau R 2021 *Nat. Commun.* **12** 3168
[48] Wang Z, Combs S A, Brand R, Calvo M R, Xu P, Price G, Golovach N, Salawu E O, Wise C J, Ponnapalli S P, Clark P M 2022 *Sci. Rep.* **12** 6832
[49] Chen C, Zhou J, Wang F, Liu X, Dou D 2023 arXiv: 2204.04213[cs.LG]
[50] Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, Zhang L, Ke G 2022 DOI: 10.26434/chemrxiv-2022-jjm0j-v4
[51] Su J, Han C, Zhou Y, Shan J, Zhou X, Yuan F 2023

bioRxiv: 2023.10.01.560349[Bioinformatics]
[52] Su J, Li Z, Han C, Zhou Y, Shan J, Zhou X, Ma D, OPMC T, Ovchinnikov S, Yuan F 2024 bioRxiv: 2024.05.24.595648 [Bioinformatics]
[53] Hu M Y, Yuan F J, Yang K K, Ju F S, Su J, Wang H, Yang F, Ding Q Y 2022 arXiv:2206.06583 [q-bio.QM]
[54] Abramson J, Adler J, Dunger J, et al. 2024 *Nature* **630** 493
[55] Wang L, Liu H, Liu Y, Kurtin J, Ji S 2022 arXiv: 2207.12600[cs.LG]
[56] Somnath V R, Bunne C, Krause A 2021 arXiv: 2204.02337[cs.LG]
[57] Gainza P, Sverrisson F, Monti F, Rodola E, Boscaini D, Bronstein M M, Correia B E 2020 *Nat. Methods* **17** 184
[58] Wu F, Jin S, Jiang Y, Jin X, Tang B, Niu Z, Liu X, Zhang Q, Zeng X, Li S Z 2022 arXiv: 2204.08663[CE]
[59] Meyer T, D'Abramo M, Rueda M, Ferrer-Costa C, Pérez A, Carrillo O, Camps J, Fenollosa C, Repchevsky D, Gelpí J L, Orozco M 2010 *Structure* **18** 1399
[60] Zhang N, Bi Z, Liang X, Cheng S, Hong H, Deng S, Lian J, Zhang Q, Chen H 2022 arXiv: 2201.11147[q-bio.BM]
[61] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H 2021 arXiv: 2007.15779[cs.CV]
[62] Zhou H Y, Fu Y, Zhang Z, Bian C, Yu Y 2023 arXiv: 2301.13154[cs.LG]
[63] Xu M, Yuan X, Miret S, Tang J 2023 arXiv: 2301.12040 [q-bio.BM]
[64] Singh J, Hanson J, Paliwal K, Zhou Y 2019 *Nat. Commun.* **10** 5407
[65] Singh J, Paliwal K, Zhang T, Singh J, Litfin T, Zhou Y 2021 *Bioinformatics* **37** 2589
[66] Wang J, Mao K, Zhao Y, Zeng C, Xiang J, Zhang Y, Xiao Y 2017 *Nucleic Acids Res.* **45** 6299
[67] Wang J, Xiao Y 2017 *Current Protocols in Bioinformatics* **57** 5
[68] Wang J, Wang J, Huang Y, Xiao Y 2019 *Int. J. Mol. Sci.* **20** 4116
[69] Tan Y L, Wang X, Shi Y Z, Zhang W, Tan Z J 2022 *Biophys. J.* **121** 142
[70] Zhou L, Wang X, Yu S, Tan Y L, Tan Z J 2022 *Biophys. J.* **121** 3381
[71] Wang X, Tan Y L, Yu S, Shi Y Z, Tan Z J 2023 *Biophys. J.* **122** 1503
[72] Li J, Zhu W, Wang J, Li W, Gong S, Zhang J, Wang W 2018 *PLoS Comput. Biol.* **14** e1006514
[73] Fu L, Cao Y, Wu J, Peng Q, Nie Q, Xie X 2022 *Nucleic Acids Res.* **50** e14
[74] Pearce R, Omenn G S, Zhang Y 2022 bioRxiv: 2022.05.15.491755[Bioinformatics]
[75] Baek M, McHugh R, Anishchenko I, Baker D, DiMaio F 2022 bioRxiv: 2022.09.09.507333[Bioinformatics]
[76] Zhang J, Lang M, Zhou Y, Zhang Y 2024 *Trends in Genetics* **40** 94
[77] Li J, Zhou Y, Chen S J 2024 *Curr. Opin. Struct. Biol.* **87** 102847
[78] Chen J, Hu Z, Sun S, Tan Q, Wang Y, Yu Q, Zong L, Hong L, Xiao J, Shen T, King I, Li Y 2022 arXiv: 2204.00300[q-bio.QM]
[79] Chen K, Zhou Y, Ding M, Wang Y, Ren Z, Yang Y 2023 bioRxiv: 2023.01.31.526427[Bioinformatics]
[80] Babjac A N, Lu Z, Emrich S J 2023 *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* New York, United States, September 3–6, 2023 p1
[81] Chu Y, Yu D, Li Y, Huang K, Shen Y, Cong L, Zhang J, Wang M 2024 *Nature Machine Intelligence* **6** 449
[82] Yang Y, Li G, Pang K, Cao W, Li X, Zhang Z 2023 bioRxiv: 2023.09.08.556883[Bioinformatics]
[83] Zhang Y, Lang M, Jiang J, Gao Z, Xu F, Litfin T, Chen K, Singh J, Huang X, Song G, Tian Y, Zhan J, Chen J, Zhou Y 2024 *Nucleic Acids Res.* **52** e3
[84] Wang X, Gu R, Chen Z, Li Y, Ji X, Ke G, Wen H 2023 bioRxiv: 2023.07.11.548588[Bioinformatics]
[85] Wang N, Bian J, Li Y, Li X, Mumtaz S, Kong L, Xiong H 2024 *Nature Machine Intelligence* **6** 548
[86] Akiyama M, Sakakibara Y 2022 *NAR Genomics and Bioinformatics* **4** lqac012
[87] Shen T, Hu Z, Peng Z, Chen J, Xiong P, Hong L, Zheng L, Wang Y, King I, Wang S, Siqi S, Yu L 2022 arXiv: 2207.01586[q-bio.QM]
[88] Li Y, Zhang C, Feng C, Pearce R, Lydia Freddolino P, Zhang Y 2023 *Nat. Commun.* **14** 5745
[89] Ferruz N, Schmidt S, Höcker B 2022 *Nat. Commun.* **13** 4348
[90] Wang J, Lisanza S, Juergens D, Tischer D, Watson J L, Castro K M, Ragotte R, Saragovi A, Milles L F, Baek M, Anishchenko I, Yang W, Hicks D R, Expòsit M, Schlichthaerle T, Chun J H, Dauparas J, Bennett N, Wicky B I M, Muenks A, DiMaio F, Correia B, Ovchinnikov S, Baker D 2022 *Science* **377** 387
[91] Trippe B L, Yim J, Tischer D, Baker D, Broderick T, Barzilay R, Jaakkola T 2022 arXiv: 2206.04119[q-bio.BM]
[92] Anishchenko I, Pellock S J, Chidyausiku T M, Ramelot T A, Ovchinnikov S, Hao J, Bafna K, Norn C, Kang A, Bera A K, DiMaio F, Carter L, Chow C M, Montelione G T, Baker D 2021 *Nature* **600** 547
[93] Wicky B I M, Milles L F, Courbet A, Ragotte R J, Dauparas J, Kinfu E, Tipps S, Kibler R D, Baek M, DiMaio F, Li X, Carter L, Kang A, Nguyen H, Bera A K, Baker D 2022 *Science* **378** 56
[94] Anand N, Achim T 2022 arXiv: 2205.15019[q-bio.QM]
[95] Luo S, Su Y, Peng X, Wang S, Peng J, Ma J 2022 *Advances in Neural Information Processing Systems* **35** 9754
[96] Cao L, Coventry B, Goreshnik I, et al 2022 *Nature* **605** 551
[97] Kuhlman B, Bradley P 2019 *Nat. Rev. Mol. Cell Biol.* **20** 681
[98] Pan X, Kortemme T 2021 *J. Biol. Chem.* **296** 100558
[99] Khakzad H, Igashov I, Schneuing A, Goverde C, Bronstein M, Correia B 2023 *Cell Systems* **14** 925
[100] Malbranke C, Bikard D, Cocco S, Monasson R, Tubiana J 2023 *Curr. Opin. Struct. Biol.* **80** 102571
[101] Kortemme T 2024 *Cell* **187** 526
[102] Notin P, Rollins N, Gal Y, Sander C, Marks D 2024 *Nat. Biotechnol.* **42** 216
[103] Listov D, Goverde C A, Correia B E, Fleishman S J 2024 *Nat. Rev. Mol. Cell Biol.* **25** 639
[104] Ingraham J, Garg V K, Barzilay R, Jaakkola T 2019 *Proceedings of the 33rd International Conference on Neural Information Processing Systems* Vancouver, BC, Canada, December 8–14, 2019 p15820
[105] Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte R J, Milles L F, Wicky B I M, Courbet A, de Haas R J, Bethel N, Leung P J Y, Huddy T F, Pellock S, Tischer D, Chan F, Koepnick B, Nguyen H, Kang A, Sankaran B, Bera A K, King N P, Baker D 2022 *Science* **378** 49
[106] Hsu C, Verkuil R, Liu J, Lin Z, Hie B, Sercu T, Lerer A, Rives A 2022 bioRxiv: 2022.04.10.487779[Systems Biology]
[107] Sohl-Dickstein J, Weiss E A, Maheswaranathan N, Ganguli S 2015 arXiv: 1503.03585[cs.LG]
[108] Ho J, Jain A, Abbeel P 2020 *Advances in Neural Information Processing Systems* **33** 6840
[109] Watson J L, Juergens D, Bennett N R, et al 2023 *Nature* **620** 1089
[110] Song Y, Sohl-Dickstein J, Kingma D P, Kumar A, Ermon S, Poole B 2020 arXiv: 2011.13456[cs.LG]

[111] Lee J S, Kim J, Kim P M 2023 *Nature Computational Science* **3** 382
[112] Liu Y, Chen L, Liu H 2023 bioRxiv: 2023.11.18.567666 [Bioinformatics]
[113] Zheng Z, Deng Y, Xue D, Zhou Y, YE F, Gu Q 2023 arXiv: 2302.01649[cs.LG]
[114] Yang K K, Zanichelli N, Yeh H 2023 *Protein Eng. Des. Sel.* **36** gzad015
[115] Kaplan J, McCandlish S, Henighan T, Brown T B, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D 2020 arXiv: 2001.08361[cs.LG]
[116] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R 2021 arXiv: 2111.06377[cs.CV]
[117] Chen T, Kornblith S, Norouzi M, Hinton G 2020 arXiv: 2002.05709[cs.LG]
[118] Wang Z, Wang Z, Srinivasan B, Ioannidis V N, Rangwala H, Anubhai R 2023 arXiv: 2310.03320[cs.LG]
[119] Von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, Kirsch B, Walczak M, Pfrommer J, Pick A, Ramamurthy R, Garcke J, Bauckhage C, Schuecker J 2021 *IEEE Trans. Knowl. Data Eng.* **35** 614
[120] Bao L, Zhang X, Jin L, Tan Z J 2015 *Chin. Phys. B* **25** 018703
[121] Qiang X W, Zhang C, Dong H L, Tian F J, Fu H, Yang Y J, Dai L, Zhang X H, Tan Z J 2022 *Phys. Rev. Lett.* **128** 108103
[122] Dong H L, Zhang C, Dai L, Zhang Y, Zhang X H, Tan Z J 2024 *Nucleic Acids Res.* **52** 2519

SPECIAL TOPIC—Machine learning in biomolecular modelling

Progress in protein pre-training models integrating structural knowledge*

Tang Tian-Yi [1)] Xiong Yi-Ming [1)] Zhang Rui-Ge [1)] Zhang Jian [1)2)†]
Li Wen-Fei [1)2)] Wang Jun [1)2)] Wang Wei [1)2)‡]

1) (*School of Physics*, *Nanjing University*, *Nanjing* 210093, *China*)

2) (*Institute of Brain Science*, *Nanjing University*, *Nanjing* 210093, *China*)

(Received 7 June 2024; revised manuscript received 12 July 2024)

Abstract

The AI revolution, sparked by natural language and image processing, has brought new ideas and research paradigms to the field of protein computing. One significant advancement is the development of pre-training protein language models through self-supervised learning from massive protein sequences. These pre-trained models encode various information about protein sequences, evolution, structures, and even functions, which can be easily transferred to various downstream tasks and demonstrate robust generalization capabilities. Recently, researchers have further developed multimodal pre-trained models that integrate more diverse types of data. The recent studies in this direction are summarized and reviewed from the following aspects in this paper. Firstly, the protein pre-training models that integrate protein structures into language models are reviewed: this is particularly important, for protein structure is the primary determinant of its function. Secondly, the pre-trained models that integrate protein dynamic information are introduced. These models may benefit downstream tasks such as protein-protein interactions, soft docking of ligands, and interactions involving allosteric proteins and intrinsic disordered proteins. Thirdly, the pre-trained models that integrate knowledge such as gene ontology are described. Fourthly, we briefly introduce pre-trained models in RNA fields. Finally, we introduce the most recent developments in protein designs and discuss the relationship of these models with the aforementioned pre-trained models that integrate protein structure information.

Keywords: protein foundation model, protein multi-modal model, protein structure, machine learning

PACS: 87.10.Vg, 87.16.A–, 87.14.E–, 87.15.A– **DOI:** 10.7498/aps.73.20240811

* Project supported by the Science and Technology Innovation Project of the Ministry of Science and Technology (Grant No. 2030-2021ZD0201300) and the National Natural Science Foundation of China (Grant No. 11934008).

† Corresponding author. E-mail: jzhang@nju.edu.cn

‡ Corresponding author. E-mail: wangwei@nju.edu.cn

物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

融合结构知识的蛋白质预训练模型进展

汤天一 熊翊名 张睿格 张建 李文飞 王骏 王炜

Progress in protein pre-training models integrating structural knowledge

Tang Tian-Yi Xiong Yi-Ming Zhang Rui-Ge Zhang Jian Li Wen-Fei Wang Jun Wang Wei

引用信息 Citation: Acta Physica Sinica, 73, 188701 (2024) DOI: 10.7498/aps.73.20240811
在线阅读 View online: https://doi.org/10.7498/aps.73.20240811
当期内容 View table of contents: http://wulixb.iphy.ac.cn
