特约评述

DOI: 10.12211/2096-8280.2025-016

统计物理与人工智能驱动的蛋白质结构生物信息学

夏辰亮[1]，张泽成[2]，管星悦[3]，唐乾元[2]
（[1] 三江学院数理部，江苏 南京 210012；[2] 香港浸会大学物理系，香港 999077；[3] 南京大学物理学院，江苏 南京 210093）

摘要：结构生物信息学聚焦于生物分子的三维结构及其功能，蛋白质的结构是其核心研究对象。深度学习引发的蛋白质结构预测革命，特别是AlphaFold2的突破，实现了仅凭氨基酸序列即可达到原子精度的蛋白质结构预测，从根本上重构了该领域的数据生态。统计物理学与大数据分析方法的深度融合，使研究者能够突破传统个案研究的局限，从海量数据中系统性揭示蛋白质设计的普适性规律。大规模蛋白质结构数据的积累为定量化研究蛋白质动力学中的长程关联及其与进化的对应关系奠定了重要基础，这不仅为理解蛋白质的结构、动力学、功能与进化提供了统一的理论框架，其揭示的普适规律与设计原则也为人工蛋白质设计提供了关键指导。在此基础上，基于AlphaFold数据库的跨物种蛋白质结构对比统计分析，突显了数据驱动方法在揭示蛋白质进化过程中随生物复杂性增加而呈现的普适统计规律方面的核心作用，为理解生命进化的分子机制提供了全新视角。鉴于蛋白质功能的实现往往依赖于多种构象状态间的动态转换，蛋白质动力学的精确预测已成为当前研究的核心方向。统计物理与人工智能相结合的研究范式将持续引领蛋白质科学的创新发展，通过提升高通量筛选和理性设计效率，加速从基础发现到实际应用的转化，为合成生物学、精准医学等领域开辟新的可能性。

关键词：统计物理；人工智能；蛋白质结构；蛋白质动力学；结构生物信息学；AlphaFold数据库

中图分类号：Q615　　**文献标志码：**A

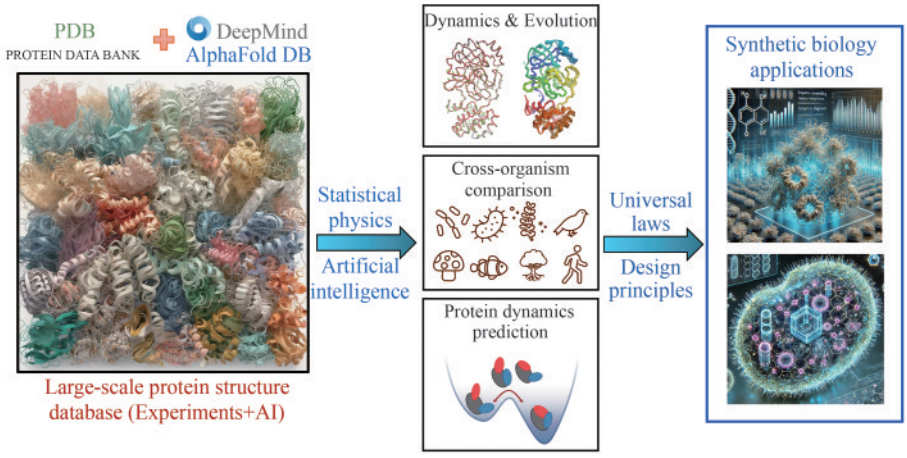
Protein structural bioinformatics empowered by statistical physics and artificial intelligence

XIA Chenliang[1], ZHANG Zecheng[2], GUAN Xingyue[3], TANG Qianyuan[2]
([1]*Department of Mathematics and Physics, Sanjiang University, Nanjing 210012, Jiangsu, China*; [2]*Department of Physics, Hong Kong Baptist University, Hong Kong 999077, China*; [3]*School of Physics, Nanjing University, Nanjing 210093, Jiangsu, China*)

Abstract: Structural bioinformatics focuses on the computational study of three-dimensional biomolecular structures

收稿日期：2025-03-17 修回日期：2025-04-15
基金项目：国家自然科学基金（12305052）；江苏省高等学校自然科学研究项目（22KJD14005）；香港研究资助局杰出青年学者计划（22302723）；香港浸会大学资助项目（RC-FNRA-IG/22-23/SCI/03）
引用本文：夏辰亮，张泽成，管星悦，唐乾元. 统计物理与人工智能驱动的蛋白质结构生物信息学[J]. 合成生物学，2025，6(3)：547-565
Citation：XIA Chenliang，ZHANG Zecheng，GUAN Xingyue，TANG Qianyuan. Protein structural bioinformatics empowered by statistical physics and artificial intelligence[J]. Synthetic Biology Journal，2025，6(3)：547-565

and their functions, with protein structures as its core research object. Traditional research in this field relied on protein structure databases of experimentally determined proteins but was constrained by the high cost and low-throughput nature of experimental methods. The revolution in protein structure prediction driven by deep learning, particularly AlphaFold2's breakthrough, has fundamentally transformed the field's data landscape by achieving atomic-level prediction accuracy from amino acid sequences alone. The deep integration of statistical physics with big data analysis methodologies has enabled researchers to overcome limitations of traditional case-by-case studies, systematically revealing universal principles of protein design from massive datasets. The accumulation of extensive protein structure data provides a crucial foundation for quantifying long-range correlations in protein dynamics and their evolutionary correspondence, revealing universal principles rooted in the interplay between sequence variability, structural constraint, and functional optimization. These principles not only offer a unified framework for understanding protein structure, dynamics, function, and evolution but also serve as the basis for predictive models and *de novo* protein design in engineering applications. Building upon this foundation, statistical analyses based on the AlphaFold Database highlight the crucial role of data-driven methods in uncovering universal statistical laws and dimensionality reduction principles in protein evolution across increasing organismal complexity, offering fresh perspectives on the fundamental constraints and convergent patterns driving molecular evolution. Recognizing that protein functions often depends on transitions between multiple conformational states, precise prediction of protein dynamics has become a core research direction. These advances are propelling protein engineering into an era of precise rational design where researchers can predict and manipulate conformational change pathways to regulate enzyme activity, optimize ligand specificity, and design allosteric responses with unprecedented precision. The research paradigm combining statistical physics and artificial intelligence continues to drive innovation in protein science, enhancing high-throughput screening and rational design efficiency to accelerate translation from basic discoveries to practical applications. As computational capabilities advance and AI models evolve, the field progresses from single protein design toward complex biological system construction, opening new frontiers in synthetic biology, precision medicine, and other applications.



Keywords: statistical physics; artificial intelligence; protein structure; protein dynamics; structural bioinformatics; AlphaFold database

结构生物信息学与传统以序列分析为核心的生物信息学不同，其主要聚焦于生物分子的三维结构及其功能。蛋白质的结构是这一领域的核心研究对象，研究高度依赖于海量的结构数据，主要采用大数据分析、统计建模、机器学习以及计算模拟等方法，以揭示蛋白质结构与功能之间的

复杂关系[1]。近年来，深度学习技术的突破，特别是从氨基酸序列实现原子级精度的蛋白质结构预测，彻底重塑了该领域的数据生态。以AlphaFold2为代表的人工智能（AI）工具，不仅为蛋白质结构预测带来了革命性进展[2]，其建立的AlphaFold数据库（AlphaFold database，AFDB）更提供了涵盖从细菌到人类等多个物种的超过2亿个蛋白质结构数据，为研究者从统计角度理解蛋白质的动力学、功能与进化规律创造了前所未有的机遇[3]。2024年，AlphaFold2的主要开发者John Jumper及Demis Hassabis与蛋白质设计领域的先驱David Baker共同荣获诺贝尔化学奖，以表彰以AlphaFold为代表的AI技术在蛋白质结构预测与计算设计中的革命性影响。

合成生物学作为一门新兴交叉学科，旨在通过工程化设计生命系统实现特定功能，其中蛋白质的理性设计与功能优化是核心研究课题[4]。无论是代谢途径中的酶分子、生物传感器的受体蛋白，还是分子机器的结构单元或细胞工厂的调控元件，蛋白质设计能力直接决定了合成生物系统的性能。然而，理性设计和优化蛋白质以实现预期功能仍是当前面临的核心挑战，其关键在于深入理解蛋白质的结构、动力学特性及其进化规律。

统计物理方法为解决上述问题提供了新的方案，其在蛋白质科学研究中早已发挥重要作用[5-6]，而AI的发展进一步加速了这一领域的进步。通过将统计物理学原理与蛋白质结构大数据（其中也包含由AI预测得到的海量蛋白质结构）的分析相结合[7-9]，研究者能够超越单一蛋白质的个案研究，从整体上把握蛋白质设计的普遍原则。本文将系统阐述相关方法在解析蛋白质科学基本问题方面的潜力。这种基于数据和统计物理的研究方法正在为合成生物学领域的蛋白质设计开创新的范式，推动该领域向更精准、更可控的方向发展。

1 蛋白质的动力学与进化的对应关系：数据驱动研究

1.1 蛋白质动力学：分子机制与研究方法

蛋白质是生命体系中执行各种功能的分子机器的重要组成部分，在催化、免疫、运输、能量转化和各种生命活动的调控中都发挥着关键作用[10]。研究蛋白质的动力学特性对于理解其功能机制至关重要，因为生物体内蛋白质功能的执行往往需要依赖于构象变化[11-12]。从物理环境来看，生命体系存在于一个高度涨落的环境中。虽然蛋白质的天然态结构对应于自由能最低的高度特异性的稳定状态，但这种天然态并非固定的静态结构——生物体内的蛋白质分子始终处于动态的涨落之中，有时甚至会相对于天然态发生大尺度的构象变化[13-14]。这些丰富的动力学特性反映出蛋白质分子的柔性，体现了蛋白质作为软物质体系的典型特征[15]。蛋白质的动力学特性在各种生物学过程中扮演着关键角色，例如：酶催化过程中的底物结合和产物释放、信号蛋白响应环境变化时的构象转变、离子通道蛋白的开关功能等等[9, 16]。因此，揭示蛋白质动力学的基本规律，不仅有助于深入理解生命过程，也为蛋白质的功能设计和调控提供理论基础。

传统的蛋白质动力学研究主要通过实验测量和理论计算两种方法开展。实验方面，核磁共振（nuclear magnetic resonance，NMR）可以探测蛋白质在溶液中的动态变化，但其时间尺度和样品适用性有限，并非所有动力学过程都能测量[17-18]；X射线晶体学和冷冻电镜虽然能够提供高精度的静态结构信息，但直接获取动力学信息的能力有限[19-20]。时间分辨率冷冻电镜（time-resolved EM，trEM）在一定程度上提高了观察的精度，但仍受到时间分辨率、样品异质性和数据处理等因素的限制，特别是对非平衡态的研究[21-22]。此外，质谱技术也为蛋白质动力学的研究提供了重要补充[23-24]。尽管这些实验方法在蛋白质动力学的研究方面提供了强有力的工具，但它们通常只能捕捉动力学过程的某一方面，难以同时兼顾时间与空间分辨率。

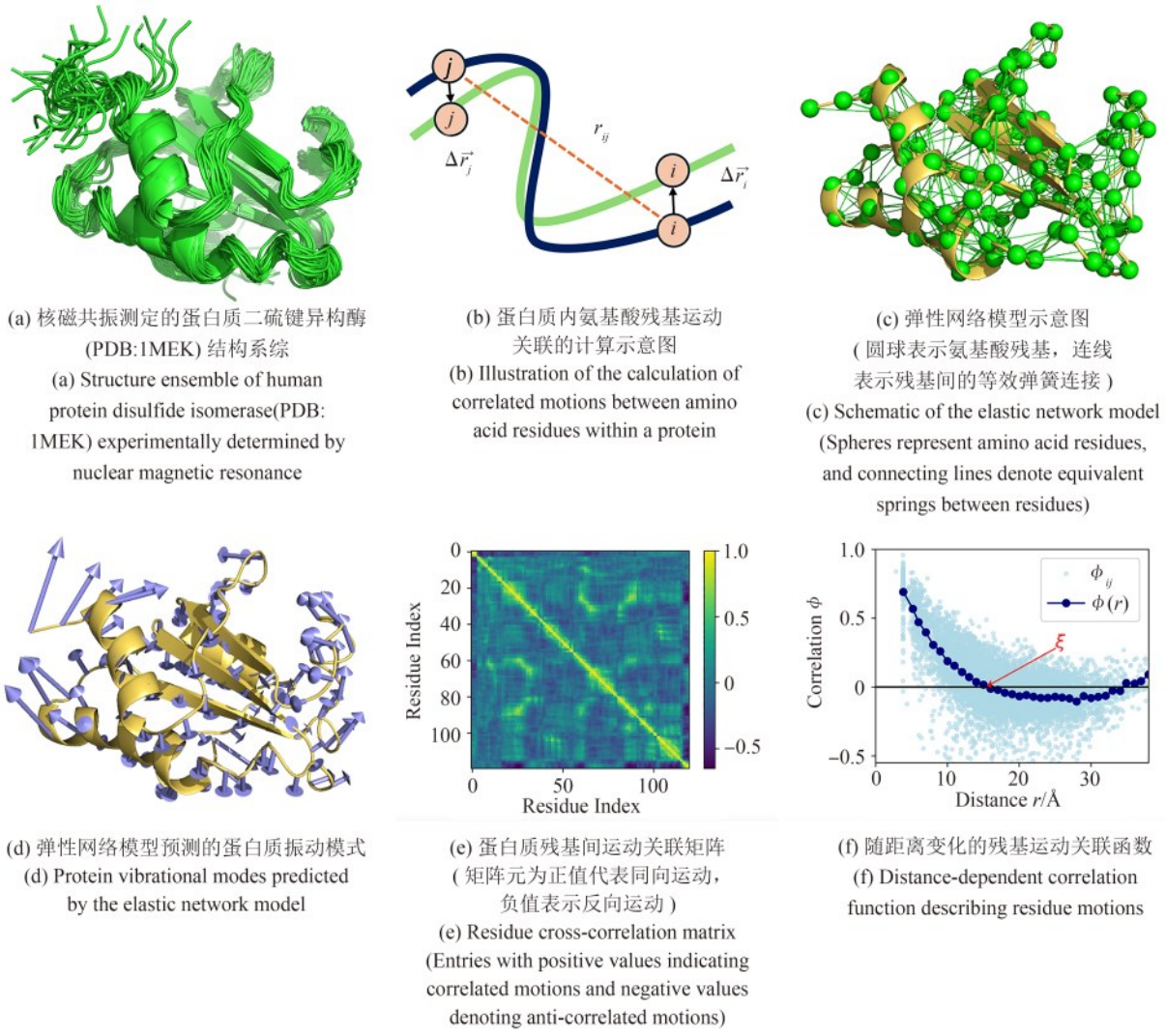
在理论计算方面，分子动力学模拟作为一种重要的研究手段，能够在原子尺度上描述蛋白质的运动细节[11]，为全面揭示蛋白质动力学的普遍规律提供了重要途径。然而，受限于计算资源，分子动力学模拟难以实现对大量不同蛋白质的系统研究。随着结构生物学和生物信息学的发展，

数据驱动的统计研究方法为理解蛋白质动力学提供了新的视角[5-6]。这种方法不再局限于研究具体的蛋白质，而是致力于发现蛋白质动力学的普适性质。在物理学研究中，普适性具有特殊的重要性：不同体系可能表现出相似的宏观行为，这种普适性往往反映了更深层的物理规律。

1.2 蛋白质动力学中的长程关联与临界性

随着大规模蛋白质结构数据的积累，基于统计物理学的研究方法揭示了蛋白质动力学的普适规律，推动蛋白质动力学研究进入定量化、高通量、系统化的新阶段。在蛋白质动力学的刻画中，氨基酸残基的关联运动（correlated motion）描述了残基之间通过协同运动实现功能调控的物理机制。这种关联不仅体现在局域结构的热涨落（如α螺旋的伸缩或β折叠的扭曲）引发的构象变化，而且局域扰动可通过“长程关联”（long-range correlations）传递到空间远端的残基，驱动蛋白质整体构象发生重排。这种跨越空间尺度的动态协同性，是蛋白质执行催化、变构调控、分子识别等多样化生物功能的物理基础[14, 25]。

下面具体介绍氨基酸残基间关联运动的定量刻画。图1(a)展示了由NMR测定的结构系综（即NMR测定的蛋白质在天然态结构附近的构象集合）。如图1(b)所示，以Cα原子坐标代表残基的



(a) 核磁共振测定的蛋白质二硫键异构酶(PDB:1MEK)结构系综
(a) Structure ensemble of human protein disulfide isomerase(PDB: 1MEK) experimentally determined by nuclear magnetic resonance

(b) 蛋白质内氨基酸残基运动关联的计算示意图
(b) Illustration of the calculation of correlated motions between amino acid residues within a protein

(c) 弹性网络模型示意图（圆球表示氨基酸残基，连线表示残基间的等效弹簧连接）
(c) Schematic of the elastic network model (Spheres represent amino acid residues, and connecting lines denote equivalent springs between residues)

(d) 弹性网络模型预测的蛋白质振动模式
(d) Protein vibrational modes predicted by the elastic network model

(e) 蛋白质残基间运动关联矩阵（矩阵元为正值代表同向运动，负值表示反向运动）
(e) Residue cross-correlation matrix (Entries with positive values indicating correlated motions and negative values denoting anti-correlated motions)

(f) 随距离变化的残基运动关联函数
(f) Distance-dependent correlation function describing residue motions

图1 蛋白质动力学分析方法示意图

Fig. 1 Methods and characterizations in protein dynamics analysis

坐标，将其中第i个残基的坐标涨落记为$\Delta\vec{r}_i$（$\vec{r}_i$是相对于平均坐标的偏移量），则i与j两个残基的协同运动的方向交叉相关可以计算为：

$$\phi_{ij}=\frac{\Delta\vec{r}_i\cdot\Delta\vec{r}_j}{|\Delta\vec{r}_i||\Delta\vec{r}_j|} \quad (1)$$

式中，$|\Delta\vec{r}_i|$代表涨落$\Delta\vec{r}_i$的大小。

本文重点讨论该方向相关性（即运动方向关联性），而运动幅度相关性（主要取决于残基定位差异，如表面与内部残基的差异）不在讨论范围内。除了基于NMR构象系综的统计分析以外，也可以将蛋白质的X射线晶体学结构表示为弹性网络，如图1(c)所示，在弹性网络模型中，氨基酸残基被描述为一系列的节点，以相应残基的Cα的坐标表示，当两个节点的距离小于给定的截断距离时，这两个节点被视为以弹簧连接（节点间存在接触），弹簧的弹性系数可以取为固定的常数或者根据残基之间的相互作用强度来选取[26-28]。对此弹性网络模型进行振动模式分析，便可以预测蛋白质在天然态附近的动力学，如图1(d)所示，图中箭头反映的是特定振动模式下氨基酸残基的运动方向。

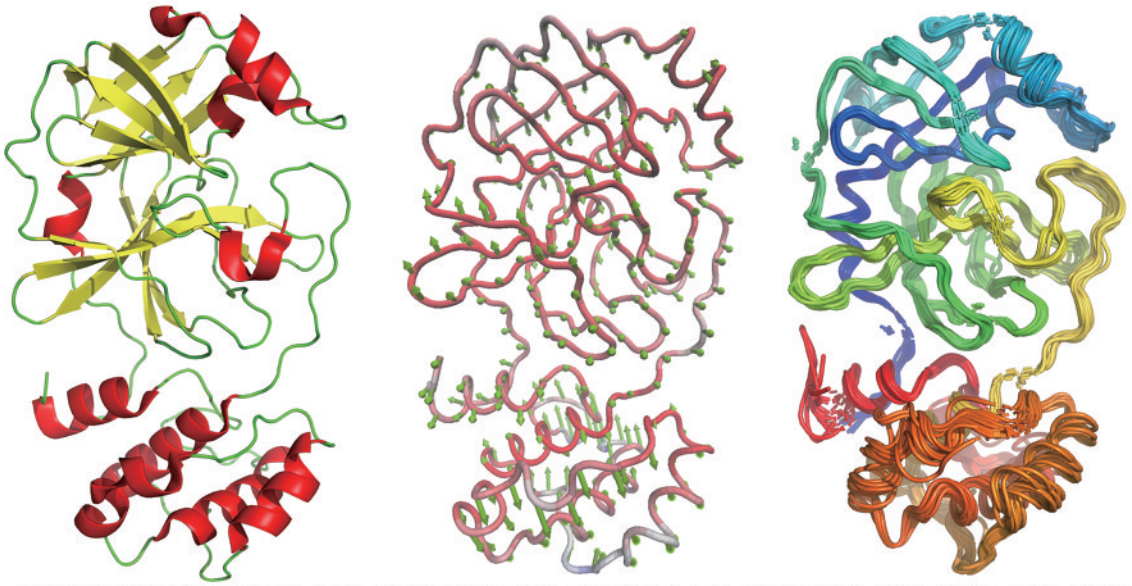
NMR研究与弹性网络模型的预测结果均可以观察到残基运动中相似的模式[29]。将如图1(e)所示的残基运动交叉关联矩阵的矩阵元ϕ_{ij}按照对应的残基对之间的距离r_{ij}分组后进行平均操作，可以得到随距离变化的交叉关联函数$\phi(r)$，使得$\phi_{ij}\approx\phi(r_{ij})$。关联函数$\phi(r)$的规律如图1(f)所示，随着两个残基之间距离的增加，残基之间的运动首先呈现随距离衰减的正相关性（同向运动），当残基之间的距离达到某一特定的长度（记作关联长度ξ）时，运动关联会经过一个零点（此时两残基的运动呈现统计意义上的无相关性），而随距离继续递增，残基的运动关联呈现反相关性（反向运动）。统计不同蛋白质的天然态涨落可以发现，涨落的关联长度与蛋白质分子的尺寸成正比，这正是长程关联的体现，也是有限尺寸物理体系处于临界态的典型特征。此外，长程关联所揭示的非局域性以及蛋白质振动谱中的幂律分布[30-32]也都是蛋白质处于临界态的有力证据[6, 33]。

蛋白质长程关联的形成是分子尺度自然选择的体现。通过分析两个互补的结构参数：形状因子（基于结构分析，表征蛋白质整体形状偏离密堆椭球形的程度）[33]与结构模块度（基于网络拓扑分析，量化网络结构被分割为模块的难易程度）[34]，我们发现，统计分布中最高频率出现的形状因子或结构模块度，与蛋白质在动力学达到最高敏感性（susceptibility）的最优形状或模块度恰好是一致的[33]，同时，天然态蛋白质结构与三维空间中的密堆结构或随机几何图不同，会表现出更低的分形维度[29, 31]。这表明进化压力普遍选择具有特定构型的蛋白质，在结构稳定性与功能必需的柔性间达到最优平衡。该平衡与蛋白质尺寸密切相关：短链蛋白倾向于形成紧凑的球状结构，而长链蛋白往往形成多结构域以实现构象柔性[29, 33]。这项研究为蛋白质结构与动力学之间的关系提供了一个新的视角，也为揭示不同尺寸蛋白质的进化原理提供了理论依据。

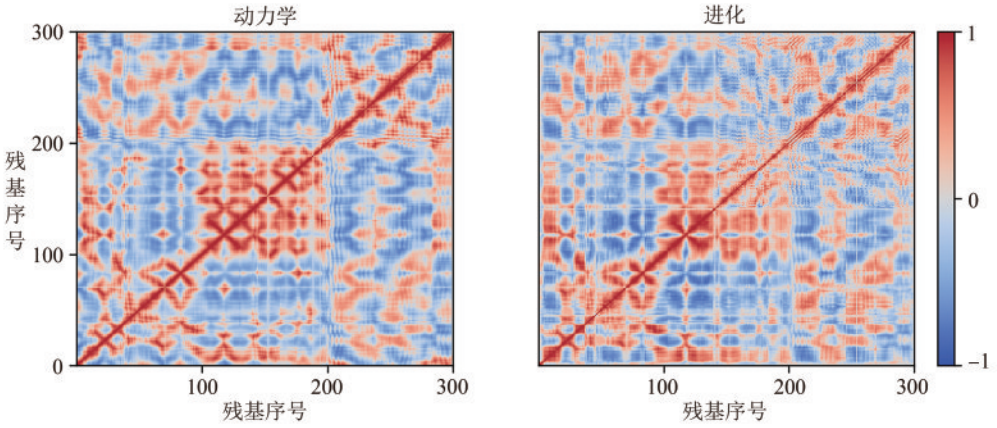
从统计物理学的角度来看，蛋白质天然态结构同时满足稳定性与高敏感性，这反映了蛋白质“有序”与“无序”之间的微妙平衡。直观而言，稳定性和敏感性这两个需求是相互制约的：增强天然结构的稳定性会抑制体系的涨落，而维持系统的敏感性又可能削弱特定天然结构的稳定性。这种矛盾在进化过程中通过自然选择得到了巧妙解决。天然态蛋白质体系表现出类似于物理学中“相变临界点”的特征，一方面，天然态结构能够保持足够的稳定性以确保分子的结构完整性与功能特异性；另一方面，蛋白质可以保持适度的构象柔性，以实现对外部信号的高灵敏响应和必要的构象调控，从而精确执行复杂的生物学功能。这种双重特性恰恰是临界点的独特属性。

1.3 蛋白质动力学与进化的对应关系

蛋白质的动力学特性与进化过程之间存在着深刻的内在联系，这种联系是理解其生物功能与环境适应性的关键。进化选择了那些动力学特性能够最优支持生物功能的蛋白质序列及其所编码的结构变形，使得蛋白质动力学表现出的结构变化模式与其功能需求达到精确匹配。因此，生物体内的蛋白质展现出如图2(a)所示的两种不同时间尺度下的“变形”机制：在较短的时间尺度上，



(a) 冠状病毒蛋白酶 (PDB:1P9S) 的蛋白质天然态结构（左）、动力学模式（中）及同家族蛋白质的结构差异（右）示意图
(a) Schematic diagram of the native state structure of coronavirus main proteinase (PDB: 1P9S) (left), dynamic patterns (middle), and structural variations within the same family (right)



(b) 残基运动关联（左）与突变引起的结构变化关联的交叉相关矩阵（右）[51]
(b) Cross-correlation matrices describing the thermal fluctuations of the residues (left) and structural variations induced by random mutations (right)[51]

图2 关联分析方法在蛋白质结构动力学及突变引起的结构变化研究中的应用

Fig. 2 Application of correlation analysis in protein structural dynamics and mutation-induced structural variations

热力学涨落驱动蛋白质构象在天然态附近发生涨落，这种动力学特性与蛋白质的功能实现密切相关[35]；而在漫长的时间尺度上，氨基酸序列通过突变和自然选择实现进化，导致结构发生适应性改变[36]。尽管这两种过程分别对应着截然不同的时间尺度：前者涉及细胞内快速发生的生物化学反应，维持生命的日常活动[37]；后者则建立在随机突变和自然选择的基础上，需要经年累代的积累[38]。但越来越多的研究表明，它们具有许多相似性：蛋白质热涨落与结构突变都具有一定的低维特征[39-40]，热涨落幅度较大的残基同样是突变率较高的残基[41]，两种来源不同的结构改变会引起相似的力学性质改变等[42]。一种经典观点认为：蛋白质的三维结构比其氨基酸序列具有更强的进化保守性，这种保守性在蛋白质的核心区域尤为显著[43-44]。数据驱动方法为揭示蛋白质动力学与进化之间的内在联系提供了新的定量分析视角[45-46]，对多种蛋白质家族的大规模统计分析还发现，蛋

白质的稳定性与进化速度之间存在显著关联，这种关联受到选择压力、蛋白质表达水平等多因素的复杂调控[47-49]。

作者利用包含数十万个蛋白质结构（来自上百个蛋白质家族）的大型数据库[50]，系统分析了蛋白质的天然态动力学与突变引起的结构变化中残基对的运动方向所出现的关联模式[51]，所采用的计算方法（包括交叉关联矩阵与关联函数）与上一节所述方法一致。对数据集中的各个蛋白质家族，分别计算其代表蛋白内数百个残基的热扰动的交叉关联矩阵（动力学）与刻画突变引起的结构变化（进化）的交叉关联矩阵［图2(b)］，结果发现，在不同蛋白质家族中，蛋白质残基在动力学中与在突变中出现的结构变化关联呈现高度相似性，从而揭示了二者间的深层对应关系。这种相似性具体表现在以下几个方面：首先，在同一家族的不同蛋白质结构中，对热扰动敏感、易发生较大涨落的残基，也更易因序列突变而发生显著结构变化；其次，描述蛋白质动力学的关联矩阵与描述蛋白质进化的关联矩阵的特征值分布与特征向量均高度相似；此外，动力学中的随距离变化的交叉关联函数$\phi(r)$也表现出相似性，均呈现显著的长程关联模式[52]。这种进化中的长程关联机制是对前文提到的蛋白质动力学中的长程关联的重要补充，具有重要的生物学意义：其一，残基的局域扰动可以对远端残基产生影响，促进蛋白质的大尺度构象变化；其二，单个位点的突变可导致远端位点的协同变化；其三，上述两种变化之间存在定量对应关系[53]。这些发现为理解蛋白质功能动力学以及分析其进化约束条件提供了统一的理论框架。

尽管蛋白质的动力学和进化都涉及高维自由度（数百甚至上千个残基的涨落或突变），但主成分分析等数据驱动的降维方法，可将复杂的高维动力学约化到低维流形。这种描述揭示了蛋白质体系的“准低维”特征：虽然系统具有极高的自由度，但功能相关的构象变化被限制在由主成分分析确定的低维流形上[54-55]，这使得蛋白质能够在保持结构稳定性的同时，通过有限维度的构象涨落高效执行复杂生物功能[56]。

在低维描述框架下，数据驱动研究进一步证实了蛋白质动力学与进化的高度吻合[51, 57]。一方面，蛋白质天然态结构对热扰动表现出整体稳定性，同时保留特定敏感的运动方向；这些特定方向对应于动力学主成分，是功能实现的关键[27-28, 58]。另一方面，在蛋白质序列发生突变时，蛋白质承担生物功能的运动方向仍能够保持相对稳定，避免功能剧变导致的适应度降低[59-60]。这种“准低维”特征的限制既保证了蛋白质功能的稳定性，又扩充了对应于特定结构的序列空间，提高了蛋白质的可设计性[61-62]。

综上，数据驱动方法通过整合大规模蛋白质结构数据与统计物理模型，系统揭示了蛋白质动力学与进化间的内在联系。基于弹性网络分析与主成分降维的研究表明，天然态蛋白质普遍具有临界态特性。统计物理框架下的关联分析进一步证明，局域热涨落与远端残基的协同进化均受相同物理规律支配。这些发现不仅为理解蛋白质的动力学、功能与进化提供了统一理论模型，更为人工蛋白质设计确立了普适性的准则：通过模拟临界态的长程关联模式，可为优化人工设计蛋白的构象调控能力与进化适应性提供有力的指导，从而突破传统设计的思路限制。

2 从蛋白质的结构预测到基于结构预测的统计分析

2.1 蛋白质结构预测概述

结构生物信息学的传统研究主要依赖实验解析的蛋白质结构数据库（如PDB）[63]，然而由于实验方法的高成本与低通量特性，相关研究极大程度受到结构数据库的规模制约。截至2024年，PDB数据库中通过实验解析得出的蛋白质结构仅约23万个[63]，而UniProt数据库已收录超过2.46亿条蛋白质序列[64]，这种显著的鸿沟促使科学家们开发更高效的方法以基于序列信息预测蛋白质结构。而深度学习技术引发的蛋白质结构预测革命，特别是AlphaFold2（AF2）等AI工具的涌现，通过基于氨基酸序列实现原子级精度的结构预测，重构了该领域的数据生态与研究范式[2-3]。当前，主流的蛋白质结构预测方法可分为三大范式：

其一为同源建模方法：基于“相似序列编码相似结构”的基本原理，利用已知结构的同源蛋白作为模板预测目标蛋白结构[65]。该方法计算效率高，对具有同源模板的蛋白质（序列相似性>30%）预测较准确[66]，但缺乏适合模板时性能显著下降。典型工具包括MODELLER和SWISS-MODEL[67-68]。

其二为从头建模（*ab initio*）方法：该方法模拟蛋白质折叠的物理过程，通过搜索构象空间中能量最低状态预测结构，早期方法采用晶格模型，用能量函数关联序列与折叠[69-70]，如代表性方法Rosetta[71-72]。该方法核心在于构建准确的势能函数以描述残基间相互作用，以及开发高效的构象空间采样算法。尽管从头建模适用范围广，但蛋白质构象空间巨大的自由度导致计算成本非常高。

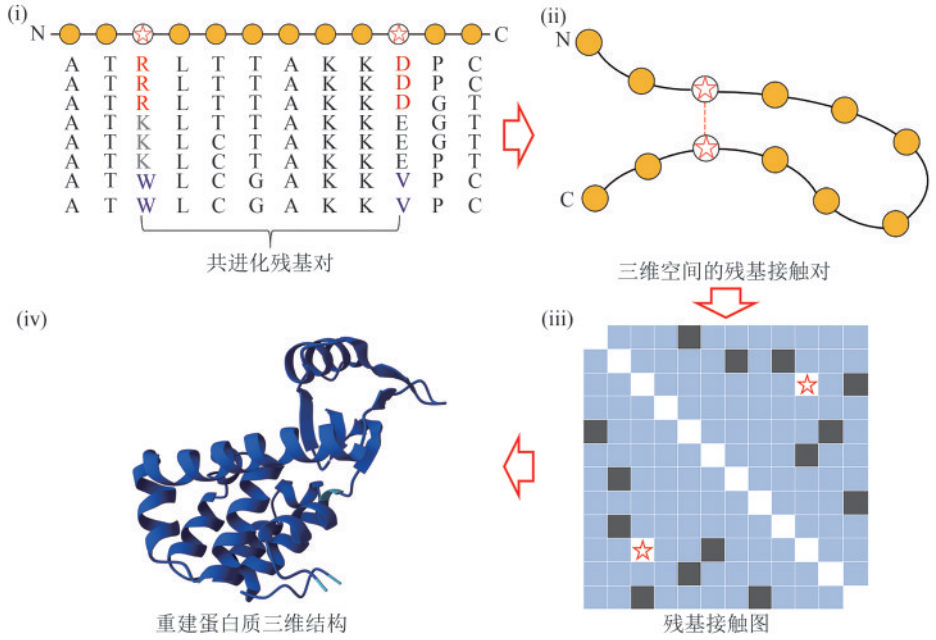
其三为机器学习方法：基于机器学习的预测方法在近年取得了突破性进展。例如，AlphaFold2将Transformer架构与进化信息相结合，整合了注意力机制（用于处理序列中的长程相互作用）、几何约束（用于引导构象搜索）以及多尺度预测（从全局折叠到原子级定位）等创新方法，从而实现了原子级精度的蛋白质结构预测，对大多数单体蛋白质的预测结果达到了接近实验解析的精度（TM-score>0.9）[2, 73-74]，在本文2.2节中，将更具体地介绍其工作原理。此外，RoseTTAFold引入三轨交互网络实现了多层次结构信息的并行处理，显著提高了计算效率[75-76]，而ESMFold基于蛋白质语言模型，通过大规模序列预训练实现了高精度预测，特别适用于宏基因组数据的快速分析，为未知蛋白质的功能解析提供了高效工具[77]。机器学习方法为蛋白质科学带来革命性变革，使对海量蛋白质开展结构生物信息学研究成为可能[78]。

2.2 AlphaFold系列模型的工作原理

作为蛋白质结构预测领域的里程碑式突破，AlphaFold系列模型的成功源于其对进化信息与深度学习技术的创新性整合。在深入探讨AlphaFold的工作原理之前，有必要回顾其理论基础——基于共进化分析的结构预测方法。共进化分析源于对蛋白质序列中协同突变模式的观察：序列中协同突变的位点在空间上通常紧密接触且存在物理相互作用。如图3（a）所示，当一个残基位点发生突变时，与之在三维空间接近的其他位点往往在进化选择的压力之下需要发生补偿性突变，从而维持蛋白质结构与功能稳定性。这种现象为从序列信息推断结构信息提供了可能性。然而，两个位点的相关性可能源于它们与第三位点的相互作用，而非彼此间的直接物理接触，为解决这一问题，研究者开发了直接耦合分析（direct coupling analysis，DCA）方法[53, 79]。DCA通过统计物理方法分离直接相互作用和间接相互作用，通过同源序列中各位点残基的联合分布推断位点间的直接耦合强度。该方法首先基于多序列比对（multiple sequence alignment， MSA）统计每个位点的氨基酸残基出现频率以及位点之间的残基对组合频率，然后通过最大熵原理构建能重现这些观测频率的最简统计模型，模型中的直接耦合参数反映了残基在三维结构中形成接触的可能性。

AlphaFold系列模型展现了共进化信息提取方式的不断深化。第一代AlphaFold采用卷积神经网络处理DCA衍生的特征，在CASP13中表现出色[80]。而AF2的核心突破在于设计了一种专为结构预测问题优化的Transformer架构——Evoformer，直接处理原始多序列比对MSA数据，实现了MSA表示（提取共进化信息）与结构表示（提取残基配对信息）的协同优化。AF2的基本工作原理如图3（b）所示，这种架构突破了传统特征分离处理的局限性，使模型可以在利用自注意力机制的长程建模能力学习MSA中的共进化模式的同时，确保残基间的距离和方向建模符合局部立体化学规则。同时，AF2通过3次循环迭代逐步修正预测结构，实现精度累积提升；它还采用了自蒸馏技术，利用高置信度预测结果构建增强训练集，提升对低同源度序列的泛化能力[2]。

2024年发布的AlphaFold 3（AF3）基于扩散模型架构，实现了生物分子复合物的全原子联合预测，进一步拓展了蛋白质结构预测的应用边界[81]。AF3采用三阶段架构：输入模块新增构象生成功能；Pairformer模块减少对MSA的依赖，并通过跨蒸馏训练增强泛化能力；扩散模块以原子级表征替代AF2中的几何约束。相较于AF2的确定性预测框架，AF3通过扩散过程的逐步迭代去噪生成结构，显著提升了柔性区域的建模能力，在蛋白质-配体

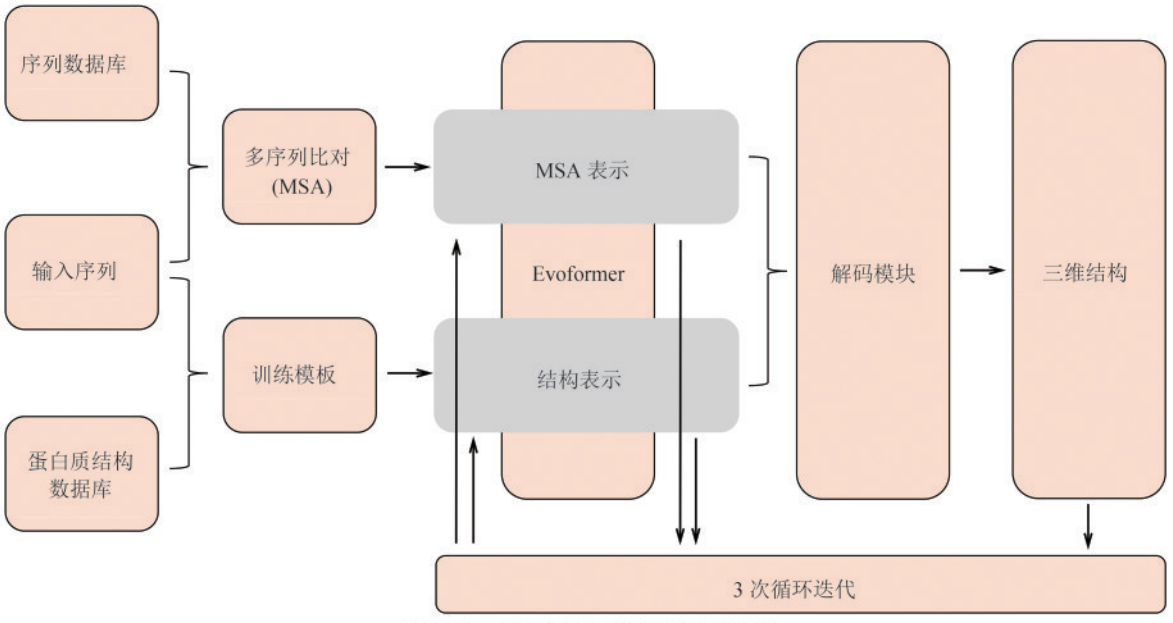


(a) 基于共进化数据预测蛋白质残基接触的基本原理示意图

(i) 多序列比对揭示协同进化残基对；(ii) 残基共进化模式对应于残基在空间上的邻近关系；(iii) 通过统计建模将共进化强度量化为接触概率，构建残基接触图；(iv) AlphaFold2 通过几何优化将接触网络解码为三维结构

(a) Coevolution-to-structure prediction pipeline

(i) MSA reveals co-evolving residue pairs; (ii) Coevolution patterns imply spatial proximity; (iii) Statistical modeling quantifies coevolution strength as contact probabilities, constructing the contact map of the protein; (iv) AlphaFold2 reconstruct 3D structure through geometric optimization



(b) AlphaFold2 的基本工作原理示意图 [2]

(b) Schematic diagram of the basic working principles of AlphaFold2[2]

图 3　基于共进化的残基接触预测与 AlphaFold2 蛋白质结构预测模型架构示意图

Fig. 3　Schematic illustration of the coevolution-based residue contact prediction and model architecture of AlphaFold2 for protein structure prediction

结合、抗原-抗体复合物结构的预测中表现出显著优势。这一模型标志着AI驱动的结构预测从单一蛋白质向多分子互作网络的范式转变，对药物设计和蛋白质工程具有直接的应用价值。

尽管AF等机器学习方法在蛋白质结构预测领域取得了革命性进展，但仍面临若干关键挑战[82-83]。例如，在多结构域蛋白预测方面，由于结构域间仅存在弱进化关联（表现为MSA冗余度低），导致预测结构容易出现较大的偏差，此外，长链蛋白质结构预测中的注意力计算存在计算瓶颈[84]，这些问题可以利用分治式组装策略与基于扩散模型的混合架构来提升预测精度[85-86]。对于内禀无序结构的采样挑战，目前主要采用融合PAE/pLDDT置信度指标与分子动力学模拟的方法来捕获其动态构象系综[87]。在蛋白质复合物结构预测领域，AlphaFold-Multimer[88]通过改进AF2损失函数与训练流程提升了预测精度，此外，基于Monte-Carlo树搜索[89]和扩散模型[90]的新方法也在不断优化预测性能。此外，针对模型的可解释性问题，结合高通量实验与力学模型分析可显著提升AF2对突变诱导的力学响应变化的预测能力[91-92]。综上所述，要突破这些挑战，需要构建“生成式AI+物理约束+实验验证”的协同创新体系，全面推动蛋白质结构预测从静态结构解析向动态功能研究的新范式转变。

2.3 基于AlphaFold数据库的统计研究

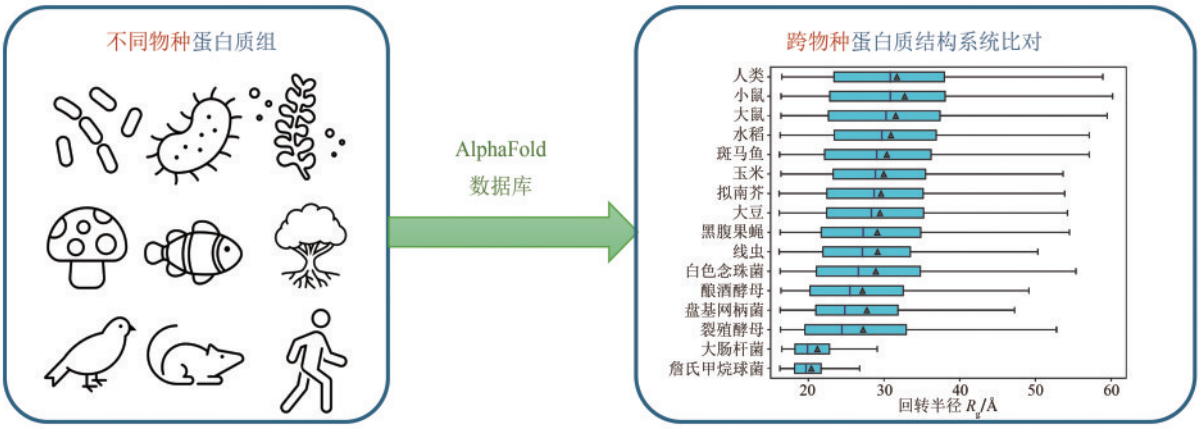
AFDB的建立是结构生物信息学领域的一个重要进展[4]。这一规模空前的蛋白质结构数据库为从统计角度系统性理解蛋白质结构、功能与进化提供了坚实基础，并促进了多种创新性研究方法的发展。

在结构与功能方面，研究者以结构预测结果为桥梁，将海量序列与功能联系起来，从而提高了“从序列到功能”的注释精度[93]。例如，基于结构比对的FoldSeek工具通过对蛋白质结构域进行聚类分析，系统性地组织和分类了AFDB中的结构[94-96]。研究者利用AFDB中的结构数据，揭示了全新的β-花折叠（β-flower fold）等此前未知的折叠结构，并将多个未知序列家族归类到已知结构超家族中[97]；同时，针对内禀无序蛋白或无序片段，研究者也得以通过AFDB的数据对传统方法难以定量刻画的结构特征进行分类与统计[98]。这些成果表明，通过数据驱动的手段挖掘统计规律，不仅增进了我们对“蛋白质宇宙”中“暗物质”的了解，还能帮助识别全新模式或功能元件，为合成生物学设计提供更为丰富的元件库[99]。

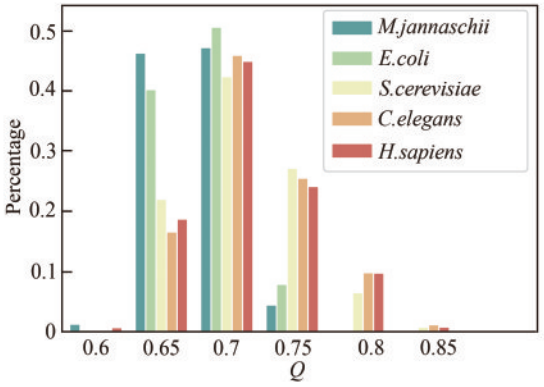
而在进化研究方面，AFDB也带来了全新的研究视角。如图4(a)所示，传统的蛋白质进化研究通常聚焦于单一蛋白质，进行跨物种或家族内的序列或结构分析。而随着AFDB的出现，研究者能够将研究范围扩展到大规模蛋白质集合，甚至可以直接对比不同物种的蛋白质组，从中挖掘统计规律。这种从单一家族到跨物种、从局部到全局的研究视角扩展，为理解蛋白质进化提供了更全面的框架。尽管AF2等AI蛋白质结构预测工具在个体蛋白质的结构预测中仍存在一定误差，但这些结果在统计研究和趋势挖掘方面依然非常可靠，因为统计规律往往不依赖于每个蛋白质结构预测的绝对准确性。基于AFDB，我们对40多种生物的蛋白质组进行了结构对比研究，揭示了蛋白质进化中的统计规律[100]。随着物种复杂性的提高，如图4(b)所示，物种体内的蛋白质在结构方面趋向于具有更高的柔性和模块化程度；在序列方面表现出更显著的亲疏水片段分隔；同时，蛋白质的功能专能性也不断提高。这些基于AFDB的统计研究成果在分子进化和物种进化之间建立了重要联系，为理解生物复杂性的起源与进化提供了新的视角。

此外，在物种复杂度的分析中，研究者也尝试引入基于AFDB的蛋白质动力学视角，即利用弹性网络模型对不同物种中链长相似的蛋白质进行了对比分析。结果显示，随着物种复杂性的提升，蛋白质天然态运动的主成分比例发生显著变化。例如，在大肠杆菌中，蛋白质运动的第1主成分与第2主成分的相对大小较为接近，而在更复杂的生物体内(例如人体中)，两者之间的差距显著增大。

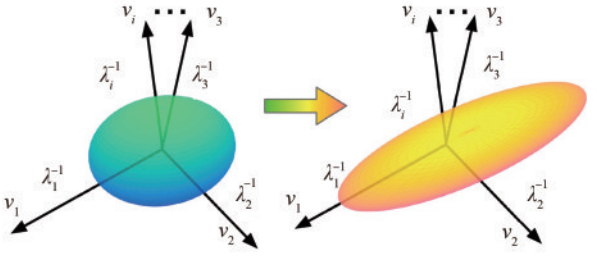
进一步的统计分析表明，从简单到复杂的物种演化过程中，物种体内的蛋白质动力学也呈现出“降维”趋势，增强了蛋白质动力学中的准低维特征，使蛋白质特定的功能运动模式更加突出。这种“进化降维”现象在其他理论生物学研究中也有类似报道[101-103]，如图4(c)所示，在复杂性更高的生物体



(a) AlphaFold 数据库为跨物种蛋白质组的结构与功能比较研究提供了系统化平台
(跨物种蛋白质结构对比分析结果显示，在链长相近时，蛋白质的回转半径随物种复杂度增加而增大)
(a) The AlphaFold database provides a platform for systematic comparative studies of structure and function across proteomes of different organisms.
(The analysis reveals that proteins with similar chain lengths exhibit a positive correlation between their radius of gyration and the increase in organismal complexity)



(b) 不同物种中相似链长的蛋白质的模块度分布对比
(b) Modularity distribution of proteins with similar chain lengths across different organisms



(c) 蛋白质天然态动力学随物种复杂度增加（箭头方向）呈现降维趋势：其涨落模式（即结构热运动的主导振动方向）从各向同性椭球向各向异性椭球演化，主导运动模式逐渐凸显
(c) Protein native dynamics exhibits dimensionality reduction with increasing organismal complexity (arrow): Fluctuation modes (principal components in vibrations) evolve from isotropic to anisotropic ellipsoids, with dominant modes becoming more pronounced

图 4　基于 AlphaFold 数据库研究不同复杂度物种体内蛋白质结构与动力学的统计规律 [100]

Fig. 4　Statistical trends in protein structure and dynamics across organisms of varying complexity: an analysis based on the AlphaFold database [100]

内，更多蛋白质倾向于沿着与特定功能相关的主成分方向运动，其蛋白质随进化呈现出从“通用”到“专用”的统计趋势，即高复杂性生物体内更可能出现高度功能专业化的蛋白质 [100]。

这种“专业化”的进化趋势与生物体基因组规模的扩张密切相关。较简单生物体虽然基因组较小，酶的种类有限，但其高混杂性的酶能够支持基本的生命活动。相比之下，复杂生物体拥有更大的基因组，能够编码更多样化的蛋白质，使其能够执行高度专业化的功能，从而更好地适应复杂的细胞环境，提高了对多样化外部环境的可塑性和适应能力。然而，需要强调的是，上述规律虽然具有普适性，但本质上是统计性的，在针对特定蛋白质进行定向进化和设计时，仍需具体问题具体分析。这些基于 AFDB 揭示的进化特征，充分体现了统计物理方法在解析蛋白质功能进化规律中的关键作用，为我们理解生命进化的分子机制提供了全新的视角。

3 基于人工智能的蛋白质动力学预测

3.1 从静态结构到蛋白质动力学预测

尽管AF2在预测蛋白质天然态三维静态结构方面表现出色，但蛋白质功能的实现通常依赖于多种构象状态间的动态转换[104]。对构象变化与动力学信息的预测已经成为当前AI结构预测领域的重要挑战。传统意义上，蛋白质动力学信息主要依赖于特定案例的实验观察或分子模拟，而AI结构预测的发展为这一领域带来了全新的研究思路。在药物设计领域，特别是针对变构蛋白的药物开发中，准确预测构象变化有助于设计更有效的药物分子。在蛋白质工程方面，对构象变化的深入理解有助于优化蛋白质的功能和稳定性[105]。

为应对蛋白动力学预测这一挑战，研究者们提出了两种主要策略。第一种是AI模型的架构调整，通过重构模型架构、调整参数和优化训练方法，直接提升AI算法的动力学预测能力；第二种是“提示词调整”策略，保持AI模型的框架不变，通过微调输入信息（如MSA或结构数据库）引导构象预测[106-112]。后者因其训练成本低、操作简单且能建立序列-结构定量联系的优势，成为蛋白质理性设计的实用解决方案。

“提示词调整”的基本原理在于，MSA提供的共进化信息是AF2预测天然结构的重要依据。蛋白质的不同构象（即残基不同的接触模式）间接反映的是不同的共进化信息。即使基于相同输入序列，只要MSA不同（共进化信息不同），AF2仍可预测出不同的构象状态。如图5所示，待测蛋白序列的完整MSA可直接通过AF2生成“标准”预测结构，也可结合物理特征对完整的MSA进行重采样，得到若干MSA子集，对这些不同的MSA子集进行多轮结构预测，可以获得代表蛋白质动态特性的多构象系综。这种思路的物理基础在于：通过筛选或重组MSA中可能相互矛盾的共进化信号，引导模型捕获特定构象的能量面特征。在技术实现上，研究者开发了多种MSA序列调控策略以增强预测多样性，例如欠采样法通过减少MSA序列数量，降低互斥的进化信号，促使AI模型预测出其他构象[106-108]；AF-Cluster通过将MSA序列聚类，将相互矛盾的共进化信息进行解耦，从而实现对已知变构蛋白不同构象状态的预测[109]。这些方法虽然实现方式不同，但它们都增强了现有AI模型对蛋白质构象多样性的预测能力[110-112]，为蛋白质构象动态的定向预测指明新的方向。

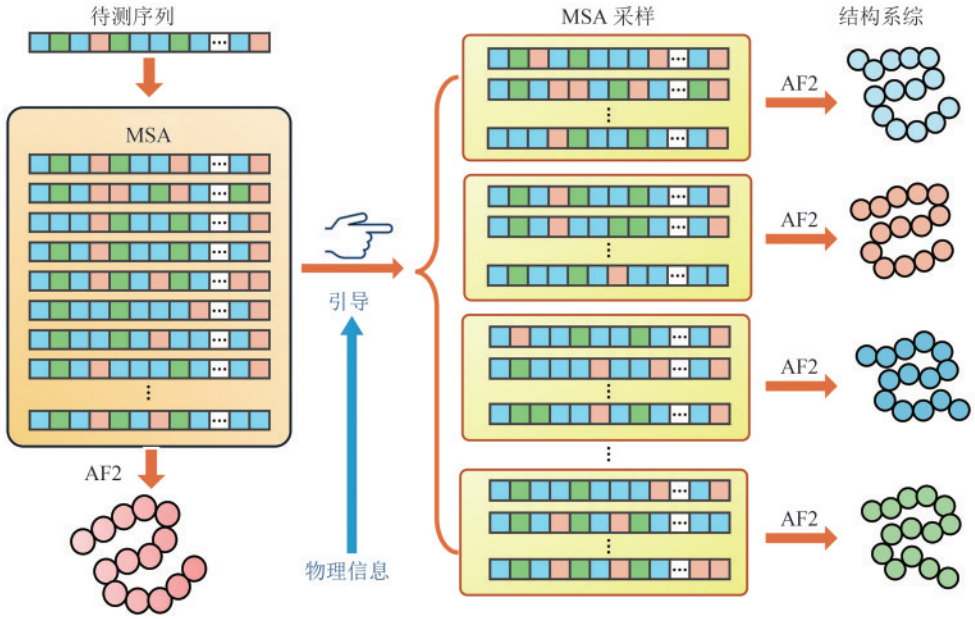


图5 基于物理学信息引导的人工智能预测蛋白质动力学方法示意图

Fig. 5 Schematic illustration of physics-prompted artificial intelligence approach for predicting protein dynamics

3.2 用物理学信息引导人工智能预测蛋白质动力学

近年来，研究者开始将物理知识引入AI结构预测框架，这一趋势源于对蛋白质能量面理论的深入理解。虽然预测动力学信息对纯数据驱动方法仍然是一大挑战，但概述蛋白质折叠与动力学行为的理论规律则早已被提出：Peter Wolynes等人提出的蛋白质能量面理论指出，蛋白质折叠可被视为在漏斗状能量面内下降至能量最低点的过程[113]，而根据“最小阻挫原理”（minimal frustration principle），蛋白质序列在进化中被优化，使能量面尽可能光滑，减少局部极小值或能垒，从而帮助蛋白质更容易找到全局能量最低的天然态构象[114-115]。然而，与自旋玻璃等物理系统类似，完全消除能量面的局部阻挫在生物学上既不可能也无必要。事实上，天然态蛋白质为执行功能，需要保留一定的“局部阻挫”（local frustration）[116]。这些局部阻挫并非缺陷，而是蛋白质灵活利用动力学以实现功能的精巧策略。例如，研究表明，变构蛋白通过局部阻挫调节构象转变，实现高效的功能运动[117-120]。局部阻挫的存在体现了结构稳定性与其功能运动灵活性之间的微妙平衡。

研究表明，大尺度构象变化所涉及的关键残基位点（即残基接触发生断开的位点）往往具有更高的阻挫[118]。因此，利用现有的阻挫分析工具Frustratometer[116]识别蛋白质中局部阻挫较高的区域，可以预测构象变化中的关键位点。基于识别出的高阻挫位点，研究者们可以筛选和操纵MSA空间中包含的序列或其蕴藏的共进化信息，从而预测出天然态结构以外的其他构象，以此来获得蛋白质的变构路径细节。例如，采用序列筛选、序列混合、位点遮盖等方法筛选或屏蔽高阻挫位点的共进化信息，可以避免AF2在相关位点附近预测出特定残基接触。这种研究思路已成功应用于预测ADK、KaiB等蛋白的不同构象甚至变构路径，且预测结果与实验及分子模拟高度一致[121]。这些发现表明，基于蛋白质能量面理论，识别高阻挫位点并在MSA空间理性筛选共进化信息，不仅能够突破AF2的静态预测限制，预测蛋白的亚稳态结构，还能获得变构路径等关键信息。

此外，近期研究进展表明，结合物理知识和机器学习方法可以更全面地理解和预测蛋白质的构象变化。研究者开始引入能量函数约束，考虑氢键网络和疏水作用等分子相互作用，并结合分子动力学模拟结果来增强采样和优化预测转变态结构[122]。这种多学科融合方法显著提高了对复杂蛋白质动力学过程的预测准确性。

3.3 人工智能蛋白质动力学预测的挑战与前景

在蛋白质动力学预测领域，AI方法仍面临诸多挑战。目前主流方法包括基于深度学习的端到端预测、基于分子动力学的混合方法以及基于能量函数优化的方法。尽管有研究表明AF能够预测多个构象态，但批评者指出，这些预测主要依赖于对数据库的“记忆”，而非对蛋白质动力学本质的深入理解[123]。更深层次的问题在于，现有AI方法尚未从能量景观和玻尔兹曼分布等统计物理角度真正理解蛋白质动力学。这一局限性在处理折叠切换蛋白（Fold-Switch）和包含内禀无序片段的蛋白质时尤为明显，因为AI模型倾向于生成与已知结构相似的结果，而忽略稀有构象[124]。这表明，在缺乏先验知识的情况下，AI模型捕获未知构象的能力仍然有限。为克服这些局限性，研究者提出了多种改进策略，例如用生成模型产生虚拟的同源序列[125]，结合实验结果指导预测[126]等，旨在增强模型对蛋白质动力学与功能的理解。

近年来，扩散模型作为一种新兴生成模型，在蛋白质动力学预测领域展现出独特优势[81，127]。与传统深度学习方法不同，扩散模型通过模拟从热力学平衡到非平衡的渐进过程，能够更好地捕捉蛋白质构象转换的连续性特征[127]。具体而言，扩散模型通过逐步向蛋白质结构添加噪声，并学习去噪过程来生成构象转换路径。这种方法不仅能够自然生成连续的构象变化轨迹，还可以模拟不同时间尺度上的构象变化，在预测无序蛋白区域的构象集合方面显示出特殊优势。然而，扩散模型的应用也面临显著挑战，诸如如何将物理约束合理整合到模型中。一个可能的解决方案是将扩散模型与分子动力学模拟结合，例如使用分子动力学生成的轨迹训练扩散模型，同时引入物理能量项约束生成过程[128]。

AI蛋白质动力学预测将引领蛋白质工程进入精确设计的新时代，指导基于动力学特征的蛋白质理性设计，通过预测并操控构象变化路径，实现酶活性或稳定性调控、配体特异性优化和变构路径的设计[129-130]；AlphaMissense等工具进一步展示了AI结构预测在致病性分析等临床应用领域的潜力[131]；在药物发现领域，AI动力学预测将推动靶向设计方法的革新，通过分析靶蛋白的动力学行为，指导设计更特异有效的药物分子，实现精准药效调控[132-133]。

4 总结与展望

本综述围绕数据驱动方法如何揭示蛋白质动力学与进化之间的关联，系统梳理了相关研究进展与理论发展。AF的出现不仅推动了结构生物信息学的快速发展，更为蛋白质动力学研究提供了新视角。伴随着AI预测结构和实验数据不断涌现，越来越多的不同模态的数据也在不断扩充[134]，为大规模的数据整合与统计分析提供了可能性。在后AF时代，统计物理学与生物大数据分析的结合，使研究者能够超越个例研究，从整体上把握蛋白质设计的普遍原则[135]，启发与指导蛋白质设计任务[136-137]。结合不断扩充的宏基因组和蛋白质组数据，数据驱动的研究范式将为我们理解生物复杂系统提供更全面的视角[138]。

在方法学层面，物理学原理与AI的深度融合将催生新型数据分析方法，并推动建立多尺度计算模拟的新框架[139-140]。这种方法学创新将显著提升对复杂生物系统的理解能力，使我们能从分子水平到系统水平全面把握蛋白质功能。随着结构数据库的快速扩充，多组学数据的整合分析将成为常态，大规模数据挖掘技术将从海量数据中提取更多有价值信息，推动“从数据到知识”的转变。

在理论与应用方面，基于AFDB的统计分析揭示的进化规律和结构-功能关系，为理性设计新型蛋白质提供了坚实的理论基础。深入了解不同复杂度生物中蛋白质的动力学特性差异，可以指导研究者针对特定应用场景设计具有适当柔性、热稳定性或功能专一性的蛋白质[141-142]。此外，从数据库挖掘出的新型结构模块和功能元件，为合成生物学家提供了更丰富的“基本元件”，用于组装全新功能的人工蛋白质[143]。这些进展将推动新型生物催化剂的开发、智能生物材料的设计以及精准医疗技术的进步，加速合成生物学从经验驱动向知识驱动的范式转变，为解决能源、环境和医疗等全球挑战提供创新解决方案。

在技术层面，AI预测与实验验证将形成高效闭环系统：高通量实验数据持续优化AI模型，而AI预测（尤其是基于最新的蛋白质语言模型的预测）[144]则指导更精准的实验设计[145-146]。这一闭环将推动自动化实验平台与AI系统的深度集成，最终构建“AI驱动的蛋白质工程工厂”，能够自主完成设计-预测-合成-测试-优化的全流程循环[147]。这种整合平台已在抗体工程和酶优化中展现出显著潜力，未来有望成为合成生物学研究的核心基础设施，大幅缩短从设计概念到实际应用的时间。此外，数据驱动方法还为高通量筛选带来了变革[148]，传统高通量筛选方法通常基于静态结构或经验规则，而整合动力学信息的AI方法能显著提高筛选精度，减少假阳性结果。更重要的是，这些方法能够针对具有特定动力学特性的候选分子，缩小实验筛选空间，从而大幅节省研究资源。

综上所述，数据驱动的结构生物信息学正为合成生物学、药物设计和精准医疗带来革命性变革。通过提升高通量筛选和理性设计的效率，这些方法加速了从基础发现到实际应用的转化过程。随着计算能力的持续提升和大型AI模型的不断发展，该领域有望在不久的将来实现重大突破，推动从蛋白质设计、蛋白质复合体与大型分子机器设计到复杂生物系统构建的技术飞跃。

参 考 文 献

[1] CHOU K C. Structural bioinformatics and its impact to biomedical science[J]. Current Medicinal Chemistry, 2004, 11(16): 2105-2134.

[2] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[3] 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-D444.

[4] KORTEMME T. *De novo* protein design-from new structures to programmable functions[J]. Cell, 2024, 187(3): 526-544.

[5] HOPF T A, COLWELL L J, SHERIDAN R, et al. Three-dimensional structures of membrane proteins from genomic sequencing[J]. Cell, 2012, 149(7): 1607-1621.

[6] MORA T, BIALEK W. Are biological systems poised at criticality? [J]. Journal of Statistical Physics, 2011, 144(2): 268-302.

[7] HALILOGLU T, BAHAR I. Adaptability of protein structures to enable functional interactions and evolutionary implications [J]. Current Opinion in Structural Biology, 2015, 35: 17-23.

[8] HALABI N, RIVOIRE O, LEIBLER S, et al. Protein sectors: evolutionary units of three-dimensional structure[J]. Cell, 2009, 138(4): 774-786.

[9] NUSSINOV R, TSAI C J. Allostery in disease and in drug discovery[J]. Cell, 2013, 153(2): 293-305.

[10] ORENGO C A, TODD A E, THORNTON J M. From protein structure to function[J]. Current Opinion in Structural Biology, 1999, 9(3): 374-382.

[11] KARPLUS M, MCCAMMON J A. Molecular dynamics simulations of biomolecules[J]. Nature Structural Biology, 2002, 9(9): 646-652.

[12] FRAUENFELDER H, CHEN G, BERENDZEN J, et al. A unified model of protein dynamics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(13): 5129-5134.

[13] BOEHR D D, NUSSINOV R, WRIGHT P E. The role of dynamic conformational ensembles in biomolecular recognition [J]. Nature Chemical Biology, 2009, 5(11): 789-796.

[14] HENZLER-WILDMAN K, KERN D. Dynamic personalities of proteins[J]. Nature, 2007, 450(7172): 964-972.

[15] DE GENNES P G. Soft matter[J]. Science, 1992, 256(5056): 495-497.

[16] CHANGEUX J P, CHRISTOPOULOS A. Allosteric modulation as a unifying mechanism for receptor function and regulation[J]. Cell, 2016, 166(5): 1084-1102.

[17] KAY L E. NMR studies of protein structure and dynamics[J]. Journal of Magnetic Resonance, 2005, 173(2): 193-207.

[18] XIE T, SALEH T, ROSSI P, et al. Conformational states dynamically populated by a kinase determine its function[J]. Science, 2020, 370(6513): eabc2754.

[19] FRASER J S, VAN DEN BEDEM H, SAMELSON A J, et al. Accessing protein conformational ensembles using room-temperature X-ray crystallography[J]. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(39): 16247-16252.

[20] MERK A, BARTESAGHI A, BANERJEE S, et al. Breaking cryo-EM resolution barriers to facilitate drug discovery[J]. Cell, 2016, 165(7): 1698-1707.

[21] FRANK J. Time-resolved cryo-electron microscopy: recent progress[J]. Journal of Structural Biology, 2017, 200(3): 303-306.

[22] HARDER O F, BARRASS S V, DRABBELS M, et al. Fast viral dynamics revealed by microsecond time-resolved cryo-EM[J]. Nature Communications, 2023, 14: 5649.

[23] KALTASHOV I A, BOBST C E, ABZALIMOV R R. Mass spectrometry-based methods to study protein architecture and dynamics[J]. Protein Science, 2013, 22(5): 530-544.

[24] LENTO C, WILSON D J. Subsecond time-resolved mass spectrometry in dynamic structural biology[J]. Chemical Reviews, 2022, 122(8): 7624-7646.

[25] BENKOVIC S J, HAMMES-SCHIFFER S. A perspective on enzyme catalysis[J]. Science, 2003, 301(5637): 1196-1202.

[26] BAHAR I, LEZON T R, YANG L W, et al. Global dynamics of proteins: bridging between structure and function[J]. Annual Review of Biophysics, 2010, 39: 23-42.

[27] BAHAR I, RADER A J. Coarse-grained normal mode analysis in structural biology[J]. Current Opinion in Structural Biology, 2005, 15(5): 586-592.

[28] TIRION M M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis[J]. Physical Review Letters, 1996, 77(9): 1905-1908.

[29] TANG Q Y, KANEKO K. Long-range correlation in protein dynamics: confirmation by structural data and normal mode analysis[J]. PLoS Computational Biology, 2020, 16(2): e1007670.

[30] REUVENI S, GRANEK R, KLAFTER J. Proteins: coexistence of stability and flexibility[J]. Physical Review Letters, 2008, 100(20): 208101.

[31] TANG Q-Y, HATAKEYAMA T S, KANEKO K. Functional sensitivity and mutational robustness of proteins[J]. Physical Review Research, 2020, 2(3): 033452.

[32] HU X H, HONG L, DEAN SMITH M, et al. The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time[J]. Nature Physics, 2016, 12(2): 171-174.

[33] TANG Q Y, ZHANG Y Y, WANG J, et al. Critical fluctuations in the native state of proteins[J]. Physical Review Letters, 2017, 118(8): 088102.

[34] NEWMAN M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(23): 8577-8582.

[35] EISENMESSER E Z, MILLET O, LABEIKOVSKY W, et al. Intrinsic dynamics of an enzyme underlies catalysis[J]. Nature, 2005, 438(7064): 117-121.

[36] STEIN A, FOWLER D M, HARTMANN-PETERSEN R, et al. Biophysical and mechanistic models for disease-causing protein variants[J]. Trends in Biochemical Sciences, 2019, 44(7): 575-588.

[37] FRAUENFELDER H, SLIGAR S G, WOLYNES P G. The energy landscapes and motions of proteins[J]. Science, 1991, 254(5038): 1598-1603.

[38] ZUCKERKANDL E, PAULING L. Evolutionary divergence

and convergence in proteins[J]. Evolving Genes and Proteins, 1965: 97-166.

[39] TAMA F, SANEJOUAND Y H. Conformational change of proteins arising from normal mode calculations[J]. Protein Engineering, 2001, 14(1): 1-6.

[40] FACCO E, PAGNANI A, RUSSO E T, et al. The intrinsic dimension of protein sequence evolution[J]. PLoS Computational Biology, 2019, 15(4): e1006767.

[41] LIU Y, BAHAR I. Sequence evolution correlates with structural dynamics[J]. Molecular Biology and Evolution, 2012, 29(9): 2253-2263.

[42] TOKURIKI N, TAWFIK D S. Protein dynamism and evolvability [J]. Science, 2009, 324(5924): 203-207.

[43] ILLERGÅRD K, ARDELL D H, ELOFSSON A. Structure is three to ten times more conserved than sequence: a study of structural response in protein cores[J]. Proteins: Structure, Function, and Bioinformatics, 2009, 77(3): 499-508.

[44] WORTH C L, GONG S, BLUNDELL T L. Structural and functional constraints in the evolution of protein families[J]. Nature Reviews Molecular Cell Biology, 2009, 10(10): 709-720.

[45] LIBERLES D A, TEICHMANN S A, BAHAR I, et al. The interface of protein structure, protein biophysics, and molecular evolution[J]. Protein Science, 2012, 21(6): 769-785.

[46] ECHAVE J, WILKE C O. Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence[J]. Annual Review of Biophysics, 2017, 46: 85-103.

[47] BLOOM J D, LABTHAVIKUL S T, OTEY C R, et al. Protein stability promotes evolvability[J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(15): 5869-5874.

[48] SEROHIJOS A W R, SHAKHNOVICH E I. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics [J]. Current Opinion in Structural Biology, 2014, 26: 84-91.

[49] DRUMMOND D A, WILKE C O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution [J]. Cell, 2008, 134(2): 341-352.

[50] HOLM L. Dali server: structural unification of protein families [J]. Nucleic Acids Research, 2022, 50(W1): W210-W215.

[51] TANG Q Y, KANEKO K. Dynamics-evolution correspondence in protein structures[J]. Physical Review Letters, 2021, 127(9): 098103.

[52] ECHAVE J, SPIELMAN S J, WILKE C O. Causes of evolutionary rate variation among protein sites[J]. Nature Reviews Genetics, 2016, 17(2): 109-121.

[53] MORCOS F, PAGNANI A, LUNT B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families[J]. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(49): E1293-E1301.

[54] HENZLER-WILDMAN K A, LEI M, THAI V, et al. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis[J]. Nature, 2007, 450(7171): 913-916.

[55] GRANT B J, GORFE A A, MCCAMMON J A. Large conformational changes in proteins: signaling and other functions [J]. Current Opinion in Structural Biology, 2010, 20(2): 142-147.

[56] NUSSINOV R, TSAI C J, LIU J. Principles of allosteric interactions in cell signaling[J]. Journal of the American Chemical Society, 2014, 136(51): 17692-17701.

[57] MARSH J A, TEICHMANN S A. parallel dynamics and evolution: protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure[J]. BioEssays, 2014, 36(2): 209-218.

[58] YANG L W, BAHAR I. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes[J]. Structure, 2005, 13(6): 893-904.

[59] MAGUID S, FERNANDEZ-ALBERTI S, ECHAVE J. Evolutionary conservation of protein vibrational dynamics[J]. Gene, 2008, 422(1-2): 7-13.

[60] TÓTH-PETRÓCZY Á, TAWFIK D S. The robustness and innovability of protein folds[J]. Current Opinion in Structural Biology, 2014, 26: 131-138.

[61] LI H, TANG C, WINGREEN N S. Nature of driving force for protein folding: a result from analyzing the statistical potential [J]. Physical Review Letters, 1997, 79(4): 765-768.

[62] ENGLAND J L, SHAKHNOVICH E I. Structural determinant of protein designability[J]. Physical Review Letters, 2003, 90(21): 218101.

[63] BURLEY S K, BHATT R, BHIKADIYA C, et al. Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank [J]. Nucleic Acids Research, 2025, 53(D1): D564-D574.

[64] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023[J]. Nucleic Acids Research, 2023, 51 (D1): D523-D531.

[65] CHOTHIA C, LESK A M. The relation between the divergence of sequence and structure in proteins[J]. The EMBO Journal, 1986, 5(4): 823-826.

[66] KRYSHTAFOVYCH A, SCHWEDE T, TOPF M, et al. Critical assessment of methods of protein structure prediction (CASP) - Round XIV[J]. Proteins: Structure, Function, and Bioinformatics, 2021, 89(12): 1607-1617.

[67] FISER A, ŠALI A. Modeller: generation and refinement of homology-based protein structure models[J]. Methods in Enzymology, 2003, 374: 461-491.

[68] 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.

[69] LAU K F, DILL K A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins[J]. Macromolecules, 1989, 22(10): 3986-3997.

[70] GO N. Theoretical studies of protein folding[J]. Annual Review of Biophysics and Bioengineering, 1983, 12: 183-210.

[71] SIMONS K T, KOOPERBERG C, HUANG E, et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions[J]. Journal of Molecular Biology, 1997, 268(1): 209-225.

[72] BRADLEY P, MISURA K M S, BAKER D. Toward high-resolution *de novo* structure prediction for small proteins[J]. Science, 2005, 309(5742): 1868-1871.

[73] MIRDITA M, SCHÜTZE K, MORIWAKI Y, et al. ColabFold: making protein folding accessible to all[J]. Nature Methods, 2022, 19(6): 679-682.

[74] AHDRITZ G, BOUATTA N, FLORISTEAN C, et al. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization[J]. Nature Methods, 2024, 21(8): 1514-1524.

[75] BAEK M, DIMAIO F, ANISHCHENKO I, et al. Accurate prediction of protein structures and interactions using a three-track neural network[J]. Science, 2021, 373(6557): 871-876.

[76] BAEK M, BAKER D. Deep learning and protein structure modeling[J]. Nature Methods, 2022, 19(1): 13-14.

[77] LIN Z M, AKIN H, RAO R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science, 2023, 379(6637): 1123-1130.

[78] TUNYASUVUNAKOOL K, ADLER J, WU Z, et al. Highly accurate protein structure prediction for the human proteome [J]. Nature, 2021, 596(7873): 590-596.

[79] WEIGT M, WHITE R A, SZURMANT H, et al. Identification of direct residue contacts in protein-protein interaction by message passing[J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(1): 67-72.

[80] SENIOR A W, EVANS R, JUMPER J, et al. Improved protein structure prediction using potentials from deep learning[J]. Nature, 2020, 577(7792): 706-710.

[81] ABRAMSON J, ADLER J, DUNGER J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3[J]. Nature, 2024, 630(8016): 493-500.

[82] WANG L, WEN Z H, LIU S W, et al. Overview of AlphaFold2 and breakthroughs in overcoming its limitations[J]. Computers in Biology and Medicine, 2024, 176: 108620.

[83] YANG Z Y, ZENG X X, ZHAO Y, et al. AlphaFold2 and its applications in the fields of biology and medicine[J]. Signal Transduction and Targeted Therapy, 2023, 8: 115.

[84] JUMPER J, EVANS R, PRITZEL A, et al. Applying and improving AlphaFold at CASP14[J]. Proteins: Structure, Function, and Bioinformatics, 2021, 89(12): 1711-1721.

[85] XIA Y H, ZHAO K L, LIU D, et al. Multi-domain and complex protein structure prediction using inter-domain interactions from deep learning[J]. Communications Biology, 2023, 6: 1221.

[86] SHOR B, SCHNEIDMAN-DUHOVNY D. CombFold: predicting structures of large protein assemblies using a combinatorial assembly algorithm and AlphaFold2[J]. Nature Methods, 2024, 21(3): 477-487.

[87] TESEI G, TROLLE A I, JONSSON N, et al. Conformational ensembles of the human intrinsically disordered proteome[J]. Nature, 2024, 626(8000): 897-904.

[88] EVANS R, O'NEILL M, PRITZEL A, et al. Protein complex prediction with AlphaFold-Multimer[EB/OL]. bioRxiv, 2021, 2021.10. 04.463034[2025-01-15]. https://doi.org/10.1101/2021.10.04.463034.

[89] BRYANT P, POZZATI G, ZHU W S, et al. Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search[J]. Nature Communications, 2022, 13: 6028.

[90] LIU S W, ZHU T, REN M L, et al. Predicting mutational effects on protein-protein binding *via* a side-chain diffusion probabilistic model[C/OL]//Advances in Neural Information Processing Systems, 2023, 36: 48994-49005[2025-01-15]. https://proceedings. neurips. cc/paper_files/paper/2023/hash/99088dffd5eab0babebcda4bc58bbcea-Abstract-Conference.html.

[91] TANG Q Y. The mechanics of protein sweet spots[J/OL]. Nature Physics, 2025. (2025-03-28) [2025-03-29]. https://doi.org/10.1038/s41567-025-02826-8.

[92] WEINREB E, MCBRIDE J M, SIEK M, et al. Enzymes as viscoelastic catalytic machines[J/OL]. Nature Physics, 2025. (2025-03-28) [2025-03-29]. https://doi.org/10.1038/s41567-025-02825-9.

[93] MA W J, ZHANG S G, LI Z, et al. Enhancing protein function prediction performance by utilizing AlphaFold-predicted protein structures[J]. Journal of Chemical Information and Modeling, 2022, 62(17): 4008-4017.

[94] VAN KEMPEN M, KIM S S, TUMESCHEIT C, et al. Fast and accurate protein structure search with Foldseek[J]. Nature Biotechnology, 2024, 42(2): 243-246.

[95] BARRIO-HERNANDEZ I, YEO J, JÄNES J, et al. Clustering predicted structures at the scale of the known protein universe [J]. Nature, 2023, 622(7983): 637-645.

[96] KIM W S, MIRDITA M, LEVY KARIN E, et al. Rapid and sensitive protein complex alignment with Foldseek-Multimer [J]. Nature Methods, 2025, 22(3): 469-472.

[97] DURAIRAJ J, WATERHOUSE A M, METS T, et al. Uncovering new families and folds in the natural protein universe[J]. Nature, 2023, 622(7983): 646-653.

[98] ALDERSON T R, PRITIŠANAC I, KOLARIĆ Đ, et al. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2[J]. Proceedings of the National

Academy of Sciences of the United States of America, 2023, 120(44): e2304302120.

[99] THORNTON J M, LASKOWSKI R A, BORKAKOTI N. AlphaFold heralds a data-driven revolution in biology and medicine[J]. Nature Medicine, 2021, 27(10): 1666-1669.

[100] TANG Q Y, REN W T, WANG J, et al. The statistical trends of protein evolution: a lesson from AlphaFold database[J]. Molecular Biology and Evolution, 2022, 39(10): msac197.

[101] SATO T U, KANEKO K. Evolutionary dimension reduction in phenotypic space[J]. Physical Review Research, 2020, 2(1): 013197.

[102] SAKATA A, KANEKO K. Dimensional reduction in evolving spin-glass model: correlation of phenotypic responses to environmental and mutational changes[J]. Physical Review Letters, 2020, 124(21): 218101.

[103] KANEKO K. Constructing universal phenomenology for biological cellular systems: an idiosyncratic review on evolutionary dimensional reduction[J]. Journal of Statistical Mechanics: Theory and Experiment, 2024, 2024(2): 024002.

[104] KARPLUS M, KURIYAN J. Molecular dynamics and protein function[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(19): 6679-6685.

[105] AMBROGGIO X I, KUHLMAN B. Design of protein conformational switches[J]. Current Opinion in Structural Biology, 2006, 16(4): 525-530.

[106] MONTEIRO DA SILVA G, CUI J Y, DALGARNO D C, et al. High-throughput prediction of protein conformational distributions with subsampled AlphaFold2[J]. Nature Communications, 2024, 15: 2464.

[107] STEIN R A, MCHAOURAB H S. SPEACH_AF: sampling protein ensembles and conformational heterogeneity with AlphaFold2[J]. PLoS Computational Biology, 2022, 18(8): e1010483.

[108] DEL ALAMO D, SALA D, MCHAOURAB H S, et al. Sampling alternative conformational states of transporters and receptors with AlphaFold2[J]. eLife, 2022, 11: e75751.

[109] WAYMENT-STEELE H K, OJOAWO A, OTTEN R, et al. Predicting multiple conformations *via* sequence clustering and AlphaFold2[J]. Nature, 2024, 625(7996): 832-839.

[110] HEO L, FEIG M. Multi-state modeling of G-protein coupled receptors at experimental accuracy[J]. Proteins: Structure, Function, and Bioinformatics, 2022, 90(11): 1873-1885.

[111] SALA D, ENGELBERGER F, MCHAOURAB H S, et al. Modeling conformational states of proteins with AlphaFold[J]. Current Opinion in Structural Biology, 2023, 81: 102645.

[112] SALA D, HILDEBRAND P W, MEILER J. Biasing AlphaFold2 to predict GPCRs and kinases with user-defined functional or structural properties[J]. Frontiers in Molecular Biosciences, 2023, 10: 1121962.

[113] WOLYNES P G, ONUCHIC J N, THIRUMALAI D. Navigating the folding routes[J]. Science, 1995, 267(5204): 1619-1620.

[114] BRYNGELSON J D, ONUCHIC J N, SOCCI N D, et al. Funnels, pathways, and the energy landscape of protein folding: a synthesis [J]. Proteins: Structure, Function, and Bioinformatics, 1995, 21(3): 167-195.

[115] FERREIRO D U, KOMIVES E A, WOLYNES P G. Frustration in biomolecules[J]. Quarterly Reviews of Biophysics, 2014, 47(4): 285-363.

[116] PARRA R G, SCHAFER N P, RADUSKY L G, et al. Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics[J]. Nucleic Acids Research, 2016, 44(W1): W356-W360.

[117] FERREIRO D U, HEGLER J A, KOMIVES E A, et al. Localizing frustration in native proteins and protein assemblies [J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(50): 19819-19824.

[118] LI W F, WOLYNES P G, TAKADA S. Frustration, specific sequence dependence, and nonlinearity in large-amplitude fluctuations of allosteric proteins[J]. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(9): 3504-3509.

[119] CHEN M C, CHEN X, SCHAFER N P, et al. Surveying biomolecular frustration at atomic resolution[J]. Nature Communications, 2020, 11: 5944.

[120] GIANNI S, FREIBERGER M I, JEMTH P, et al. Fuzziness and frustration in the energy landscape of protein folding, function, and assembly[J]. Accounts of Chemical Research, 2021, 54(5): 1251-1259.

[121] GUAN X Y, TANG Q Y, REN W T, et al. Predicting protein conformational motions using energetic frustration analysis and AlphaFold2[J]. Proceedings of the National Academy of Sciences of the United States of America, 2024, 121(35): e2410662121.

[122] XIE T Y, SONG Z L, HUANG J. Conditioned protein structure prediction[J]. PRX Life, 2024, 2(4): 043001.

[123] CHAKRAVARTY D, SCHAFER J W, CHEN E A, et al. AlphaFold predictions of fold-switched conformations are driven by structure memorization[J]. Nature Communications, 2024, 15: 7296.

[124] BRYANT P, NOÉ F. Structure prediction of alternative protein conformations[J]. Nature Communications, 2024, 15: 7328.

[125] ZHANG J, LIU S R, CHEN M Y, et al. Unsupervisedly prompting AlphaFold2 for accurate few-shot protein structure prediction[J]. Journal of Chemical Theory and Computation, 2023, 19(22): 8460-8471.

[126] LEE C Y, HUBRICH D, VARGA J K, et al. Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation[J]. Molecular Systems Biology, 2024, 20(2): 75-97.

[127] GUO Z Y, LIU J, WANG Y L, et al. Diffusion models in

bioinformatics and computational biology[J]. Nature Reviews Bioengineering, 2024, 2(2): 136-154.

[128] WU K E, YANG K K, VAN DEN BERG R, et al. Protein structure generation *via* folding diffusion[J]. Nature Communications, 2024, 15: 1059.

[129] PILLAI A, IDRIS A, PHILOMIN A, et al. *De novo* design of allosterically switchable protein assemblies[J]. Nature, 2024, 632(8026): 911-920.

[130] ECKMANN J P, ROUGEMONT J, TLUSTY T. *Colloquium*: proteins: the physics of amorphous evolving matter[J]. Reviews of Modern Physics, 2019, 91(3): 031001.

[131] CHENG J, NOVATI G, PAN J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense[J]. Science, 2023, 381(6664): eadg7492.

[132] MARCHETTI F, MORONI E, PANDINI A, et al. Machine learning prediction of allosteric drug activity from molecular dynamics[J]. The Journal of Physical Chemistry Letters, 2021, 12(15): 3724-3732.

[133] BAI Q F, LIU S, TIAN Y N, et al. Application advances of deep learning methods for *de novo* drug design and molecular dynamics simulation[J]. Wiley Interdisciplinary Reviews: Computational Molecular Science, 2022, 12(3): e1581.

[134] PERRAKIS A, SIXMA T K. AI revolutions in biology: the joys and perils of AlphaFold[J]. EMBO Reports, 2021, 22(11): e54046.

[135] MONZON V, HAFT D H, BATEMAN A. Folding the unfoldable: using AlphaFold to explore spurious proteins[J]. Bioinformatics Advances, 2022, 2(1): vbab043.

[136] ANISHCHENKO I, PELLOCK S J, CHIDYAUSIKU T M, et al. *De novo* protein design by deep network hallucination[J]. Nature, 2021, 600(7889): 547-552.

[137] YANG K K, WU Z, ARNOLD F H. Machine-learning-guided directed evolution for protein engineering[J]. Nature Methods, 2019, 16(8): 687-694.

[138] BAYLY-JONES C, WHISSTOCK J C. Mining folded proteomes in the era of accurate structure prediction[J]. PLoS Computational Biology, 2022, 18(3): e1009930.

[139] LIU X Y, XING J Y, FU H H, et al. Analyzing molecular dynamics trajectories thermodynamically through artificial intelligence[J]. Journal of Chemical Theory and Computation, 2024, 20(2): 665-676.

[140] WANG T, HE X H, LI M Y, et al. *Ab initio* characterization of protein molecular dynamics with AI2BMD[J]. Nature, 2024, 635(8040): 1019-1027.

[141] BOLON D N, GRANT R A, BAKER T A, et al. Specificity *versus* stability in computational protein design[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(36): 12724-12729.

[142] PECCATI F, ALUNNO-RUFINI S, JIMÉNEZ-OSÉS G. Accurate prediction of enzyme thermostabilization with Rosetta using AlphaFold ensembles[J]. Journal of Chemical Information and Modeling, 2023, 63(3): 898-909.

[143] ZHU J, AVAKYAN N, KAKKIS A, et al. Protein assembly by design[J]. Chemical Reviews, 2021, 121(22): 13701-13796.

[144] HAYES T, RAO R, AKIN H, et al. Simulating 500 million years of evolution with a language model[J]. Science, 2025, 387(6736): 850-858.

[145] ROMERO P A, ARNOLD F H. Exploring protein fitness landscapes by directed evolution[J]. Nature Reviews Molecular Cell Biology, 2009, 10(12): 866-876.

[146] JIANG K Y, YAN Z Q, DI BERNARDO M, et al. Rapid *in silico* directed evolution by a protein language model with EVOLVEpro [J]. Science, 2025, 387(6732): eadr6006.

[147] KING R D, ROWLAND J, OLIVER S G, et al. The automation of science[J]. Science, 2009, 324(5923): 85-89.

[148] SAVINOV A, SWANSON S, KEATING A E, et al. High-throughput discovery of inhibitory protein fragments with AlphaFold[J]. Biophysical Journal, 2024, 123(3): 55A-56A.

通讯作者：唐乾元（1989—），男，博士，助理教授。研究方向为数据驱动的生物复杂系统理论框架构建，通过深度融合机器学习、统计物理与高性能计算方法，研究包括蛋白质分子和大脑等不同时空尺度的生物复杂系统，揭示其内在的普适性组织原理与动力学规律。

E-mail：tangqy@hkbu.edu.hk

第一作者：夏辰亮（1990—），男，博士，讲师。研究方向为蛋白质动力学的统计物理研究。

E-mail：xiacl1030@qq.com

共同第一作者：张泽成（1992—），男，博士研究生。研究方向为蛋白质序列、结构与动力学的统计、AI蛋白质结构预测和生物复杂性。

E-mail：zhzece@outlook.com