



分子组装理论基础的探究——现状与机遇

瞿航¹, 童添翼^{1,2}, 雷志超³, 石沛琛¹, 杨柳林¹, 曹晓宇¹, 高毅勤⁴, 侯中怀⁵, 徐昕², 田中群^{1*}

1. 厦门大学化学化工学院化学系, 固体表面物理化学国家重点实验室 (PCOSS), 能源材料化学协同创新中心 (iChEM), 厦门 361005

2. 复旦大学化学系, 上海 200433

3. 中国科学院生物物理研究所, 生物大分子国家重点实验室, 北京 100101

4. 北京大学化学与分子工程学院, 北京分子科学国家研究中心, 北京 100871

5. 中国科学技术大学化学物理系, 合肥微尺度物质科学国家研究中心, 合肥 230026

*通讯作者, E-mail: zqtian@xmu.edu.cn

收稿日期: 2022-09-23; 接受日期: 2022-11-16; 网络版发表日期: 2022-12-29

国家自然科学基金资助项目 (编号: 21991130, 21991131, 21971216, 21971217)

摘要 分子组装是在分子以上层次创造新物质、产生新功能的重要手段。提升分子组装过程的可控性和组装体功能性是该领域的核心目标, 但由于研究复杂分子组装体系的手段和理论的匮乏, 迄今绝大部分研究只能限于了解组装过程始末的“黑箱”模式, 这成为该领域的发展瓶颈之一。本文围绕分子到亚细胞层次的分子组装体系及其理论方法与模型, 尝试将物质、能量与信息作为度量分子组装研究复杂性的三个维度, 从简单到复杂地逐级剖析探讨分子组装的理论研究现状、机遇和突破口。首先, 在物质维度上揭示分子组装过程呈现多路径、多阶段和多尺度等复杂性; 进而在物质基础上引入能量维度, 阐明熵驱动或熵焓互补现象普遍存在于分子组装体系中, 并探讨远离平衡态的分子组装体系如何产生时空有序的功能耗散结构; 继而在物质与能量基础上引入信息维度, 探讨分子组装信息网络中的正、负反馈协同, 以及如何促使系统涌现出复杂的生理功能。为了突破分子组装理论研究的瓶颈, 似乎亟需建立更大的框架。从物质、能量和信息的三个维度协同研究, 有望系统深入认知组装规律, 进而建立新理论, 发展高效精准的调控手段, 提升分子组装体系的复杂性和功能性, 并可能为生命科学和软物质科学提供新视角和新方法。

关键词 分子组装, 可控组装, 协同, 反馈, 非平衡态, 信息论

1 引言

物质科学是以认知自然物质结构和相互作用规律, 进而以创造新物质和产生新功能为主要目标的科学, 而化学合成和分子组装是两种最重要的手段 (图

1a)。化学合成历经两世纪, 着重于共价键的操控, 已发展出成熟的合成方法学, 在分子层次创制新型化合物, 极大拓展了化学合成的边界, 丰富了人类的物质生活。但是, 基于共价合成的材料在复杂性和功能性等各方面仍无法满足当今日益增长的物质需求。分子组装

引用格式: Qu H, Tong T, Lei ZC, Shi P, Yang L, Cao X, Gao Y, Hou Z, Xu X, Tian ZQ. Exploring the theoretical foundation of molecular assembly: current status and opportunities. *Sci Sin Chim*, 2023, 53: 145–173, doi: 10.1360/SSC-2022-0186

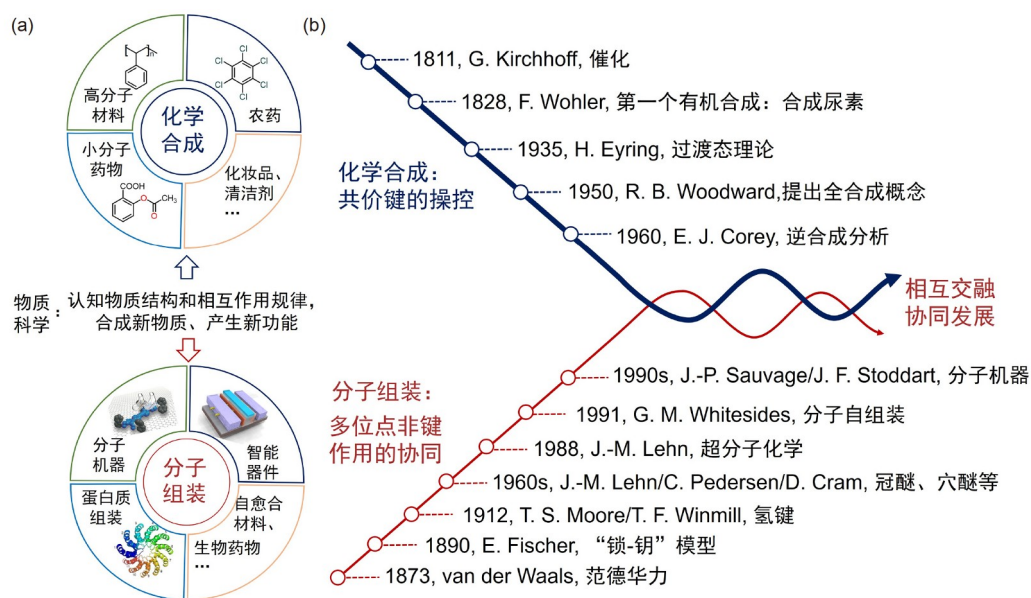


图 1 (a) 化学合成与分子组装分别在物质科学中的作用和定位. (b) 二者相互交融且协同发展的化学合成与分子组装的发展简史 (网络版彩图)

Figure 1 (a) The roles and positions of chemical synthesis and molecular assembly in matter science, respectively. (b) The brief history of chemical synthesis and molecular assembly, which is integrated with each other and developed synergistically (color online).

(molecular assembly)^[1-5]则在分子层次以上, 基于范德华力、氢键和静电相互作用等非共价键作用 (以下简称非键作用), 提供一种创制分子机器^[6,7]、智能器件^[8,9]和自愈合材料^[10,11]等功能材料和仿生材料的有效途径, 也为从分子相互作用角度理解复杂的生命现象提供指导方法^[12-14]. 因而在物质科学中, 分子组装与化学合成二者相互交融, 协同发展.

相较于成熟的合成化学研究 (图1b), 分子组装研究起步较晚, 发展尚不成熟. “分子组装”的基础和概念雏形最早可追溯至范德华力^[15] (1873年) 和“锁-钥”模型^[16] (1894年) 的提出, 随后氢键^[17]、卤键^[18,19]和亲疏水作用^[20,21]等非键作用的发现, 极大丰富了非共价作用的类型, 然而当时这些观测结果是独立分散和局部的, 缺少对分子组装领域整体性的归纳与认知, 因而限制了该领域的进一步发展. 1960年代, Charles Pedersen、Jean-Marie Lehn和Donald Cram相继合成一系列具有离子选择性识别的大环分子^[22-24], 如冠醚、穴醚等, 并因此荣膺1987年诺贝尔化学奖. 在此基础上, Lehn于1988年延伸了分子识别的概念^[25], 正式定义超分子化学为“chemistry beyond the molecule”, 指出“超分子化学是分子层次以上, 研究多种分子通过非共价相互作用形成有序功能聚集体的化学”^[2,25]. 1980

年代, 化学家发现一系列有机硫分子可在金表面自发形成有序的单层膜结构^[26,27]. 1991年, Whitesides等^[4]基于此类现象, 提出分子自组装 (molecular self-assembly) 概念, 即分子在平衡态下, 通过非共价作用自发形成稳定、规则聚集体的过程. 此外, 一系列新概念 (如“非共价合成”^[28]和“自适应”^[25]等) 的提出与碰撞, 也促使分子组装领域蓬勃发展.

分子组装的概念虽然提出较晚, 但分子组装现象广泛存在于生命体中, 如DNA双螺旋结构、蛋白质的多级结构和磷脂双分子层等, 并成为复杂、精妙的生命体功能的基石. 而人工分子组装领域在化学家的努力下和自然界的启发下, 历经三十余年, 在各方面取得长足的发展: 在结构复杂性方面, 大量具有精妙拓扑结构的组装体被制备, 如分子组结^[29-31]、共价-非共价杂化聚合物^[32]和巨分子组装体^[33]等; 在可控性方面, DNA折纸术^[34,35]、框架核酸自组装^[36,37]和框架诱导组装^[38-40]等策略被用于构筑形状、尺寸可控的组装结构; 在功能性方面, 人工光合组装体^[41]、肿瘤治疗^[42,43]和载药^[44]等实现了材料性能的跨越式发展.

分子组装通过“自下而上” (bottom-up) 的方式, 依靠分子间多位点非键作用的协同, 实现在分子尺度, 精准控制材料形貌与缺陷, 已成为传统的“自上而下”

(top-down) 之外的重要加工方法. 然而, 目前人工分子组装体系仍无法达到生命体中的分子组装体系的复杂、精妙和“智能”, 且缺乏高效、精巧且温和的调控手段. 因此, 很可能需要全面拓展乃至革新分子组装的研究范式, 深化拓展其理论基础.

目前分子组装的研究范式以设计组装基元为核心, 通过改变组装基元获得不同的组装体结构和功能. 然而, 由于针对多位点非键作用的研究手段和分子组装理论的匮乏, 尚无法详细研究大部分的组装过程. 这种只关注始末的“黑箱”研究范式极大限制了组装领域的发展. 因此, 近年来分子组装过程的机理和理论研究愈发受关注. 但是, 现有分子组装研究的理论研究主要通过延伸已有的化学合成理论^[45], 如过渡态理论^[46]、速率理论^[47]等. 关于分子组装的理论模型可追溯至19世纪初的平衡态成核生长模型^[48,49], 该模型也进一步催生出Potts模型^[50,51]、Ising模型^[52,53]和胶体模型^[54-56]等. 但这些模型极少关注分子细节和区分各类非键作用, 导致其在拟合某些具备特定结构的组装基元的组装过程时, 产生了极大的偏差. 而随着表征技术的发展, 分子组装可在实验上被定量分析, 甚至可视化. 因此, 针对具体的实验体系, 更多的分子组装机制被提出, 如超分子聚合的非经典“成核”机制^[57,58]、手性非线性放大的热力学机制^[59]、结合数学理论的分形笼结构分析方法^[60,61]等. 同时, 一些生物大分子的组装机制也被深入研究, 如阿兹海默症中关键A β 蛋白的错误折叠机制^[62,63]等. 但这些理论的针对性极强, 无法更全面、宽泛地指导如何精确操控非键作用和调控组装结构的动态特征, 也缺乏关键组装步骤的物理化学图像. 因此, 亟需发展更为普适的分子组装理论和研究框架.

为此, 本文通过探讨包括复杂生物大分子在内的分子组装体系和当前的理论方法与模型, 期望在物质科学的大框架下, 更系统、全局地审视现有分子组装的研究方法、理论基础和面临的挑战, 尝试发掘突破分子组装领域发展瓶颈的指导方法与机遇. 这要求我们从科学研究的哲学基础和认识论出发, 提升对分子组装研究的理解和认知. 在此方面, 著名物理学家Anderson在1972年提出的“More is different”思想^[64]可能为我们指明了方向, 其中文翻译为“多则异”, 核心为“在复杂性的每一个层级, 都会有崭新的性质出现, 也需要有新的定律、概念和原理”. 即虽然高层级的物质

科学也遵循底层级的定律, 但其表现出的行为和功能是无法仅凭底层级的理论所理解, 而表现出更高层级所独有的基本原理和核心规律. 因此, 物质科学研究普遍遵循由简至繁、逐级深入的研究规律, 而不同层级的物质科学研究的发展成熟度不同: 基于目前计算能力的局限, 越简单的体系越可精细研究, 越复杂的体系则一般较粗略研究, 并要采用“多则异”的原则去关注是否涌现出新特征和规律. 因此, 面向复杂分子组装研究, 可能需要首先对其分级分类.

控制论的创始人N. Wiener指出物质、能量与信息是组成世界的三大要素. 物质是能量与信息的载体, 能量与物质是系统与外界发生交换的主要方式, 信息则是维持系统稳定性和响应外界刺激的关键. 我们因此认为, 或许可将物质、能量与信息作为度量分子组装研究复杂性的三个维度, 进而对当前的分子组装的理论研究进行分级分类 (图2a). 在物质维度, 主要研究分子组装基元及溶剂等物质在平衡态下的组装过程, 如通过多物种的平衡组装过程, 可构建复杂的人工分子组结^[31] (图2b). 同时也需探讨多物种在复杂组装环境中的传质过程、相互作用模式及物理化学图像; 进而在物质维度上引入能量维度, 可使分子组装体系远离平衡态, 而体系内的非线性相互作用将微小的扰动放大, 使体系产生时空有序的功能耗散结构. 如在生命体中, ATP驱使马达蛋白沿微管蛋白定向移动而实现主动运输^[65] (图2c); 最复杂的层级是在物质和能量维度之上引入信息维度并实现三者协同, 如在细胞中的转录过程中, 贮存在mRNA中的信息在GTP的驱动下, 经由转录机器 (核糖体) 将遗传信息写入蛋白质序列^[66] (图2d). 同时复杂分子组装系统中的各物种将形成信息网络, 信息流在网络中被存储、处理及转换, 依靠正、负反馈的协同和不同层级间信息形式的转变, 进而涌现出自适应等功能. 目前分子组装研究主要关注物质维度, 对能量维度的研究刚起步, 对信息维度的探讨则鲜少涉及. 因此, 从物质扩展到能量与信息维度, 逐级深入地理清研究思路, 进而建立跨尺度、跨层级的理论模型和计算方法. 同时在每一层级需分类研究对象, 从多维度探究分子组装机制, 把握组装过程的关键物理化学参数变化, 可能有助于高效把握各个维度的关键性特征及其关联, 全面深入认知组装规律, 全面构建不同层级分子组装过程的清晰物理化学图像和规律, 进而建立新理论和实验方法, 以期指导构建具

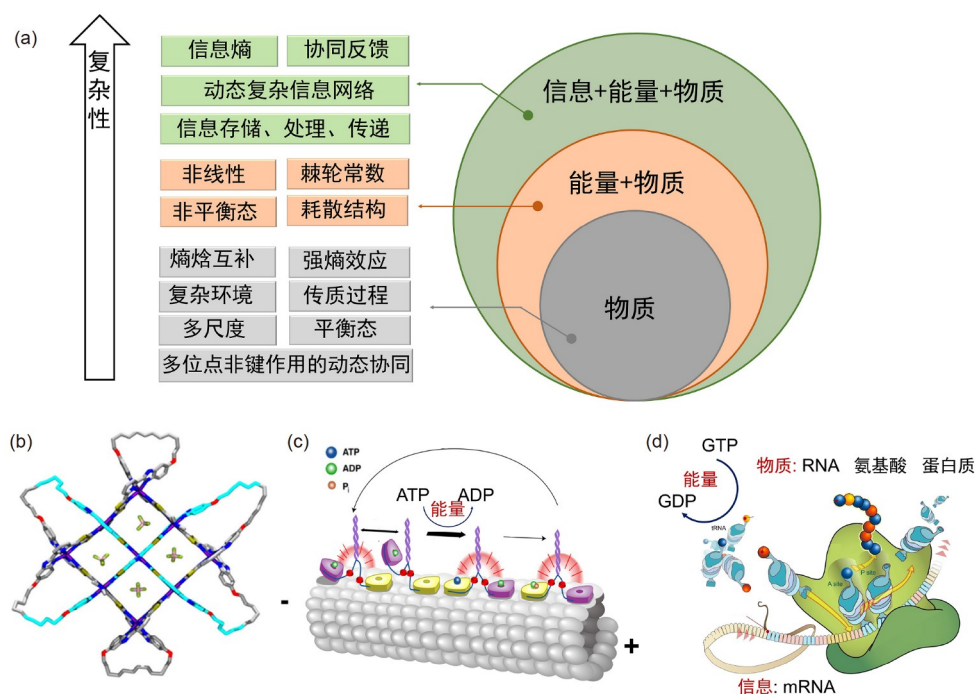


图 2 (a) 分子组装研究以物质、能量及信息的三个维度进行分级分类. 随着复杂性的提升, 分子组装体系可能涌现出新特征和规律, 并体现更强的功能. (b) 在简单的物质维度: 多物种通过平衡组装过程而形成复杂的人工分子组结^[31]. (c) 在物质基础引入能量: 在细胞中, 能量驱使马达蛋白沿微管定向移动^[65]. (d) 在物质和能量基础上再引入信息: 在细胞内的转录过程中, 存储于 mRNA 的信息在 GTP 驱动下, 经由核糖体传递至蛋白质序列^[66] (网络版彩图)

Figure 2 (a) Classification of molecular assembly research in the three dimensions of matter, energy and information. Novel characters and laws may emerge from the molecular assembly systems with the increasing complexity, which could exhibit the more powerful functions. (b) In the simple dimension of matter: multiple components can form complex artificial molecular knots through the equilibrium assembly processes^[31]. (c) In the dimension of energy beyond the matter: driven by the energy, the motor proteins travel in a certain direction along the microtubules in the cell^[65]. (d) In the dimension of the information beyond the matter and energy: in the transcription process in the cell, the information stored in the mRNA is transferred to the protein sequence driven by GTP via the ribosome^[66] (color online).

有多样性和功能性的众多分子组装体系.

2 分子组装的物质维度

组装基元是分子组装过程的核心物质, 其复杂性主要体现在组装基元间的作用力、组装过程的时间尺度和组装基元的输运等多方面. 同时组装基元周围的其他溶质和溶剂分子也将通过影响组装基元的电子结构、构象与作用力, 进而改变组装路径, 决定最终的产物分布.

2.1 组装基元间的多位点非键作用的动态协同

组装基元间作用力的复杂性主要体现在作用强度、作用力间的耦合及作用力的动态性三方面.

化学键一般具有局域性强的特征, 因此化学合成

往往仅关注单一或少数共价键的生成或断裂, 而共价键的键能通常在 200 kJ mol^{-1} 以上, 其成键或断键过程往往是不可逆的 (图3a). 与之相对的是, 分子组装更依靠组装基元间的多位点非键作用的协同. 这些非键作用的键能在 $1\sim 100 \text{ kJ mol}^{-1}$ 范围. 较弱的键能意味着非键作用的形成过程是较可逆的, 使组装体可通过不断“试错”而最终达到热力学稳定状态. 在一些仅涉及少数独立的非键作用的简单分子组装过程中, 可采用成熟的分子轨道理论, 如耦合簇理论、微扰理论和密度泛函理论等方法, 从微观电子结构角度解析这些非键作用. 但这些方法存在成本过高 (如在完备基组下的 CCSD(T) 计算) 或不够精确 (如 MP2), 甚至完全失效 (如传统的 DFA 方法) 等问题, 其主要原因在于大量密度泛函方法存在静态相关误差且自相关误差严重等问题^[67]. 因此, 发展精确、高效描述非键作用的密

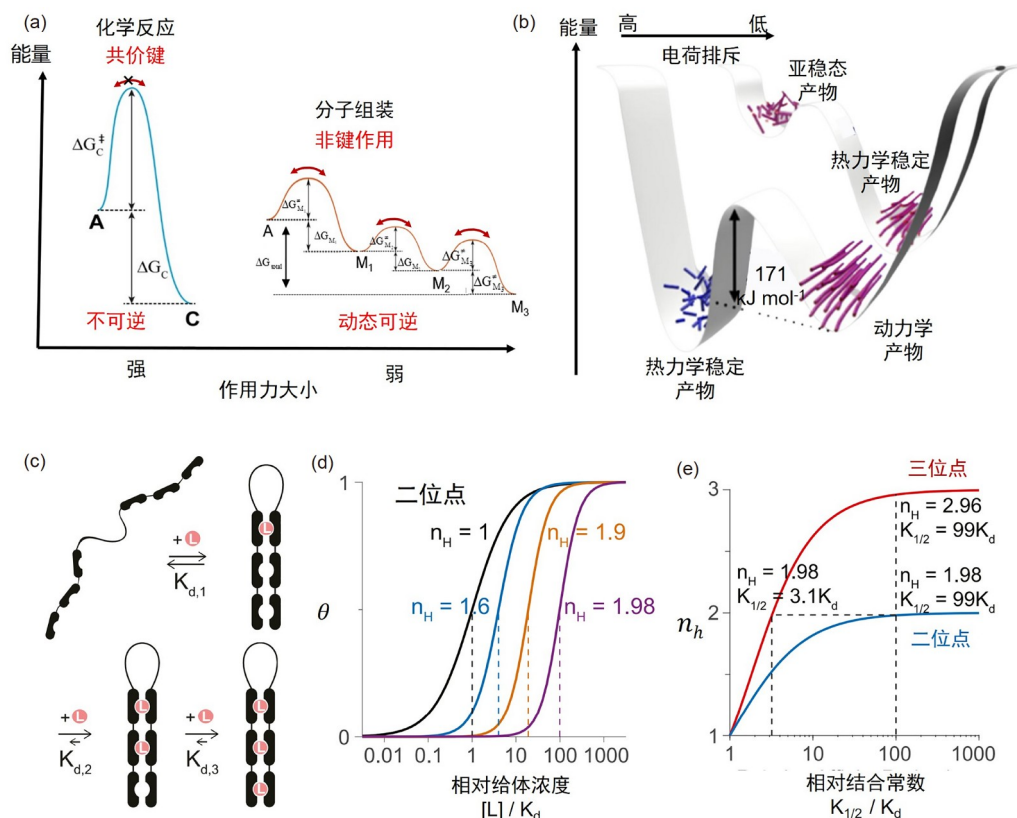


图 3 (a) 化学反应和分子组装的区别. (b) 两性多肽分子的平衡与非平衡组装路径的势能曲线图^[74]. (c) 具备三位点协同效应的受体分子组装过程示意图^[76]. (d) 具备二位点协同效应的受体分子随Hill系数的增加, 结合曲线的斜率越陡且动态范围越窄. (e) 三位点受体 (红线) 相较二位点受体 (蓝线) 可获得更高的Hill系数, 也在取得同等协同性的同时, 维持更高的亲和能力 (网络版彩图)

Figure 3 (a) The difference between the chemical reaction and molecular assembly. (b) The potential energy landscapes of equilibrium and non-equilibrium pathways of amphiphilic polypeptides ^[74]. (c) Schematic representation of synergistic assembly processes of host molecules with three binding sites ^[76]. (d) The increase in Hill coefficients will increase the steepness of binding curves of a cooperative receptor with two binding sites and narrow its dynamic range. (e) A cooperative receptor with three binding sites (red line) can achieve a higher Hill coefficient than a two-site receptor (blue line). Alternatively, it can achieve the same cooperativity (same n_H) as a two-site receptor while maintain much higher affinity (color online).

度泛函理论是分子组装领域的重要发展方向之一. 近年来, 徐昕等发展的XYG3型新泛函^[68]可高效描述非键作用, 其交换相关能的形式为:

$$E_{XC}^{XYG3} = a_1 E_x^{HF} + (1 - a_1) E_x^{LDA} + a_2 \Delta E_x^{B88} + (1 - a_3) E_c^{LYP} + a_3 E_c^{MP2} \quad (1)$$

a_1 , a_2 , a_3 为经训练集优化后的3个混合参数. XYG3在占据轨道的基础上引入了非占据轨道信息, 由HF, LDA和B88的交换能和LYP和MP2相关能两项双杂化而成, 并占据和非占据轨道信息, 可在绝热近似的框架下, 极大修正目前通用泛函在生成热、反应能垒和各类非键作用等方面的缺陷, 最终可媲美高等级从头算的结果. 同时, XYG3型泛函可结合扩展ONIOM组

合方法 (XO)^[69, 70], 可在考虑长程作用的情况下, 进一步高效描述考虑外界环境的新型非键作用, 如溶剂环境中的阴离子- π 作用^[71]等, 成为目前研究非键作用最精确的泛函之一.

非键作用通常对环境变化敏感, 呈现动态性. 如许多酶的结构受温度、溶液电解质状态和pH变化等影响而发生变构, 进而调控其空腔内的亲疏水作用力或氢键等, 改变其与底物分子间的相互作用而影响酶的活性^[72, 73]. 作用力的动态性是深入了解分子组装的重要切入点, 一方面其赋予生命体中的组装体较强的自适应能力和响应性, 另一方面也可利用这种动态性, 调控组装过程. Stupp等^[74]通过人为稀释或改变溶液离子强度的方式, 控制两性多肽分子中的两种竞争作

用力 (β -折叠的氢键和静电排斥作用), 进而使新生多肽分别通过平衡或非平衡组装路径, 获得单分散的短肽或长肽纤维 (图3b)。

分子组装过程中往往涉及多位点非键作用的协同, 即前一个作用位点被占据后, 将促进 (或抑制) 后一个作用位点的结合, 与之对应的现象称为正 (负) 协同效应。如在红细胞中, 氧气与不饱和的血红蛋白结合后会显著增加后续氧气分子的结合常数^[75]。在多位点的受体中, 单一非键作用的协同性可用Hill系数表示, 即^[76]:

$$\theta = \frac{[L]^{n_H}}{[L]^{n_H} + K_{1/2}^{n_H}} \quad (2)$$

$$K_{1/2} = \left[\prod_{i=1}^n K_{d,i} \right]^{1/n} \quad (3)$$

θ 为占据位点占受体总位点的分数, 其受到给体浓度 $[L]$ 的影响, $K_{1/2}$ 代表宏观解离平衡常数, $K_{d,i}$ 表示在 n 个位点中, 第 i 位点的解离平衡常数 (图3c)。 n_H 为衡量体系协同性的Hill系数, 在无协同性的体系中, n_H 等于1; 体系为负协同时, n_H 小于1, 且 n_H 越小则负协同性越强; 体系为正协同时, n_H 大于1, 且 n_H 越大且越接近于受体的结合位点数量 n , 正协同性越强 (图3d), 结合曲线的斜率更陡。虽然协同性可显著改善受体对给体浓度微小变化的响应, 但代价是其 $K_{1/2}$ 将远大于同等的无协同性受体的宏观解离平衡常数 (K_d)。上述缺陷可通过增加结合位点的数量, 进而调节二者间的内在平衡 (图3e)。目前, Hill系数已被广泛应用于一维超分子聚合物^[77]等研究中。但Hill系数仅能表示单一非键作用的协同性, 不适用于具有多种非键作用协同的复杂组装体系。如在生命体中的蛋白质折叠过程中, 主链间的氢键使其产生 α 螺旋或 β 折叠等二级结构, 而侧链的氨基酸残基间的疏水作用、电荷作用和氢键等作用力则使蛋白质产生三级结构和更复杂的四级结构, Hill系数可能不再适用。因此, 如何更深入揭示多位点、多种非键作用协同的本质和规律, 是组装理论研究所面临的重要挑战和发展机遇。

2.2 跨多时空尺度的分子组装模拟

一些简单的分子组装过程, 如主客体识别过程, 其时空尺度一般在纳秒和纳米量级。在该时空尺度下, 可基于量子力学方法, 如HF方法、多组态自洽场方法和

密度泛函理论 (DFT) 等, 以 1 kcal mol^{-1} 的高精度解析基元和简单组装体的电子结构和动力学, 提供其丰富的微观细节信息。然而, 高精度的第一性原理方法需耗费大量的时间和计算资源, 且通常忽略长程相互作用对分子间作用力的影响, 使其在模拟复杂的大体系时存在较大误差。对大尺度的分子组装体系, 通常采用基于经验力场模型的分子动力学模拟 (MD) 方法, 直观解析动态变化过程。同时, 在高分子和生物大分子的分子组装模拟中, 如何对多构象的采样, 包括构象熵的计算, 可能是当前发展跨尺度、高精度计算方法所面临的主要瓶颈。

复杂的组装体系往往涉及微米级别及以上的空间尺度和微秒级别及以上的时间尺度^[78], 且组装基元与组装体在尺度上可能相差几个数量级, 且涉及多尺度过程, 这导致了传统的模拟方法力有不逮。在此方面, 人们通常采用两种截然不同的策略: (1) 坚持全原子模拟并执行增强采样; (2) 采用粗粒化模型, 通过牺牲细节信息以提高速度。但若单独采用其中某种策略, 都会显示出严重的局限性。如在模拟生命体中的细胞膜组装过程时^[79] (图4a), 需以多尺度模拟探索组装体内部构象和分子间相互作用: 用全原子模型描述脂质-脂质、脂质-蛋白质和蛋白质-蛋白质的堆积方式和氢键网络; 用粗粒化模型 (CG) 探索脂质和蛋白质的动态分布及脂质的相分离; 用超粗粒化模型或连续模型解析更大尺度的复杂分子组装过程, 如脂质介导的蛋白质组装乃至全细胞建模。因此, 发展自适应的多尺度模拟方法仍面临诸多挑战: 如何确定关键的相互作用力和优化粗粒化模型中的参数, 在兼顾精度的情况下提升计算效率, 增强不同尺度模拟方法间的衔接和信息传递, 提取含罕见事件的平衡系统信息, 建立联组组装体的微观结构、介观形貌和宏观功能的模拟方法; 如何发展高效的增强采样技术, 在相空间中精准采样, 得到收敛的热力学性质, 建立针对复杂多级功能体系的热力学数据库。

针对此类问题, 高毅勤等^[80]从无监督的增强机器学习角度, 重新审视经典的增强采样方法, 并提出增强变分对抗密度估计 (RE-VADE) 方法。RE-VADE通过联合细粒化 (FG) 模型和粗粒化 (CG) 模型, 使二者可在不同尺度的模拟中串联而互相补足 (图5): 通过深度生成学习, CG模型在FG模型基础上优化, 并包含更多细节信息; 相应地, FG模型在CG模型的引导下而增

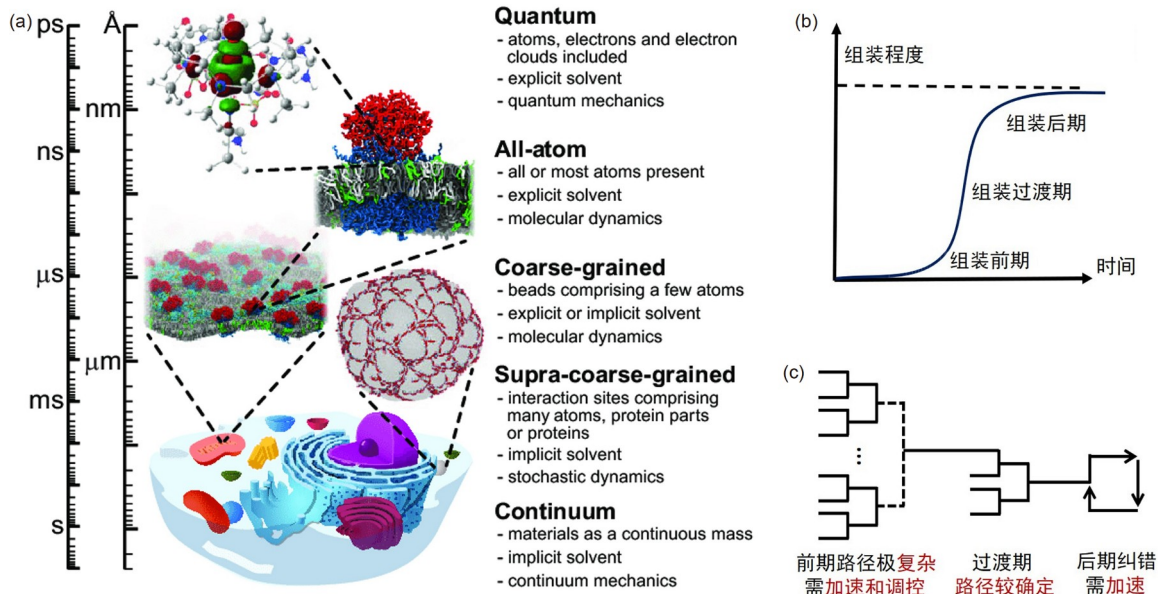


图4 复杂分子组装体系跨越多个时间与空间尺度. (a) 细胞中不同尺度组装过程的模拟方法^[79]. (b) 复杂组装过程存在明显的阶段性, 分别为前期、过渡期及后期. (c) 分子组装的每一阶段呈不同特性 (网络版彩图)

Figure 4 Complex molecular assembly systems at multiple temporal and spatial scales. (a) The simulation methods for the molecular assembly in the cell at the different scales^[79]. (b) Distinct periods of complex molecular assembly, corresponding to the earlier, transitional and later periods. (c) The different characters in each period of molecular assemblies (color online).

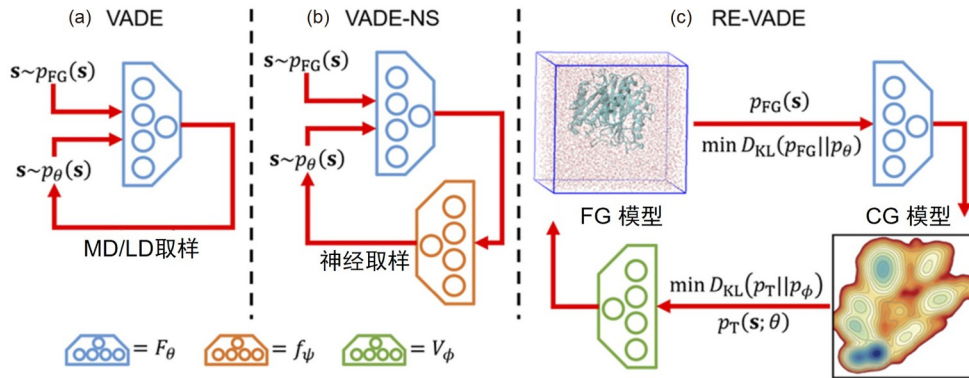


图5 VADE方法的工作流程^[80]. (a) VADE: 自由能估计 F 同时包含正样本 $[s \sim p_{FG}(s)]$ 和负样本 $[s \sim p_{\theta}(s)]$. 负样本基于 F , 通过MC或Langevin动力学(LD)模拟生成. (b) 含神经样本器的VADE (VADE-NS): 与VADE相同, 但负样本是通过神经取样器 f_{ψ} 获得. (c) RE-VADE: 给定FG模型(p_{FG})的样本和由VADE近似的CG势能 F . 红色箭头表示数据流或计算图的方向 (网络版彩图)

Figure 5 Workflow of variational adversarial density estimation (VADE)^[80]. (a) VADE: the free energy approximator F simultaneously incorporates positive samples $[s \sim p_{FG}(s)]$ and negative samples $[s \sim p_{\theta}(s)]$. The negative samples are generated by Monte Carlo (MC) or Langevin Dynamics (LD) simulations over F . (b) VADE with a neural sampler (VADE-NS): identical to VADE, except that the negative samples are drawn by using a neural sampler f_{ψ} . (c) Reinforced VADE (RE-VADE): Given simulation samples from fine-grained (FG) models (p_{FG}), a coarse-grained (CG) potential F can be approximated. Red arrows indicate the direction of data flow or the computational graph through VADE (color online).

强. 通过计算表明, 与传统CG方法相比, RE-VADE方法能仅基于平衡FG样本, 更快地变分推导出精巧的CG模型, 从而减少人为因素和降低计算成本. 此外,

RE-VADE也能显著提高生物分子构象跃迁的采样效率. 尚需指出, 如何令FG和CG模型在同一框架下自治地进行, 仍是发展跨尺度、自适应模拟方法的重要方

向之一。

复杂的分子组装过程一般具有明显的阶段性, 如类比于晶体生长理论中的成核-生长-熟化三阶段^[81], 可简单分为组装前期、过渡期与后期, 每一阶段所涉及的时间尺度也不同(图4b, 4c)。在组装前期, 复杂组装基元有多个作用位点, 可能需通过大量“试错”过程, 精确对位每个作用位点。因此组装前期耗时长, 且可能存在多种组装路径。各种非键作用通过协同的方式, 确定了主导的组装路径而进入组装过渡期, 该阶段的路径明确, 耗时相对较短, 因此也导致组装体结构的不完善而存在一定的缺陷。在组装后期, 基于组装过程的良好可逆性, 复杂组装体可能需长时间处于部分解组装—部分再组装的多次反复过程, 最终才可形成热力学稳定的组装体, 导致组装后期的耗时也较长。因此, 针对分子组装过程的各阶段的特点, 或许需要发展分阶段、多时间尺度的模拟方法和理论。

2.3 复杂、拥挤和动态的组装环境

组装基元并非是孤立的, 而是处在溶液(或溶剂)中。溶剂会显著影响溶质分子的电子结构与构象, 进而影响组装过程和产物分布^[82]。在理论计算与模拟中, 显式溶剂模型是考虑溶剂效应最直接的方法, 将溶质与周围溶剂分子作为一个体系进行处理。蒙特卡洛(Monte Carlo, MC)模拟或分子动力学(molecular dynamics, MD)模拟通常采用显式溶剂模型, 以得到系综平均和时间平均的性质, 但也大幅增加了计算量^[83]。在高精度计算中, 可将溶质与溶剂分子分开考虑, 采用不同精度的方法以降低计算量, 如QM/MM模型^[84], 中心溶质采用高精度的量子力学(QM)计算, 外围溶剂分子则采用分子力学(MM)计算。该模型可在考虑溶剂-溶质间的相互作用下, 精确计算较小体系的溶剂效应, 并得到溶质分子活性部位的电子结构。为降低计算量, 隐式溶剂模型也被广泛采用, 如连续介质模型^[85], 将整体溶剂视为连续介质, 考虑溶剂的主要性质(如介电常数和偶极矩等)对溶质分子的影响, 以减少体系自由度而减少计算量, 并可较精确处理长程静电作用。连续介质模型是溶剂的平均场模型, 仅能反映溶剂的总体效应, 无法分解为溶剂-溶质间具体的氢键、范德华力和CH- π 作用等非键作用, 且极度依赖反应场的选择。

应当特别指出, 生命体中的分子组装则处于更加

复杂的环境中(图6a)。例如, 细胞内环境挤满了大分子(占细胞体积的10%~50%), 还有各种游离的氨基酸和辅酶因子等小分子, 这些分子的浓度可高达80~400 g L⁻¹。因此, 细胞内环境是高溶质浓度、区域化、不均匀的且普遍存在体积排阻效应^[86,87]。在如此“拥挤”的细胞内环境中, 组装基元间的识别与组装将深受周围的溶质分子的影响。例如, 高浓溶液将显著影响单链DNA杂交过程的动力学^[88]。同时, 细胞代谢与外界环境的变化使其内环境的理化性质(如渗透压、pH值和温度等)时刻处于动态平衡。因此, 在复杂分子组装模拟中, 无法直接将周围环境作为连续介质且很难选择合适的反应场, 更无法体现可能存在的环境对组装分子的长程作用。因此, 采用连续介质模型将产生较大偏差。而在复杂环境下的QM/MM模拟中, 不仅需要精确计算分子组装的核心过程, 也需精确考虑周围环境中的其他溶质分子和溶剂对组装基元和过程的影响, 这无疑对复杂环境中的分子组装过程的模拟提出很大挑战。近年来, 基于机器学习的分子力场模拟^[89-91]可更精确地建模溶剂效应, 有望成为研究复杂环境中的分子组装体系的重要契机。

2.4 复杂环境下的传质过程

迄今开展的分子组装研究尚不重视传质问题, 事实上, 拥挤的细胞环境会显著影响传质过程。在稀溶液中, 化学反应和简单分子组装主要通过常规的传质方式, 如扩散、迁移与对流等。这些传质方式一般不会显著影响分子组装的路径。但在复杂分子组装过程中, 各种组装基元只有出现在合适的时间、空间位置, 才可顺利形成组装体和实现相关功能。因此, 其组装过程与传质过程紧密关联。尤其在高浓度的细胞内环境中, 各分子间的体积排斥效应虽然增加溶质的化学势, 但也严重阻碍了传统的传质过程^[66], 但很可能也提供了独特新颖的传质过程, 进而显著影响生命体中的分子组装机行为^[92](图6b)。

细胞内传质过程的主要机理包含被动扩散和主动运输两种。经典扩散模型以Fick定律为基石, 将扩散模式类比三维过阻尼系统中的布朗粒子的运动^[93]:

$$\frac{\partial}{\partial t}n(x, t) = D \frac{\partial^2}{\partial x^2}n(x, t) \quad (4)$$

n 为物质浓度, x 和 t 分别表示扩散距离与时间, D 为扩散系数。Fick定律的直观理解为扩散通量和扩散浓度梯

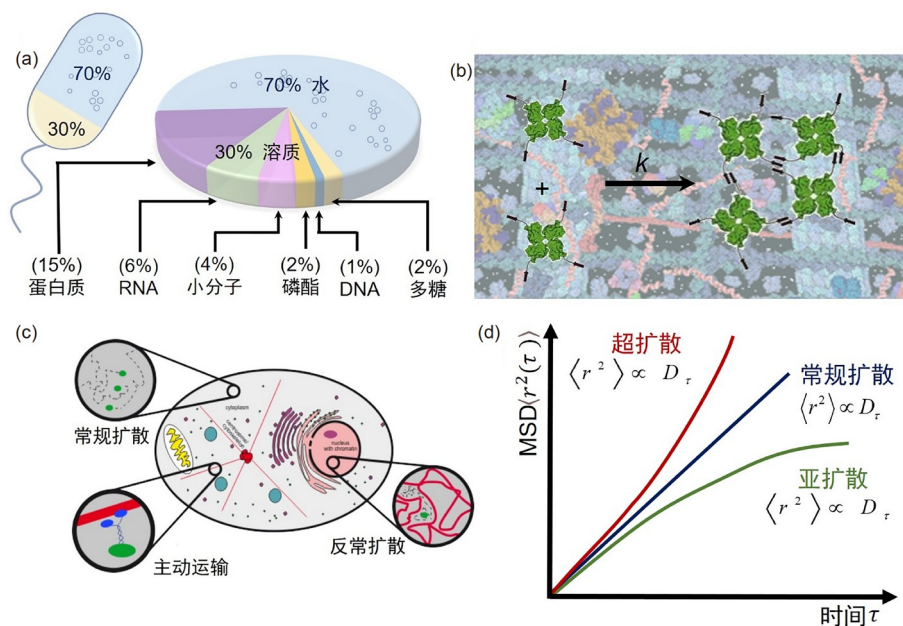


图 6 在复杂拥挤环境中的分子组装过程及传质过程. (a) 细胞环境的动态组成. (b) 在拥挤的细胞环境中, 蛋白质需要跨越长距离方可组装. (c) 细胞内三种传质过程. (d) 三种被动扩散形式^[96], 包含常规扩散和两种反常扩散 (超扩散和亚扩散) (网络版彩图)

Figure 6 Molecular assembly and mass transfer processes in the complex crowded environment. (a) The dynamic composition of cellular environment. (b) In the crowded cellular environment, proteins should travel in a long distance to assemble. (c) Three kinds of mass transfer processes in the cell. (d) Three kinds of passive diffusions [96], including conventional diffusion and two kinds of anomalous diffusions (superdiffusion and subdiffusion) (color online).

度正相关. 然而在拥挤的环境 (如生命介质) 中, 纳米颗粒 (如蛋白质和囊泡等) 存在更复杂的扩散规律和特征, 甚至呈现与经典扩散理论不一致的反常扩散现象^[94] (图6c). 目前, 通常采用分数型Brown运动模型描述此现象, 并以均方位移 $\langle r^2(t) \rangle$ (mean square displacement, MSD) 作为衡量微观颗粒扩散的重要指标^[95,96]:

$$\langle r^2(t) \rangle = 2dDt^\alpha \quad (5)$$

d 为与扩散维度有关的系数, 如 $d=3$ 表示三维扩散. α 是与扩散方式相关的系数. 当 $\alpha=1$ 时, 表现为常规扩散, 满足Fick定律; 而当 $\alpha < 1$ 时, 表现为亚扩散; 当 $\alpha > 1$ 时, 表现为超扩散 (图6d). 超扩散常发生于存在趋化作用的细胞中, 且往往伴随主动运输^[97,98], 亚扩散则常发生于拥挤环境中^[66]. 显然, MSD理论仅以 α 描述各类反常扩散现象具有较大的局限性, 一方面MSD理论无法全面反映扩散过程中的动力学细节, 另一方面, MSD理论也忽略各向异性、疏密相间的细胞内环境结构, 因而仅凭MSD模型无法有效描述复杂体系中的反常扩散现象. 为更精确地描述复杂的传质过程, 人们也发展出一系列的扩展模型, 如以受限扩散模型描述纳米粒子在

高聚物溶液中的扩散^[99]或突触囊泡在不规则神经元细胞内的移动^[100].

目前针对一些新型传质方式 (如相分离^[101~103]、主动运输^[66]、微流控中的层流、湍流^[104,105]等) 的研究可能为分子组装理论发展提供新思路. 同时, 可能需发掘更多反常扩散行为并揭示其与生物大分子的组装结构和功能间的关联, 如侯中怀等^[106]基于广义Langevin方程的动力学耦合合理理论 (kinetic mode-coupling theory), 提供一种简单而有效的理论框架, 研究大分子或高分子在拥挤环境中的扩散和构象动力学.

针对传质过程的分子组装理论或许也应注重传质过程对分子组装过程的影响. 特别是在拥挤的细胞环境中, 组装过程可能与传质过程处于同一时间和空间尺度, 各种传质过程与组装过程可能存在局域或长程的耦合作用, 将可能显著改变组装机制和路径. 近十余年来, 生物学家们也陆续发现一系列与生物分子传质过程相关的生物分子凝聚现象^[107~110]. 即在特定情况下, 细胞内的一些生物大分子会在多价相互作用的介导下, 组装凝聚而形成“相分离”, 进而建立凝聚相

和稀释相间的一种两相动态平衡, 且形成一种类似“无膜细胞器”的低熵状态^[110,111]. 不同于常规的有膜细胞器, 其为不具有固定的边界和特定结构的组装体, 但往往有更强的动态性和可逆性. 一方面, “无膜细胞器”的组装过程和传质过程相关^[111], 且传质过程可能是决速步. 另一方面, 这种“无膜细胞器”的出现使两相中的特定分子浓度可相差几百倍甚至更多, 从而使细胞无需大量合成该分子, 也可在细胞内形成浓度梯度而加速传质^[109], 进而可能提升整体分子组装过程速率. 这种独特的传质与组装相耦合的机制对仿生分子组装体系的构筑具有启示意义, 如通过“相分离”方式, 介导氨基酸或寡肽超分子纳米纤维的成核-生长过程^[112]. 同时, 揭示“相分离”的生物机制将有助于我们理解转录调控、染色质结构调控和免疫响应等多种生命过程的内在本质, 也将促进针对相关疾病的药物研发^[113,114]. 然而目前对生物相分离的机制和其他拥挤环境中的传质方式仍知之甚少, 相关的理论研究仍有待重视和大力发展.

3 分子组装的物质和能量维度

目前分子组装研究主要集中于物质维度, 但对于复杂组装体系, 仅考虑物质维度是不充分的. 在物质基础上引入能量维度, 探索分子组装过程的熵焓效应, 可揭示分子组装现象更深层次的物理化学机制. 同时, 引入外部能量也将赋予体系以非平衡特性, 并在时空尺度下调控分子组装体结构与功能.

目前, 大多数人工分子组装研究使用封闭体系, 极少与外界环境持续地发生物质与能量交换, 因此一般表现为平衡态过程^[115] (图7a), 即在该组装过程中, 其最终组装体处于一个全局自由能最小的势阱中, 是热力学上最稳定的状态. 若产物在形成期间被困于一个动力学势阱中, 短时间内无法逾越能垒到达最终的热力学稳定态, 则呈现一种相对稳定的亚稳态 (图7b). 但就理论而言, 只要给予其充分的时间, 亚稳态最终都能达到热力学稳定态.

平衡分子组装过程通常获得的产物或组装结构较单一且缺乏较强的功能性, 也无法实时、可控地调节产物结构与功能. 应指出, 复杂分子组装往往是非平衡过程, 其组装结构和功能依赖于对外部能量 (如光、声、电等外力外场或三磷酸腺苷等能量物质) 的

不断吸收、传导和耗散, 由此形成并维持时空有序的暂态结构或耗散结构. 同时, 研究非平衡分子组装过程的内在本质也将有助于理解某些重要的生物功能, 如细胞分裂^[116,117]和信号转导^[118-120]等, 也为仿生功能材料提供设计准则, 如分子马达^[65,121,122]和自适应材料^[123,124]等.

3.1 平衡态分子组装的熵和焓驱动力

在恒压的平衡态分子组装研究中, 吉布斯自由能变 ΔG ($\Delta G = \Delta H - T\Delta S$) 是封闭系统的主要平衡判据. 其中, H 为焓, 是体系内能与压力势能的总和; S 为熵, 与系统的微观状态数相关. 焓和熵是构成吉布斯自由能的重要热力学参数, 然而二者对系统的作用方式和结果迥然不同. 在焓为主要驱动力的体系中, 焓主要通过微观局部的相互作用, 直接显著地降低自由能而驱动反应或分子组装进行; 在熵为主要驱动力的体系中, 熵则主要通过宏观的整体调节, 增加系统的总微观状态数, 使整体实现熵增而驱动整个过程. 虽然焓和熵的作用方式不同, 但焓和熵的竞争与合作是决定体系最终状态的关键.

目前要表征组装过程中的各不同状态间的转移路径、能垒信息和熵焓变化, 仍面临巨大挑战, 理论计算与模拟则能补充分子组装过程中相关的重要信息, 如通过选取合适的反应坐标, 将组装体系的相空间离散化, 建立态与态之间转移的马尔科夫态模型 (Markov state model, MSM), 从而获得具有统计意义的物理化学图像. 如徐昕等^[125]通过建立MSM, 揭示二棕榈酰磷脂酰胆碱 (DPPC) 组装形成双分子层过程的热力学和动力学图像 (图8a). 基于MSM, 他们从有限时长的模拟轨迹推测体系的平衡态性质, 提出运用溶剂可及表面积 (SASA) 和均方根偏差 (RMSD) 为反应坐标, 得到该组装过程的一维和二维自由能图像 (图8b). 在分子力场模型下, 通过分析分子间不同的相互作用对自由能的贡献, 并计算组装过程中体系的熵焓变化情况, 揭示熵焓互补在组装过程中的关键作用 (图8c). 此外, 从宏观态之间的转移速率和转移概率, 也可揭示组装过程中各个亚稳态的比重随时间的变化情况. 需指出, 如何选取反应坐标及描述存在远离平衡态或多条平行组装路径的情况, 可能成为MSM分析法应用于复杂分子组装体系的挑战.

分子组装过程往往伴随着巨大的熵变, 且普遍存

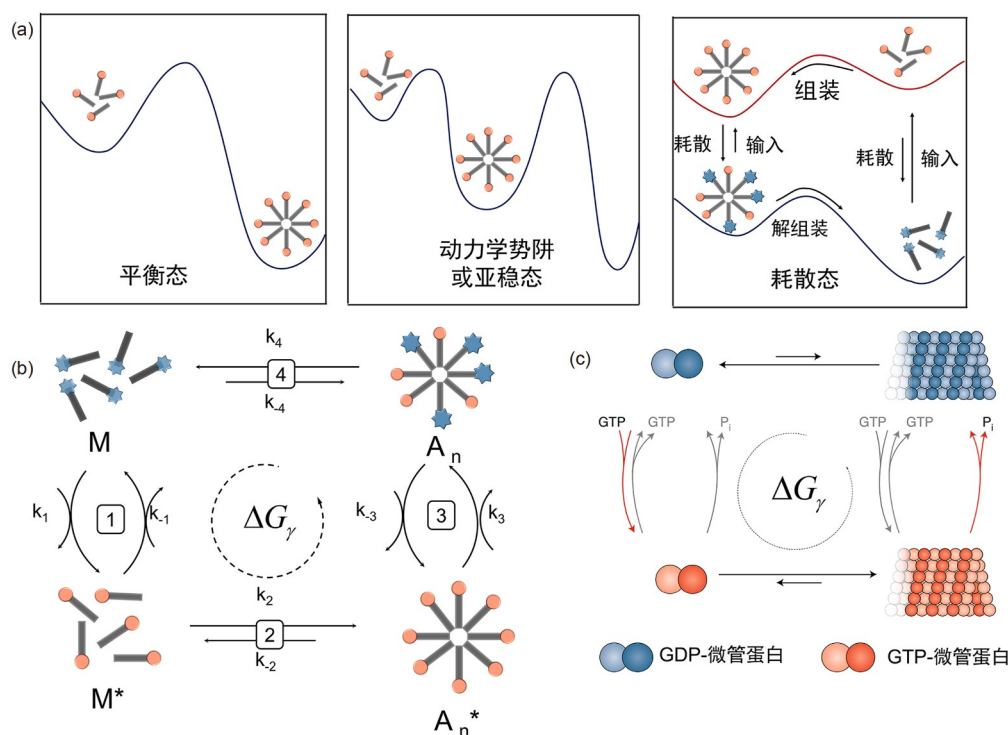


图 7 (a) 平衡和非平衡组装过程的自由能曲线对比^[115]。在平衡态, 组装体位于全局势能最低处(左)。处于动力学势阱的组装体处于局部势能面最低处, 而亚稳态的组装体则可逃逸出势阱, 因而存在有限的寿命(中)。在耗散体系中(右), 组装体位于相对较高势能面处, 只有不断输入能量或者物质才能够维持组装体结构。(b) 典型的非平衡组装过程示意图^[131]。(c) GTP驱动微管蛋白二聚体组装的示意图^[131]。GDP-微管蛋白二聚体不易发生组装, 而当GTP替代GDP, 产生GTP-微管蛋白二聚体且组装形成微管。随着GTP的不断水解, 微管也随之解体(网络版彩图)

Figure 7 (a) Comparison of the free energy landscapes of equilibrium and non-equilibrium assembly [115]. In the equilibrium assembly, the assembly resides in the global minimum (left). Kinetically trapped assemblies reside in a local minimum. By contrast, metastable assemblies can escape from the minimum and thus have a finite lifetime (middle). In dissipative assembly (right), the assemblies reside highly in the energy landscape and can only be sustained by the continuous input of external energy or matter. (b) General scheme for the non-equilibrium assembly [131]. (c) Reaction scheme for the GTP-driven assembly of tubulin dimers. GDP-tubulin dimers are not prone to assembly [131]. However, the substitution of GDP with GTP leads to the formation of GTP-tubulin dimers and microtubules. The continual hydrolysis of GTP results in the collapse of microtubules (color online).

在熵焓互补现象, 熵甚至可能成为主要驱动力^[126,127]。如Zacharias等^[128]通过全原子分子动力学模拟, 揭示丝状 β -淀粉样蛋白组装基元或部分组装体与纤维尖端结合时, 溶剂熵的增加补偿了组装过程中的构象熵减少(图8d, 8e)。这也解释随温度上升, 淀粉样蛋白稳定性增加的实验现象。需指出, 熵驱动的体系会在不同宏观状态间变化时, 更多地通过涌现的方式, 呈现熵驱动体系所不具备的新结构和新功能, 该类行为可统称为强熵效应(superentropic effect)^[129]。

虽然在强熵体系中, 熵占主导地位, 但焓的作用也不容忽略。在纯熵驱动的理想体系中(如硬核势主导的硬粒子体系^[130]), 组装基元间仅凭体积排阻效应, 使体系只能沿着熵增方向演变, 导致最终结构不具备功

能性。而焓介导的熵驱动体系在组装初期, 基元间的多种相互作用主导组装进程, 即仍表现为焓驱动过程, 此时熵主要体现为焓的制衡; 而随着体系有序度的增加, 体系从“弱”熵状态逐渐向转变为“强”熵状态, 熵则演变成自组织的关键驱动力, 也成为决定组装结构和功能的关键因素。这种熵焓交织现象在软物质体系中普遍存在, 且催生出复杂多样的生命现象。深入探究熵效应的规律及其物理化学机制, 并发展相应的熵调控策略, 将促进软物质领域相关理论框架的完善。更有意义的是, 信息熵(4.1节)作为传统熵概念的拓展, 将更有效地刻画与分析远离平衡态的分子组装体系(3.3.1节), 建构分子组装过程的信息转换及处理过程(4.1节), 因而有望成为理解复杂分子组装体系规律

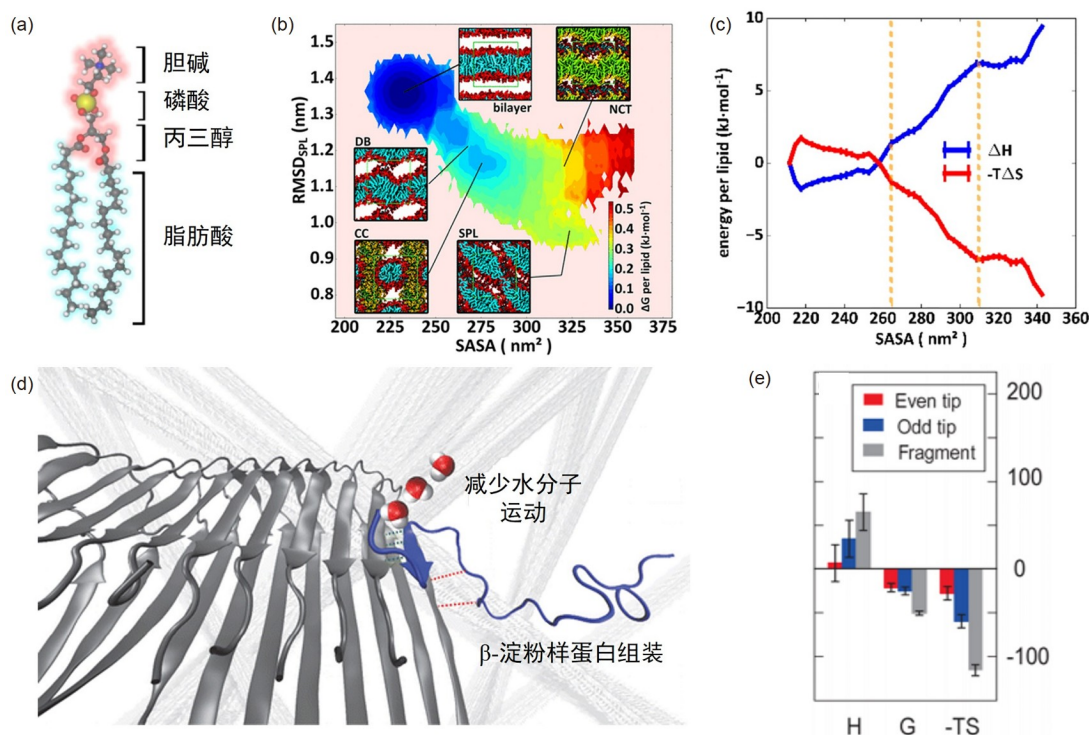


图 8 (a) DPPC 分子结构. (b) 以 SASA 和 RMSD 为坐标构建的组装过程二维自由能图^[125]. (c) 基于 SASA 坐标, 将自由能分解为焓和熵. (d) β -淀粉样蛋白组装过程示意图. (e) β -淀粉样蛋白组装过程中的吉布斯自由能 (G)、焓 (H) 和熵 (S) 的贡献^[128] (网络版彩图)

Figure 8 (a) The molecular structure of DPPC. (b) 2-D free energy landscape spanned on SASA and RMSD for the self-assembly process of DPPC [125]. (c) Decomposition of the free energy profile against SASA into the enthalpy and entropy terms. (d) Schematic diagram of self-assembly process of β -amyloid. (e) The contribution of Gibbs energy (G), enthalpy (H) and entropy (S) in the self-assembly process of β -amyloid [128] (color online).

的重要突破口。

3.2 非平衡分子组装过程的分子机制

与基于封闭系统的平衡分子组装过程不同, 生命体中的组装过程往往是开放体系, 需与外界环境发生净能量或物质交换, 而促使体系远离平衡态. 这种非平衡特性赋予生命体中的组装结构以独特的性质与功能. 虽然非平衡分子组装过程普遍存在于生命体系中, 但目前人工非平衡分子组装体系的研究仍处于初级阶段, 如何设计非平衡分子组装体系及揭示其能量耗散的调控机制是当前该领域面临的主要挑战之一. Prins 等^[131]提出一类典型的非平衡耗散组装循环机理 (图7a 右和7b), 并从势能曲线上区分平衡与非平衡过程: 组装前体 (M) 无法自发组装, 且位于所历经的势能面中的最低处, 而前体可接收能量并活化转变为组装基元 (M^*), 其势能也因外界的能量输入而提升至更高的能量状态. 此时, 组装基元存在两个竞争过程: 组装基元

发生相互作用, 进一步形成组装体 A_n^* ; 部分组装基元通过能量耗散返回前体. 因为组装体仍从外界环境中获得能量, 其结构 (A_n) 得以维持. 若外界再无能量输入时, 耗散结构则将解组装, 最终回到最初的组装前体. 在生命体中, A_n^* 与 A_n 均可能是具备功能的耗散结构. 通常可用棘轮常数 γ 描述整个耗散循环的方向性:

$$\gamma = \left(\frac{k_1}{k_{-1}}\right)^n \left(\frac{k_2}{k_{-2}}\right) \left(\frac{k_3}{k_{-3}}\right) \left(\frac{k_4}{k_{-4}}\right) \quad (6)$$

k_i 与 k_{-i} 分别为 i 步骤向前或向后的速率常数. 若 $\gamma > 1$, 耗散循环才能正向进行. 值得注意的是, 若终止流经体系的能量时, 耗散结构仍需一段时间发生解组装回到组装前体, 表明体系在耗散能量的同时也在储存能量于耗散结构中, 此能量存储的过程正是由于耗散循环的不对称性导致的. 这部分存储的能量 (ΔG_γ) 可用 γ 量化: $\Delta G_\gamma = -RT \ln \gamma$. 该模型也说明步骤3和4的动力学必须慢于化学燃料的催化转化速率 (图7b), 否则耗散循

环将在化学燃料耗尽后迅速平衡, 而所有能量仅是耗散而不会被储存.

若推广至一般的体系, 在如图9所示的 M 个物种 P_i ($i = 1, 2, \dots, M$) 参与的耗散组装循环中^[132,133], 将在某一时刻 t 下, 每个物种的浓度记为 $c_i(t)$, 并进一步将其可能性记作 $p_i(t) = c_i(t) / \sum_i c_i(t)$, 则有:

$$\frac{dp_i}{dt} = \sum_j (p_j k_{ji} - p_i k_{ij}) \quad (7)$$

$$J_{ij} = p_j k_{ji} - p_i k_{ij} \quad (8)$$

$$\Delta G_{ij} = k_B T \ln \frac{p_j k_{ji}}{p_i k_{ij}} \quad (9)$$

k_{ij} 表示组分 $i \rightarrow j$ 的速率常数. J_{ij} 表示 i 物种和 j 物种间的概率流. 当体系尚未稳定时, 组分的概率不断发生变化, 各组分间存在概率流, 即 $J_{ij} \neq 0$; 当体系达到定态时, 各组分的概率 p_i 不再随时间变化, 即 $J_{ij} = 0$. 此时:

$$\gamma = \prod_i \frac{k_{i,i+1}}{k_{i+1,i}} \quad (10)$$

$$\Delta G_\gamma = \sum_i \Delta G_{i,i+1} = -k_B T \ln \gamma \quad (11)$$

理论上, 棘轮常数 γ 与耗散循环的动力学不对称性相关, 即单体更易活化而耗散结构的解组装过程须为决速步, 而与其相关的存储能量 ΔG_γ 可用于量化衡量耗散结构远离平衡的程度.

鸟苷三磷酸 (GTP) 驱动的微管蛋白组装是生命体系中典型的非平衡组装过程^[134-136] (图7c). 微管蛋白及其组成的微管网络是细胞骨架的主要成分, 并在细胞内物质运输和细胞分裂中期的染色体分裂过程中发挥作用. 这些功能要求微管蛋白的组装必须是动态的, 可及时调整形貌以满足不同功能的需求. 组成微管的微管蛋白有两类: α 和 β 微管蛋白. 两种微管蛋白通过

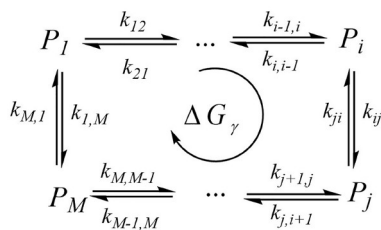


图9 M 个物种参与的耗散组装过程

Figure 9 The dissipative assembly process involving M kinds of species.

不断将GTP水解成鸟苷二磷酸 (GDP) 而形成头对尾的二聚体结构, 进而构筑横截面为13个微管蛋白的微管. GTP水解过程在微管中被极大加速, 而形成一个富含GDP且热力学不稳定的微管, 周围富含GTP的微管蛋白不断参与微管的组装, 从而有效避免其解体. 当体系中的GTP被消耗殆尽时, 微管也解组装而不复存在. 在整个过程中, 约有 21 kJ mol^{-1} 的能量存储于丝状组装体中^[131,137].

3.3 非平衡分子组装的理论模型

分子组装的平衡态往往对应于可逆过程, 而非平衡态则往往对应不可逆过程. 在平衡态附近, 系统趋向于平衡态或有类似平衡态行为的线性非平衡定态, 并伴随无序的增加和宏观结构的破坏. 而在远离平衡态的非线性区间, 非平衡定态可能通过涨落和突变而产生宏观有序的耗散结构. 因而, 研究分子组装的非线性非平衡的热力学和动力学, 是当前非平衡理论发展的重要方向.

3.3.1 非平衡态热力学

非平衡态热力学不同于只关注始末状态的平衡态热力学, 需更关注其物理化学过程的时间进程, 因其系统的状态参量在空间和时间上往往是连续但非均匀的, 如细胞中各区域浓度随时间动态变化, 但其浓度仍是连续变化且无明确的浓度边界. 这使非平衡态理论中的热力学广延量 (如焓和熵等) 需定义为平衡态中的相应广延量的空间密度, 通过相应参数对时间和空间的求导获得. 通过实验获得这些参数的实际数值十分困难, 需通过足够小且响应迅速的测量探头在局域区间内精确测量获得, 这也为非平衡组装体系的设计与表征提出挑战. 同时, 非平衡过程是非线性的, 组装基元间由于非线性关联而引发协同效应, 使无序的组装基元转化为有序的耗散结构, 同时微小的环境扰动 (涨落) 可通过系统内的非线性过程而迅速放大, 使系统失稳或使组装结构快速转化而适应外界环境, 产生复杂的自适应现象. 这些特征对非平衡态热力学的理论建立提出十分苛刻的要求. 为将传统的平衡态热力学扩展至非平衡态体系, 人们建立了一些近似模型, 如局域平衡假设^[138,139]和扩展不可逆热力学^[140]等. 但非平衡热力学理论模型仍处于初期发展阶段, 还需要更多新概念、新模型和新理论充实和发展非平衡态热

力学理论.

熵及熵产生是非平衡态热力学中非常重要的概念^[139,141]. 在耗散组装循环中, Gibbs熵是衡量系统演化方向与过程的重要物理量, 其本质是信息熵的一种特殊表现形式 (4.1节), 其定义为^[133,141]:

$$S = -k_B \sum_i^n p_i \ln p_i \quad (12)$$

k_B 为波尔茨曼常数, p_i 表示第*i*物种的可能性, 其定义可见式 (7). 以Gibbs熵对时间*t*求导, 并应用式 (7), 可得Gibbs熵的变化率为:

$$\begin{aligned} \frac{dS}{dt} &= -k_B \sum_i^n \frac{dp_i}{dt} \ln p_i = f + \sigma \\ &= -\frac{k_B}{2} \sum_{i,j} (k_{ij} p_i - k_{ji} p_j) \ln \left(\frac{k_{ij}}{k_{ji}} \right) \\ &\quad + \frac{k_B}{2} \sum_{i,j} (k_{ij} p_i - k_{ji} p_j) \ln \left(\frac{k_{ij} p_i}{k_{ji} p_j} \right) \end{aligned} \quad (13)$$

将式 (13) 拆分成熵流 (f , entropy flow) 和熵产生 (σ , entropy production) 两项, 进一步应用式 (8)和(9), 可得:

$$f = -\frac{k_B}{2} \sum_{i,j} J_{ij} \ln \frac{k_{ji}}{k_{ij}} \quad (14)$$

$$\sigma = \frac{1}{2T} \sum_{i,j} J_{ij} \Delta G_{ij} \quad (15)$$

熵流*f*是由外部的热力学力与系统耦合而产生, 阻止系统达到平衡状态. σ 是系统随时间所产生的熵, 反映系统耗散能量或物质的能力, 是判断耗散结构生成与选择的重要指标. 如在非平衡凝胶组装过程中^[142], 熵产生可用于确定其动力学参数的最优值; Grzybowski等报道在相同实验条件下, 基于磁流体的非平衡体系^[143,144] (图10c) 存在两种平稳的耗散结构 (图10d), 分别对应低或高能量耗散状态. 他们发现两种状态的出现概率*P*与 σ 相关:

$$P(\sigma) \sim \exp(-k\sigma) \quad (16)$$

k 是拟合常数. 该研究提示, 虽然生命体系倾向于形成低能量耗散 (能量效率更高) 的结构, 但也有可能选择高能量耗散的结构. 然而, 目前熵产生与耗散结构的形成之间的联系仍不清晰, 一些理论认为维持非平衡态结构和提升非平衡过程的效率, 需最大化流入或穿过体系的能量, 以便形成较大的负熵流抵消体系的熵产生^[145]. 但其能量利用效率与熵产生间无相关性, 取决

于体系外部条件和特定结构的要求, 也可能与组装过程动力学和组装体的最佳功能相关^[146,147].

3.3.2 非平衡态动力学模型

当体系远离平衡态且处于非平衡的非线性区间时, 单凭纯粹的非平衡热力学将无法确定体系的发展方向, 需同时考虑其动力学的详细行为. 目前, MD方法、MC方法、动力学蒙特卡洛 (Kinetic Monte Carlo, KMC) 方法和Langevin方法等已分别成功应用于不同尺度的非平衡分子组装的动力学模拟中. 其中, MC方法适用于纳米尺度的非平衡动力学过程模拟^[148,149], 如扩散控制或反应控制的团簇-团簇组装过程^[150]. 然而, 更大尺度的非平衡动力学过程已远超MC方法的模拟范围. 而相较于传统的MC方法, KMC方法则更着眼于体系的组态变化, 而将模拟的时间跨度提升至组态跃迁的尺度. 因此, KMC方法更适用于模拟系统的演化过程^[151]及描述长时间尺度的非平衡动力学过程^[152], 如细胞中复杂的分子组装网络^[153]. 而在更大尺度的Langevin动力学模拟中, 可利用内部和外部作用力, 如静电、局部温度梯度、磁场、电场或温度场, 控制缺乏相互作用的粒子的分子组装过程. 目前Langevin动力学模拟已成功应用于分析活性粒子的随机运动^[154]及自驱动粒子形成不同形状结构^[155]等过程.

还应注意, 一些基于反应扩散方程和质量作用定律的动力学模型则可避免经典MC、MD和Langevin方法的缺陷, 如长计算时间、受限的时空间尺度和噪声项的使用^[148]. 虽然如反应扩散、磁-流体动力学和介观非平衡理论等动力学模型无法给出足够的耗散结构形成的微观信息, 它们却是建立微观和宏观动力学间联系的重要桥梁.

反应扩散模型 (reaction-diffusion model) 源于一系列确定性的非线性偏微分方程^[156], 适用于描述系统随时间演化而产生的动力学图纹结构 (图10a), 如Turing图^[157,158]和Liesegang结构^[159,160]等. Liesegang结构的形成可分为成核、生长和沉淀三个关键的基元步骤^[159]. 而在生长的决速步中, 颗粒的生长速率 R_s 可分为两部分:

$$R_s = \begin{cases} k(T)C_o C_s^2 & C_o < C_{lim} \\ k(T)C_o C_s^2 + k(T)C_e C_o (C_s - C_{lim}) & C_o \geq C_{lim} \end{cases} \quad (17)$$

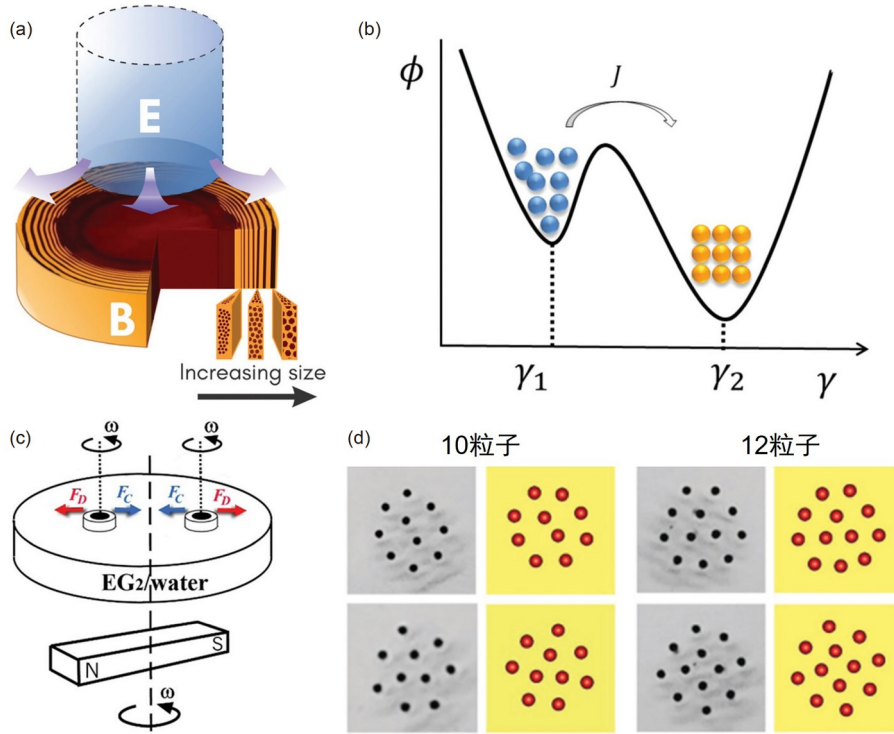


图 10 (a) 具备准周期沉淀环的Liesegang结构^[159]. 电解质E扩散到基质中并与电解质B组装, 形成介观Liesegang结构, 且结构尺寸随径向位置的变化而增大. (b) 介观非平衡热力学模型的势能面示意图. 在初始状态, 基元可被活化但呈无序状态, 而最终状态则呈有序结构^[148]. (c) 磁流体非平衡组装体系示意图^[144]. 外接磁力搅拌控制小磁体的旋转, 小磁体在气液界面运动. 红色箭头表示排斥力, 蓝色箭头表示吸引力. (d) 10个或12个小磁体在到达定态后, 分别呈现两种不同的排列图纹 (网络版彩图)

Figure 10 (a) Illustration of the Liesegang patterns consisting of quasi-periodic precipitation rings ^[159]. Electrolyte E diffuses into the matrix and interacts with electrolyte B to trigger the self-assembly of meso-structures of the Liesegang patterns. Structure size increases as a function of the radial position. (b) Illustration of potential energy landscape of mesoscopic non-equilibrium thermodynamic models ^[148]. In the initial state, the building blocks are already activated but disordered. The final state corresponds to the self-assembled structure. (c) Illustration of magneto-hydrodynamic assembly systems: an external magnet drives the rotations of small magnetic particles immersed in a viscous liquid at a liquid-air interface ^[144]. Repulsive forces are indicated by red arrows; the confining magnetic forces are indicated by blue arrows. (d) 10 or 12 small magnetic particles exhibit two kinds of arrangement modes in their steady states (color online).

$k(T)$ 是局域温度 T 下的速率常数. C_e 、 C_0 、 C_s 和 C_{lim} 分别为电解质浓度、核浓度、固体颗粒浓度和核的临界浓度, 这些参数共同决定Liesegang结构的生长模式.

磁流体动力学模型 (magneto-hydrodynamic models, MHD) 的提出是为解决磁场条件下的单分散磁体的非平衡组装问题. 近年来, MHD模型也被拓展至在外界能量驱动下的界面非平衡体系, 如气-液界面处的磁性粒子在旋转磁场下非平衡组装过程^[143,144] (图10c). Navier-Stokes方程是经典MHD模型的理论核心^[161,162], 该方程考虑由于磁体旋转而施加在流体上的力 f 和多体相互作用:

$$\rho \left[\frac{\partial \bar{u}}{\partial t} + \bar{u} \cdot \nabla \bar{u} \right] = -\nabla p + \mu \nabla^2 \bar{u} + \bar{f} \quad (18)$$

ρ 和 μ 分别为流体的密度和粘度, \bar{u} 为流体速率, \bar{f} 则可表示为:

$$\bar{f}_i = \sum_n F_i^{(n)} \Delta[x - y^{(n)}] + G_{ij}^{(n)} \quad (19)$$

N 为磁体数目, $y^{(n)}$ 、 $F^{(n)}$ 分别表示第 n 个磁体的位置和 在流体中受到的力, Δ 是有限尺寸磁体的密度函数. 式(19)中的第一项代表力磁单极子, $G_{ij}^{(n)}$ 则表示力偶极子^[162]. 该方法可用于研究在旋转磁场作用下, 流体表面的磁体分布规律, 并揭示磁体结构与能量耗散之间的关系, 有望为相关分子组装体系的研究提供思路.

介观非平衡模型 (mesoscopic non-equilibrium thermodynamic models, MNET) 研究非平衡体系在涨落情况下的非线性动力学, 描述系统状态量的概率密

度随时间的演化过程^[138,163]。在非平衡组装过程中, 组装基元需要外部能量方可跨越能垒, 从无序状态转化为有序组装状态。若系统状态的切换是经由连续构象转变, 则不同的构象可用反应坐标 γ 表示。此时, 系统从状态 γ_1 转变至 γ_2 的过程可类比为体系状态沿反应坐标, 以类似扩散的方式跨越势能面^[148] (图10b)。因此, 在多个状态转化的非平衡过程, 其局部熵产生 σ 可表示为:

$$\sigma(\gamma) = -\frac{1}{T} \sum_j J_j \frac{\partial \mu_j}{\partial \Gamma_j} \quad (20)$$

Γ 是分子组装发生的相空间, j 表示第 j 个基元过程, Γ_j 对应于第 j 个过程的反应坐标, J_j 表示对应的扩散流。而状态概率 $p(\gamma, t)$ 随时间演化的动力学过程可应用Fokker-Plank方程表述:

$$\frac{\partial p(\Gamma, t)}{\partial t} = \sum_j D_j \frac{\partial}{\partial \Gamma_j} \left[\frac{\partial p_j}{\partial \Gamma_j} + \frac{p_j}{k_B T} \frac{\partial \phi_j}{\partial \Gamma_j} \right] \quad (21)$$

ϕ_j 表示第 j 个过程的势能面, D_j 则为对应过程的扩散系数。目前, MNET模型已成功应用于分析蛋白质离子通道^[164]的输运过程和非平衡的成核^[165,166]等过程。

4 分子组装的物质、能量和信息维度的协同

目前, 基于物质与能量维度的分子组装研究已逐渐受到关注, 有望成为物质科学的前沿领域之一, 但似乎仅从物质和能量维度, 尚不足以搭建分子组装理论的全部框架。我们认为, 进一步引入信息维度, 将可能从全新的角度和更大格局, 探寻更深层次的分子组装机制^[167]。基于生命科学的启示, 分子组装过程可能涉及大量信息的储存、交换与反馈过程: 组装基元和溶剂分子中蕴含结构与作用位点等信息, 它们通过非共价相互作用构成复杂的信息网络, 如氢键阵列和给受体序列等, 实现信息交换与反馈, 同时周边环境也可以通过信息输入的方式调控该过程^[1,168,169]。尤其是生命科学中的分子组装体系将信息的作用发挥到极致, 如仅包含4种碱基的DNA序列通过碱基互补配对的形式, 可存储大量遗传信息; 细胞内的大量分子构成信息识别网络, 可辨明蛋白质表面的细微差异^[169]。当组装体系处于复杂的化学网络中, 大量的信息交换产生巨大的信息流, 微小的信息扰动通过信息流的反

馈而迅速放大, 显著改变组装体结构与功能。此时, 外部输入物质或能量对信息网络的调控或许尤为重要。

4.1 Clausius熵、Boltzmann熵和信息熵的概念辨析

1865年, Clausius 首次提出热力学第二定律^[170], 并引申出熵的概念及其计算标准:

$$dS_C \geq \frac{dQ}{T} \quad (22)$$

即体系在平衡态下, 始末状态间的熵增量 dS 大于等于体系热量差 dQ 与系统温度 T 之商, 因此Clausius熵 (S_C) 也被称为热温熵。然而, Clausius熵仅是对热力学熵的模糊描述。Boltzmann则根据统计力学提出熵与孤立系统微观状态数的关系^[171], 即

$$S_B = k_B \ln W \quad (23)$$

W 为与系统宏观状态所对应的微观状态数。Boltzmann熵 (S_B) 是对Clausius熵在统计学上的补足, 也奠定熵在统计力学方面的地位, 但Boltzmann熵也并非是对熵本质的理解。1948年, Shannon受热力学熵的启发, 提出“信息熵” (Information entropy, S_I)^[172], 因此也被称之为香农熵 (Shannon entropy), 其定义为:

$$S_I = -\sum_{i=1}^n p(x_i) \log_2 p(x_i) \quad (24)$$

其单位为bit, $p(x_i)$ 表示随机变量 x_i 出现的概率, 在二进制的信息科学中, 公式中的对数通常取2。信息熵是度量各种可能事件发生的不确定性, 是信息理论中的核心概念, 也是现代通信理论的基石, 被广泛应用于计算机领域和通信工程等领域, 但它们在分子科学领域的深入应用亟待开发。

Clausius熵、Boltzmann熵和信息熵在不同角度提出对熵的理解, 三者的适用范围也不尽相同^[173]。Clausius熵仅适用于平衡态热力学过程, 局限于粒子的热运动, 不适用于存在能量转换的非热力学过程。Boltzmann熵则进一步延拓Clausius熵的应用范围至非热力学系统和非平衡体系, 但也限制各种微观状态是等概率的。信息熵则不受等概率、平衡态和热力学系统的约束, 因此信息熵的适用范围应更宽泛。三者关系为信息熵 \supset Boltzmann熵 \supset Clausius熵。同时, 三者在一定情况下是等价且可相互转化。从定义上看, Boltzmann熵是信息熵的一种等概表示, 二者间的关系为:

$$S_B = k_B \ln 2 \cdot S_I \quad (25)$$

进一步根据式(22)则有:

$$k_B T \ln 2 \leq \frac{-Q}{S_I} \quad (26)$$

$-Q/S_I$ 即表示获取1 bit信息所需的能量. 1961年, Landauer根据式(26), 提出能量与信息间可相互转化, 并提出其定量关系: 消除1 bit信息熵需要至少 $\zeta_{\min} = k_B T \ln 2$ 的能量, 单位为 J bit^{-1} , 该数值也被称为Landauer极限^[174].

4.2 分子组装中的信息传输过程

香农信道是信息科学中的经典信息传输模型, 包含信源、发射器、信宿、接收器、信道及作用于信道上的噪音六大要素^[172,175,176] (图11a). 其中, 信源是产生信息的实体, 信宿是信息接收端, 信息通过信源编码转化并经由发射器送出, 并通过信道传输而达到接收器, 接收器经解码而将信息重现于信宿端. 整个过程中, 噪音不断干扰信息传输过程而使信息失真. 如前文所述, 分子组装过程似乎可类比于信息传输过程. 例如在生命体系中, “中心法则”是分子组装中信息储存、传递和处理的高级形式 (图11b), 展示基因到相应蛋白质的信息流途径. 遗传信息贮存在DNA中, 其所蕴含的信息量则可用信息熵表示, 如在非特定的无修饰DNA序列的单个位点中, 四个碱基是等概率出现, 因此应用式(24)可知DNA单个位点中的最大信息熵为2 bit. 信息经由转录而传递至mRNA, mRNA通过核膜进入细胞质到达核糖体, 通过遗传密码子解码成多肽序列. 在该过程中, 温度、溶剂及周围的溶质分子等可改变分子组装过程精准性的环境因素也许可视为于信号通道中的噪音. 这些环境因素可能扮演着重要角色, 既可造成组装体缺陷, 也可与物质、能量协同促进或调控组装过程. 因此, 如何建立数学模型, 度量系统获取信息的效率及环境因素的影响或许将成为理解复杂分子组装信息传输过程的关键之一.

在信息科学中, 信道容量 (C , 单位 bit s^{-1}) 是反映有扰信道传输能力的特征指标, 是指单位时间内在有扰信道中无差错传输的最大信息量, 即信道能够传输的最大平均信息速率. 香农信道的信道容量则可通过香农公式描述^[172,176]:

$$C = W \log_2 (1 + P/N) \quad (27)$$

W 是信道带宽, P 和 N 分别是信道内传输信号和高斯白

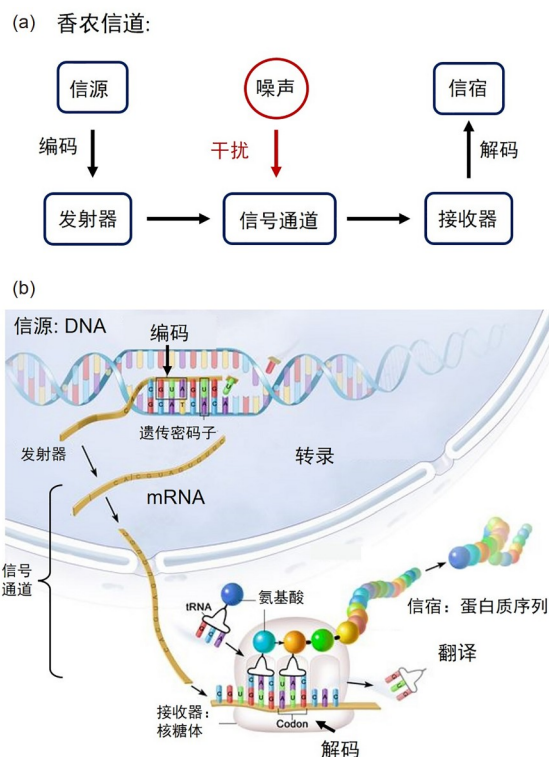


图11 (a) 香农信道示意图. (b) DNA转录及翻译过程可类比于香农信道 (网络版彩图)

Figure 11 (a) Scheme of the Shannon channel. (b) The analogy between the transcription and translating of DNA and the Shannon channel (color online).

噪声的平均功率, P/N 表示信噪比. 其物理意义为当信源的信息速率小于或者等于 C , 则可通过信息编码, 使信源的输出能以近无差错的概率通过信道传输.

在此基础上, Schneider拓展香农信道公式^[177~180], 将其应用于描述分子机器转换状态的过程. Schneider将分子机器所有可区分的状态设为一个完整的信息相空间, 分子机器在不同状态间转化即可视为在整个信息相空间中的检索与处理过程. 若信息相空间中仅包含两种等概率状态, 分子机器在每次状态转换过程中, 于二者中取其一, 则说明该过程获得1 bit的信息. 因此, 定义分子机器在某次状态转换过程中所能获得的最大信息量为“分子机器的信道容量 C_y ”:

$$C_y = d_{\text{space}} \log_2 (1 + P_y/N_y) \quad (\text{bit/次}) \quad (28)$$

d_{space} 表示分子机器可独立作用的有效位点. P_y 表示组装基元间特异性识别后释放的能量, N_y 表示高斯热噪声, 其值为 $k_B T d_{\text{space}}$. Schneider进一步定义该次状态转

化中, 消耗单位能量所获取信息量 ξ :

$$\xi = \frac{P_y}{C_y} \text{ (J bit}^{-1}\text{)} \quad (29)$$

显然, 当 $P_y \rightarrow 0$ 时, ξ 应逐渐逼近Landauer极限 (ξ_{\min}). 换言之, 当 P_y 固定, 分子机器的理论最大信道容量 $C_y(\max) = P_y/\xi_{\min}$. Schneider借鉴卫星通信的信息传输效率^[181], 定义分子机器的信息获取效率 η 为:

$$\eta = \frac{\xi_{\min}}{\xi} = \frac{C_y(\max)}{C_y} = \frac{\ln(1 + P_y/N_y)}{P_y/N_y} \times 100\% \quad (30)$$

以大肠杆菌中的EcoRI限制性内切酶特异性识别DNA中5'GAATTC3'序列为例^[182], 该过程分为两步(图12a): (1) EcoRI通过非特异性识别结合DNA; (2) EcoRI特异性识别5'GAATTC3'序列. 从能量上看, EcoRI处于较高能态, 在环境热噪声的驱动下, 通过不断“试错”(非特异性结合)的方式在DNA上检索. 当EcoRI检索到5'GAATTC3'序列, 通过特异性结合方式释放能量 P_y , 而处于稳定的较低能态, 此时只有施加足够大的环境热噪声才可将其与该位点解离. 因此, EcoRI在众多位点中选取特异性位点的过程, 即为释放能量获取信息的过程. 根据信息熵的可加性, 理论可知本次EcoRI特异性识别5'GAATTC3'序列中6个位点的实际所获的总信息熵(C_y)为12 bit. 该过程所释放能量 P_y 可直接通过实验量热法获得^[174], 则单次EcoRI在识别过程中可能获得的最大的信息量 $C_y(\max)=17.3$ bit. 因此根据式(30), 该过程的信息获取效率 η 为69.4%. 有趣的是, 目前实验所测得的其他DNA结合蛋白、细菌视紫红质和光敏视黄蛋白等的 η 均接近70%^[183-185], 而该效率已

超过目前人工信息传输技术的标准^[177].

为解释该现象, Schneider基于式(30), 发现当 $P_y > N_y$ 时, 随着 $P_y \rightarrow N_y$, η 则逐渐增大, 且当 $P_y = N_y$ 时达最大(69%), 与生命体系中的信息获取效率相当(图12b). 而当 P_y 进一步减小且小于 N_y 时, 分子机器的初始和终末状态可在环境热噪声的驱动下相互转化而不能区分, 因此无法有效地通过耗散能量而获取信息. 若温度上升, EcoRI很可能无法区分5'GAATTC3'及其近似序列5'TAATTC3', 使EcoRI可切割这两段相似序列, 而造成细菌自身DNA的灾难性损坏. 该研究指出, 历经长久的进化过程, 生命体系中的信息处理效率已达最优, 且过程中的能量损失是由于环境的热噪声导致, 而环境噪声也是导致基因突变和生物进化的主要诱因.

4.3 复杂的化学信息网络

具备复杂功能的分子组装系统通常是由多组分及相应的组装体所构成的多个可控网络耦合形成的多级次集合体. 在此类化学信息网络中, 组分之间、组分与组装体间及组装体之间通过化学反应或非键作用, 实现反馈和协同而体现出复杂功能. 然而, 目前所研究的人工分子组装体系通常包含有限的简单组分, 这些组分间缺乏反馈和协同, 难以形成可控的化学网络. 同时, 缺乏内在的信息处理过程的体系也难以实现复杂功能, 因此所得组装体的功能简单且单一. 而生命体系中的组装体系的强大功能很可能依赖于动态乃至具有长程作用的化学网络, 如生殖支原体细胞内存在28个网络^[186](图13), 各个子网络间相互耦合而产生复杂的生命现象. 尽管细胞网络看似错综复杂, 但可用有限

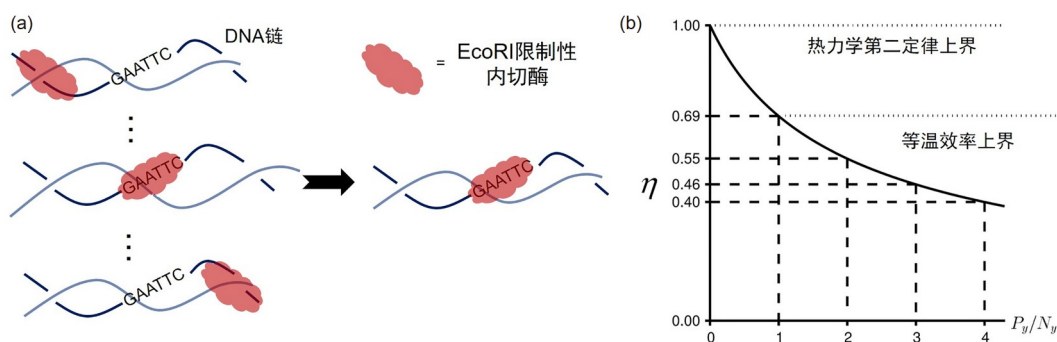


图12 (a) EcoRI可在DNA上, 以非特异性结合的方式滑动(左), 而当EcoRI特异性识别5'GAATTC3'后(右), 将释放能量并结合^[179]. (b) 等温分子机器的效率曲线^[177] (网络版彩图)

Figure 12 (a) EcoRI will non-specifically bind with a DNA molecule and scan along it. After specifically recognizing the sequence 5'GAATTC3', EcoRI will dissipate the energy and bind with the DNA molecule [179]. (b) Efficiency curve for isothermal molecular machines [177] (color online).

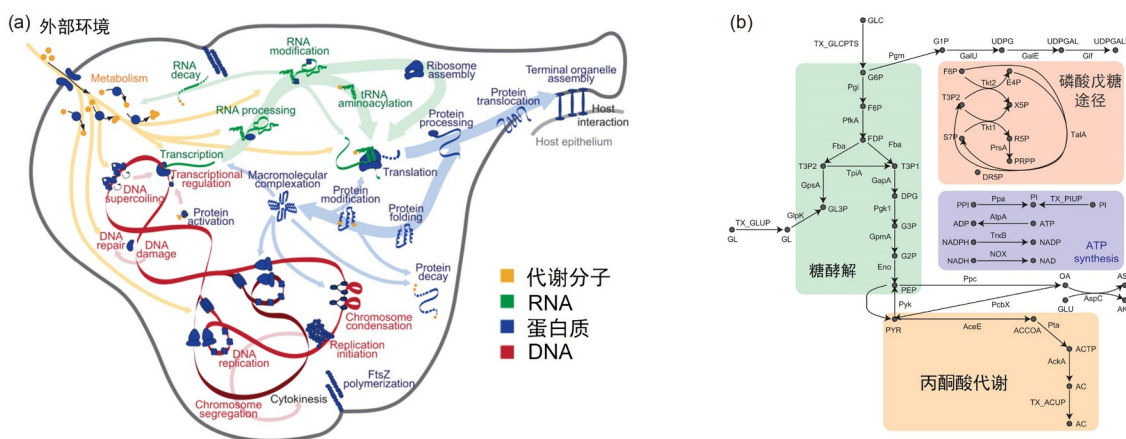


图 13 (a) 集成28个子模型的生殖支原体的全细胞模型^[186]. (b) 生殖支原体的简化网络模型图 (网络版彩图)
Figure 13 (a) The whole-cell model *M. genitalium* integrates 28 submodels^[186]. (b) The related simplified models for the networks in *M. genitalium* (color online).

数量的拓扑网络简化和描述其内在作用规律^[187,188], 如以节点替代这类动态网络的各组分, 以连线符号代表各节点间的促进、抑制和调节等信息反馈过程, 便可构建细胞内的信息反馈网络模型图 (图14和15). 探究化学信息网络的作用原理是理解细胞分化^[189]和生物节律^[190]等细胞活动机制的关键. 化学信息网络的典型特征是具有周期振荡性、时空间位置的精确性等, 如在有丝分裂的细胞周期中, 微管蛋白和附着在微管上的马达蛋白在准确的时间与空间位置, 组装形成具有

复杂结构的纺锤体, 辅助细胞分裂, 同时在细胞分裂末期发生解组装而完成细胞周期.

在化学信息网络中, 控制各节点间的反馈过程是实现这些特征的核心. 而“反馈”是控制论和信息论中的核心概念, 也可能是基于更高层级的“智能”控制

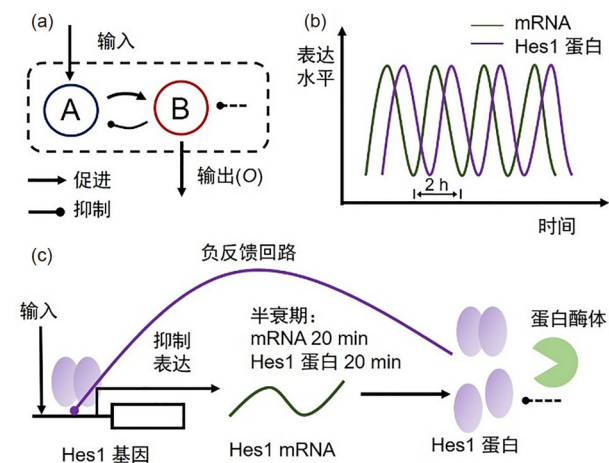


图 14 (a) 二元节点的反馈回路示意图. (b) 成纤维细胞中, Hes1 mRNA和Hes1蛋白呈周期振荡现象. (c) Hes1基因周期性表达的机理图^[195] (网络版彩图)
Figure 14 (a) Scheme of the feedback loop with two nodes. (b) The oscillation of Hes1 mRNA and Hes1 protein in fibroblasts. (c) The mechanism of periodic expression of Hes1 gene^[195] (color online).

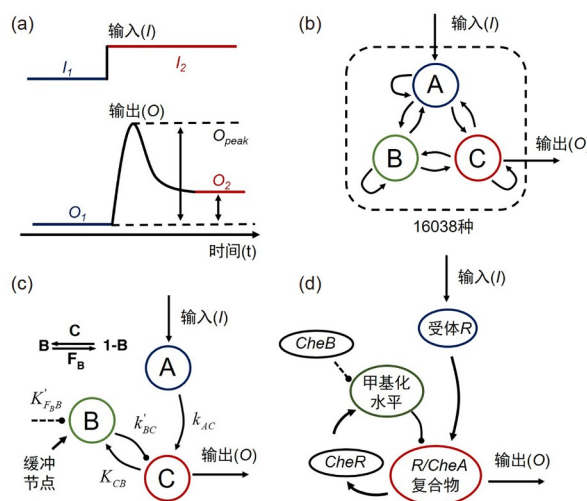


图 15 (a) 以输入 (I) 和输出 (O) 曲线定义自适应现象及其灵敏度和精确度^[187]. (b) 三元节点中所有可能的相互作用形式. (c) 一种具备高灵敏度自适应性能的、带缓冲节点的负反馈回路. (d) Barkai-Leibler模型揭示大肠杆菌的趋化性机理^[196] (网络版彩图)

Figure 15 (a) Defining the sensitivity and precision in self-adaptation based on input-output curves^[187]. (b) Possible interactions among three nodes. (c) A highly sensitive adaptive network composed of negative feedback loops with a buffer node. (d) Barkai-Leibler models for the chemotaxis mechanism of *E. coli*^[196] (color online).

系统的关键特征. 它使系统可对外界各种不确定性扰动产生响应, 并维持自身稳定性. 反馈机制分为正、负反馈两种类型, 正反馈可使系统对外界刺激快速响应, 但也容易使系统失衡, 而负反馈则主要使系统趋于稳定. 因此, 正、负反馈的协同可能帮助化学系统维持动态平衡, 也可能促使新体系从旧体系中生成而实现系统进化. 在不同类型的细胞中, 生化细节和进化历程使各细胞中的化学网络大相径庭, 但其可能具备相同的网络拓扑核心集^[191-193]. 因此, 构建统一的功能-拓扑映射是理解细胞相关功能的基础之一, 如二元节点是生物振荡最基础的反馈回路, 而三元节点则是自适应现象最核心的反馈回路.

4.3.1 二元节点反馈回路

生物振荡是指生物过程随时间呈现周期性变化的现象. 生物振荡在生物系统中的各个水平上均有体现, 是细胞内各种化学信息网络的重要特性. 而二元节点是实现振荡现象最简单的反馈回路^[194], 如以图14a所示的二元节点为例: 节点A对节点B为正反馈, 而节点B则对节点A为负反馈, 两个节点的动力学方程可描述为:

$$\begin{aligned} \dot{u}(t) &= \frac{\alpha_1 v^h(t - \tau_1)}{k_1^h + v^h(t - \tau_1)} - \beta_1 u(t) \\ \dot{v}(t) &= \frac{\alpha_2 k_2^h}{k_2^h + u^h(t - \tau_2)} - \beta_2 v(t) \end{aligned} \quad (31)$$

u 和 v 为两个节点的归一化浓度且 $u, v \in [0, 1]$, τ_1 和 τ_2 是信号传输过程的延迟时间, α_1 和 β_1 是正/负反馈分别对应的时间常数, h 为非线性Hill指数项. 通过改变 τ_1 或 τ_2 的值, 即可实现系统中节点的浓度的周期性动态变化, 实现振荡现象(图14b).

在胚胎干细胞基因中, Hes1基因^[195]是非常重要的转录抑制子, 包含一个与DNA结合的N端和一个招募协同抑制子的C端. Hes1基因是典型的信号受体, 接收信号后表达生成Hes1蛋白质, 通过抑制表达, 控制分化的靶基因, 调控胚胎干细胞的生理行为(图14c). 但Hes1蛋白质可进一步结合Hes1基因的活性端, 抑制其转录过程, 从而形成负反馈. Hes1蛋白被蛋白酶体降解, 所释放出Hes1基因的活性端, 使其可进行下一轮的转录过程, 实现周期振荡现象(图14b). 该过程也可简化成上述二元节点的反馈回路, 其振荡的周期在成

纤维细胞中约为2 h.

4.3.2 三元节点反馈回路

生物的自适应能力是指生物系统在外界信息输入的情况下, 可响应外界刺激并返回至预先水平的能力. 灵敏度(S)和精确度(P)是衡量自适应能力的两个重要指标(图15a):

$$\begin{aligned} S &= \left| \frac{(O_{\text{peak}} - O_1) / O_1}{(I_2 - I_1) / I_1} \right| \\ P &= \left| \frac{(O_2 - O_1) / O_1}{(I_2 - I_1) / I_1} \right|^{-1} \end{aligned} \quad (32)$$

I 和 O 分别代表输入和输出强度. 汤超等指出三元节点反馈回路可能是实现自适应现象的最少节点数^[187]: 节点A接收信号, 节点C负责信号输出, 节点B负责调节信号. 但即使仅存在三个节点, 节点间的相互作用可能导致16038种可能的组合方式(图15b). 以其中一种具备高灵敏度自适应性能的、带缓冲节点的负反馈回路为例, 其中节点A仅为信号从输入到节点C的简单中继, 节点B负责信号缓冲, 节点C对节点B为负反馈过程. 该信号通路的动力学过程可表示为:

$$\begin{aligned} \frac{dA}{dt} &= Ik_{IA} \frac{(1-A)}{(1-A) + K_{IA}} - F_A k'_{FAA} \frac{A}{A + K'_{FAA}} \\ \frac{dB}{dt} &= Ck_{CB} \frac{(1-B)}{(1-B) + K_{CB}} - F_B k'_{FBB} \frac{B}{B + K'_{FBB}} \\ \frac{dC}{dt} &= Ak_{AC} \frac{(1-C)}{(1-C) + K_{AC}} - Bk'_{BC} \frac{C}{C + K'_{BC}} \end{aligned} \quad (33)$$

F_A 和 F_B 为节点A和B中执行可逆反应的物质浓度, A, B, C 分别为各节点自身信号强度, I 为输入信号强度, K 为Michaelis-Menten常数, 如 K_{CB} 和 K'_{FBB} 分别表示节点C对B信号增长的促进和基础酶对B信号增长的抑制. 而仅在B节点中的物质浓度处于饱和浓度下, 如 $(1-B) \gg K_{CB}$ 和 $B \gg K'_{FBB}$, 该反馈回路方可实现自适应功能, 此时节点C最终信号输出的定态解 C^* 为:

$$C^* = F_B k'_{FBB} / k_{CB} \quad (34)$$

显然, 最终的信号响应 C^* 独立于输入信号 I , 即虽然开始时, 节点C会迅速对信号 I 做出响应, 但在节点B的缓冲下, 节点C的信号输出在一段时间后将稳定至与先前类似的水平, 也即信号通路为 $A \rightarrow C \rightarrow B \rightarrow C$. 该三元反馈回路是目前解释自适应现象的最广为接受的机

理, 也揭示具备自适应功能的信息网络的主要功能特征是维持一个独立于输入信号的稳态信号输出。

细菌的趋化性是生物自适应现象的宏观表现之一。其中, 大肠杆菌 (*E. coli*) 趋化性的细胞机制已被广泛研究, 其可用Barkai-Leibler模型描述^[196,197] (图15d)。在*E. coli*中, 受体R结合化学信号, 而组氨酸激酶CheA则会结合R形成复合物R/CheA, 将酶CheY磷酸化, 从而调节*E. coli*的鞭毛活性而产生趋化性。而R/CheA的活性受其甲基化水平调控, 因而又受到去甲基化酶CheB和甲基化酶CheR的调控。在Barkai-Leibler模型中, CheR在通常饱和浓度下维持恒定的甲基化速率, 且和R/CheA的甲基化水平无关, 而CheB只作用于活化的R/CheA, 而使去甲基化速率依赖于系统输出 (R/CheA的活性)。

5 总结与展望

分子组装现象普遍存在于各个层级和尺度的化学和生物体系中, 是创制复杂的功能材料、了解奇妙的生命现象的基石之一。因此, 分子组装领域的人们憧憬将分子组装建成联通化学与生命的桥梁, 并从数学、信息、物理和生物等领域引入诸多概念, 如耗散、非平衡态、信息、系统和进化等, 使分子组装领域历久弥新, 成为多学科汇聚的高地。迄今该领域的主流尚关注于较简单的人工分子组装体系, 这与生命体中的组装体系相比, 在可控性和功能性方面仍相去甚远。我们认为, 目前尚为初步且碎片化的分子组装理论框架及研究方法可能是使该领域发展较为迟缓的重要阻力之一。

本文根据分子组装理论的复杂度, 将研究对象从化学分子组装体系, 上升至简单的生物分子组装体系, 并期望迈向复杂的生物分子组装体系。进而在Anderson的“多即异”思想的指导下, 本文试图从物质、能量和信息三个维度及其协同的高度, 对这些分子组装体系进行分级分类。通过聚焦于三个层次上各自的分子组装过程及其机制, 探讨其理论基础与挑战, 寻找遵循客观规律、由简至繁地全面建立分子组装理论框架的正确方向和途径, 以期寻求突破分子组装发展瓶颈的思路乃至指导理论。然而, 在撰写本文的过程中, 我们愈发感到在庞大且盘根错节的分子组装理论体系中, 尤其是对生命体中的诸多复杂分子组装过程的微

观机制认识尚不清晰, 我们可能仅似盲人摸象般地观一隅而不知全貌。本文也尚未深入到亚细胞及以上的尺度的分子组装系统, 但包括生物分子组装过程在内的诸多复杂体系势必成为未来分子组装领域发展的重要方向之一。因此, 我们认为面向这类复杂分子组装体系的理论研究仍需从物质、能量和信息三个维度及其协同的高度, 分级分类地理清研究思路:

在物质维度, 针对动态的多种非键作用的协同, 需注重开发适合复杂体系、能均衡描述各种非键作用的组合方法, 更准确地描述组装基元浓度和外界条件 (如pH值、温度和溶剂等) 对组装过程和组装体结构的影响; 针对涉及多时空间尺度的模拟, 需发展准确的多尺度模型和隐式溶剂模型, 加强对组装体系的相空间取样, 提升相空间的搜索效率, 特别需要引入大数据和人工智能的相关工具, 对数据进行存储、管理和挖掘, 提升计算效率; 针对复杂环境中的传质过程, 需发展多尺度传质模型, 整合微观和宏观传质过程, 理解复杂分子组装体系传质以及传质-组装耦合机理。

近年来, 表征技术 (如冷冻电镜技术等) 的迅猛发展也为解析生命体中的重要生物大分子的组装过程及结构开启了一个新时代。如激活态次要剪接体的结构解析从分子层面揭示剪接体催化RNA剪接两步反应的工作机理^[198]。同时, 随着更多新算法的出现和算力的日益强大, 尤其是多种机器学习方法的兴起, 人工智能 (AI) 与分子组装领域的有机融合有望成为突破分子组装研究瓶颈的重要契机之一。例如, AlphaFold算法^[199]基于当前已知蛋白质的序列和结构数据集, 从而构建氨基酸分子间的非键作用与蛋白质片段的演化关系, 进而高成功率预测未知蛋白质的三级结构。当前, 一系列以深度学习为代表的AI技术也被应用于分子模型或力场的优化中, 有效提高分子模拟的计算能力。而AI技术和表征技术的相互交融也可进一步促进解析复杂生命组装过程及其组装体结构^[200-204], 如在冷冻电镜技术和AlphaFold的相辅相成下, 一系列跨空间尺度的核孔复合体被成功建构^[202,203]。这种人工智能、理论计算、表征技术和实验的交互模式也有望革新分子组装的研究范式, 拓展深化分子组装理论。

在物质和能量维度, 可能需要更重视熵效应在生物大分子及生命体系中的决定性作用, 分析分子组装过程的自由能, 选取合适的反应坐标并建立刻画自由能、定量分析不同阶段组装阶段的熵-焓贡献的有效

手段, 进而构建“强熵”效应的物理化学图像. 基于庞大且有待发展的非平衡分子组装的热力学和动力学理论, 本文仅较粗浅地提供小部分非平衡理论成功应用实例. 目前非平衡分子组装研究仍面临诸多挑战. 例如: 如何设计与表征非平衡组装体系, 并定量描述非平衡组装体系中的熵产生、能量转化、耗散结构稳定性等关键参数间的关系; 如何建立普适性的系统演化判据, 揭示系统初始条件的微小差异如何经非线性过程而放大, 促使不同耗散结构甚至“混沌 (chaos)”的产生; 如何寻找出构建抽象物化模型和直面真实的复杂生命体中的非平衡组装过程的平衡点.

在物质、能量加信息维度, 虽然我们试图将分子组装过程与信息论中的各要素作简单类比, 但这种类比仅为跳跃性的联想而非严格的推论, 因而需从更深层次的物理内涵揭示二者间的联系. 例如: 如何建立信息熵与组装基元的信息量间的定量分析方法; 生物信息如何在分子或组装体水平被编码或解码; 环境因素 (噪声) 如何直接或间接影响信息网络动力学^[205]和信号传导过程等. 同时, 信息反馈网络也将成为信息论应用于分子组装系统的另一主战场. 例如: 生命系统如何通过反馈、控制等机制, 实现信息流处理和交互, 进而涌现 (emergence) 或演化 (evolution) 出“智能”; 多个长距离的分子组装信息网络如何通讯而产生类似共振和高效的时空同步现象; 如何应用生物体中的通讯机制构建人工生物通讯系统, 或将自然生物通讯系统接入数字通讯系统. 更有趣的, 信息论将熵与状态的概率分布挂钩, 且认为可能存在一种隐形的“熵力”驱使动力学系统总是更经常从小概率状态往大概率状态移动^[206], 而如何辨析熵在生命系统中的可能的“主宰”地位也许成为理解生命现象的全新角度.

特别需要指出, 物质、能量和信息三个维度并非分立的, 反而越高层级的分子组装研究越需要考虑物质、能量和信息的耦合协同. 如在生命体中, 离子等物质可通过离子通道, 在能量驱动下, 集体性地定向运动而形成离子/分子超流体, 使信息在神经细胞和脑细胞间高效传递. 在该过程中, 离子通道的宏观量子态则可视作信息载体: 离子通道与它们所释放的光子的耦合可能产生环境波 (如脑波、声波), 其又可反过来调节通道中的离子振荡至相干态, 而退相干或一致性的状态可能分别对应于生命体睡眠或活动状态^[207]. 更全面深入理解物质、能量和信息如何从微观的协同机

制上升至宏观的群集效应, 也许可能成为解释生命现象的重要突破口, 如生命系统如何通过超低的能量消耗, 实现高效的能量转换和信息传输等过程.

更关键的是, 分子组装理论的发展需根据每个层级各自主导的规律, 构建与之相对应的描述符. 如在简单的物质层级, 可应用Arrhenius方程中的活化能作为阐明分子组装体系动力学的最重要的描述符. 但当分子组装体系涉及复杂多步骤的组装过程时, 该方程则难以全面、准确描述其动力学, 对熵影响的重视程度也严重不足. 而基于经典过渡态 (Transition State) 理论的Eyring方程则以活化自由能、活化焓和活化熵作为描述符, 分析每一步基元过程的微观动力学机制, 也补足了Arrhenius方程在熵方面的短板. 虽然经典的过渡态理论在低温或低能的情况下是严格的^[208], 但可能对一些关联性强、包含多步骤、能垒较低的复杂分子组装过程并不适用. 因其多个低能垒的步骤间易存在牵连关系, 即当某一基元过程能垒变化时, 将有可能牵一发而动全身, 显著改变多个过程的能垒乃至整体的势能面, 这可能导致Eyring方程产生较大偏差. 同时, Eyring方程也不适合有能量输入等非平衡过程. 因此, 当复杂度上升至物质和能量耦合协同的层级时, 将可能有该层级所特有的描述符. 如系统处于近平衡态的线性区间时, 线性不可逆热力学以熵产生作为关键描述符: 即当体系熵产生最小时, 判定系统处于非平衡定态^[209,210]. 然而, 当体系处于更复杂的非平衡非线性区间时, 目前尚无普适且准确的关系和方程, 用于揭示耗散结构的组装规律. 当分子组装理论进一步上升至物质、能量和信息大协同的更高层级时, 相关的理论研究迄今仍极度匮乏. 因此, 面向该层级的理论研究则需建立新模型, 考虑多位点弱键的协同反馈机制和多重过渡态, 方可探索物质、能量和信息间的相互作用机制, 并使用非平衡统计理论和非平衡信息热力学等手段, 构建与能量、时间、空间相关的全面深刻描述的关系式. 当考虑同一层级内或不同层级间的耦合协同机制, 寻找一些长程作用 (如长程范德华引力和静电作用力等) 与短程作用的耦合程度的描述符, 似乎成为揭示复杂分子组装过程和体系特性的必由之路.

还应当指出, 分子组装领域的主流思想可能仍受“自组装 (self-assembly)”概念中“自”的束缚. 复杂和高级的分子组装体系, 尤其是包括了能量输入和信息

交换的体系,基本不是自发的,而需要从更大的系统和尺度,在具体的组装环境中,外部介入高效调控手段。虽然目前共组装和场助组装等方式已可初步调控和加速组装过程,但目前尚难找到如化学合成方法学的手段,可高效调控组装过程。在化学历史进程中,催化概念的提出和应用,曾极大推进了化学合成科学和工业的发展。我们借鉴化学反应中的催化概念,提出了用于调控和加速分子组装过程的“催组装”新思路^[92,211,212]。催组剂是一类物质,它能在不改变总吉布斯自由能变化的条件下,加速和调控组装过程。我们认为,催组剂可能从物质、能量和信息三个维度,高效发挥其调控分子组装过程的作用。例如:在物质维度,催组剂可调控局域非键作用和长程相互作用,通过跨尺度协同机制,优化组装路径。同时,在复杂、拥挤和动态的组装环境中,可能存在一类催组剂在时间和空间上高效、精确地协助物质输运而提高传质效率,从而调控分子组装结构与功能。在物质加上能量的维度,催组剂可能在非平衡催组过程中,介导能量耗散和

转化,从而促进耗散结构的生成,并提升能量的利用效率。在物质加能量加信息的更高维度,催组剂可能在复杂信息网络中,通过介导反馈过程,调控正、负反馈速率而调节高级分子组装系统乃至生命体系的周期、功能及鲁棒性。我们有理由相信,随着催组概念的不断深入和催组应用的不拓展,催组剂的类型也将更加丰富,具有更复杂的作用机制。因为越复杂的组装过程越需要多种催组剂的协同。催组有望辅助揭示生命体系的运作机制,并发展相应的高效仿生和调控方法。我们将在该文的后续,另文全面阐述催组机理及物理化学图景,并详细提出并讨论其理论基础。

总之,分子组装理论研究面临众多艰巨挑战和重大发展机遇,若能引进并汇聚多学科、多领域的先进概念与研究方法,尤其是集成包括人工智能在内的信息科学,通过新颖的协同研究模式,有望为生命科学和软物质科学提供新视角和新方法,进而推动分子组装领域乃至整个物质科学的发展。

致谢 感谢国家自然科学基金委“催组研究方法 with 理论基础”重大项目(21991130, 21991131)、“可控分子自组装及其功能化”重大研究计划和国家自然科学基金项目(21971216, 21971217)的支持。衷心感谢张希、梁好均、吴家睿、郭雷、罗毅、刘鸣华、张文彬、王忻昌、樊春海、李国红、李伟、王鹏业、王梅祥、刘东生、丁松园、程俊、陈鸽等专家学者的有益讨论和宝贵建议。

参考文献

- 1 Lehn JM. *Angew Chem Int Ed*, 2013, 52: 2836–2850
- 2 Lehn JM. *Supramolecular Chemistry: Concepts and Perspectives*. Weinheim: VCH Publishers, 1995
- 3 Lehn JM. *Chem Soc Rev*, 2007, 36: 151–160
- 4 Whitesides GM, Mathias JP, Seto CT. *Science*, 1991, 254: 1312–1319
- 5 Lehn JM. *Angew Chem Int Ed*, 2015, 54: 3276–3289
- 6 Collier CP, Wong EW, Belohradsky M, Raymo FM, Stoddart JF, Kuekes PJ, Williams RS, Heath JR. *Science*, 1999, 285: 391–394
- 7 Erbas-Cakmak S, Leigh DA, McTernan CT, Nussbaumer AL. *Chem Rev*, 2015, 115: 10081–10206
- 8 Saha S, Stoddart JF. *Chem Soc Rev*, 2007, 36: 77–92
- 9 Browne WR, Feringa BL. *Nat Nanotech*, 2006, 1: 25–35
- 10 Cordier P, Tournilhac F, Soulié-Ziakovic C, Leibler L. *Nature*, 2008, 451: 977–980
- 11 Burnworth M, Tang L, Kumpfer JR, Duncan AJ, Beyer FL, Fiore GL, Rowan SJ, Weder C. *Nature*, 2011, 472: 334–337
- 12 Moore PB. *Annu Rev Biophys*, 2012, 41: 1–19
- 13 Karsenti E. *Nat Rev Mol Cell Biol*, 2008, 9: 255–262
- 14 Mitchison TJ, Field CM. *Annu Rev Cell Dev Biol*, 2021, 37: 23–41
- 15 Van der Waals J. *Over de continuïteit van den gasen vloeïstoftoestand (On the Continuity of the Gas and Liquid State)*. The Netherlands: Universiteit Leiden, 1873
- 16 Fischer E. *Ber Dtsch Chem Ges*, 1894, 27: 2985–2993

- 17 Moore TS, Winmill TF. *J Chem Soc Trans*, 1912, 101: 1635–1676
- 18 Guthrie F. *J Chem Soc*, 1863, 16: 239–244
- 19 Mulliken RS. *J Am Chem Soc*, 1950, 72: 600–608
- 20 Kauzmann W. *Adv in Protein Chem*, 1959, 14: 1–63
- 21 Chandler D. *Nature*, 2002, 417: 491
- 22 Pedersen CJ. *J Am Chem Soc*, 1967, 89: 7017–7036
- 23 Dietrich B, Lehn JM, Sauvage JP. *Tetrahedron Lett*, 1969, 10: 2885–2888
- 24 Newcomb M, Cram DJ. *J Am Chem Soc*, 1975, 97: 1257–1259
- 25 Lehn JM. *Angew Chem Int Ed Engl*, 1988, 27: 89–112
- 26 Bain CD, Troughton EB, Tao YT, Evall J, Whitesides GM, Nuzzo RG. *J Am Chem Soc*, 1989, 111: 321–335
- 27 Nuzzo RG, Allara DL. *J Am Chem Soc*, 1983, 105: 4481–4483
- 28 Prins LJ, Timmerman P, Reinhoudt DN. *Pure Appl Chem*, 1998, 70: 1459–1468
- 29 Dietrich-Buchecker CO, Sauvage JP. *Angew Chem Int Ed Engl*, 1989, 28: 189–192
- 30 Danon JJ, Krüger A, Leigh DA, Lemonnier JF, Stephens AJ, Vitorica-Yrezabal IJ, Woltering SL. *Science*, 2017, 355: 159–162
- 31 Leigh DA, Danon JJ, Fielden SDP, Lemonnier JF, Whitehead GFS, Woltering SL. *Nat Chem*, 2020, 13: 117–122
- 32 Yu Z, Tantakitti F, Yu T, Palmer LC, Schatz GC, Stupp SI. *Science*, 2016, 351: 497–502
- 33 Huang M, Hsu CH, Wang J, Mei S, Dong X, Li Y, Li M, Liu H, Zhang W, Aida T, Zhang WB, Yue K, Cheng SZD. *Science*, 2015, 348: 424–428
- 34 Dey S, Fan C, Gothelf KV, Li J, Lin C, Liu L, Liu N, Nijenhuis MAD, Saccà B, Simmel FC, Yan H, Zhan P. *Nat Rev Methods Primers*, 2021, 1: 13–37
- 35 Rothmund PWK. *Nature*, 2006, 440: 297–302
- 36 Liu X, Zhang F, Jing X, Pan M, Liu P, Li W, Zhu B, Li J, Chen H, Wang L, Lin J, Liu Y, Zhao D, Yan H, Fan C. *Nature*, 2018, 559: 593–598
- 37 Qian L, Wang Y, Zhang Z, Zhao J, Pan D, Zhang Y, Liu Q, Fan CH, Hu J, He L. *Chin Sci Bull*, 2006, 51: 2860–2863 (in Chinese) [钱璐璐, 汪颖, 张钊, 赵健, 潘敦, 张益, 刘强, 樊春海, 胡钧, 贺林. 科学通报, 2006, 51: 2860–2863]
- 38 Dong Y, Yang YR, Zhang Y, Wang D, Wei X, Banerjee S, Liu Y, Yang Z, Yan H, Liu D. *Angew Chem Int Ed*, 2017, 56: 1586–1589
- 39 Dong Y, Sun Y, Wang L, Wang D, Zhou T, Yang Z, Chen Z, Wang Q, Fan Q, Liu D. *Angew Chem Int Ed*, 2014, 53: 2607–2610
- 40 Dong Y, Yang Y, Lin C, Liu D. *Acc Chem Res*, 2022, 55: 1938–1948
- 41 Li XB, Gao YJ, Wang Y, Zhan F, Zhang XY, Kong QY, Zhao NJ, Guo Q, Wu HL, Li ZJ, Tao Y, Zhang JP, Chen B, Tung CH, Wu LZ. *J Am Chem Soc*, 2017, 139: 4789–4796
- 42 Li S, Jiang Q, Liu S, Zhang Y, Tian Y, Song C, Wang J, Zou Y, Anderson GJ, Han JY, Chang Y, Liu Y, Zhang C, Chen L, Zhou G, Nie G, Yan H, Ding B, Zhao Y. *Nat Biotechnol*, 2018, 36: 258–264
- 43 Huang P, Wang D, Su Y, Huang W, Zhou Y, Cui D, Zhu X, Yan D. *J Am Chem Soc*, 2014, 136: 11748–11756
- 44 Hou X, Zaks T, Langer R, Dong Y. *Nat Rev Mater*, 2021, 6: 1078–1094
- 45 Whitelam S, Jack RL. *Annu Rev Phys Chem*, 2015, 66: 143–163
- 46 Truhlar DG, Garrett BC. *Annu Rev Phys Chem*, 1984, 35: 159–189
- 47 Hänggi P, Talkner P, Borkovec M. *Rev Mod Phys*, 1990, 62: 251–341
- 48 Becker R, Döring W. *Ann Phys*, 1935, 416: 719–752
- 49 Stranski IN, Totomanow D. *Z für Physikalische Chem*, 1933, 163A: 399–408
- 50 Duff N, Peters B. *J Chem Phys*, 2009, 131: 184101–184108
- 51 Maibaum L. *Phys Rev Lett*, 2008, 101: 256102–256104
- 52 Odriozola G, Lozada-Cassou M. *Phys Rev Lett*, 2013, 110: 105701–105705
- 53 Gao T, Blackwell R, Glaser MA, Betterton MD, Shelley MJ. *Phys Rev E*, 2015, 92: 062709
- 54 Romano F, Sciortino F. *Nat Commun*, 2012, 3: 975–981
- 55 Chen Q, Bae SC, Granick S. *Nature*, 2011, 469: 381–384
- 56 Romano F, Sciortino F. *Soft Matter*, 2011, 7: 5799–5804
- 57 De Greef TFA, Smulders MMJ, Wolfs M, Schenning APHJ, Sijbesma RP, Meijer EW. *Chem Rev*, 2009, 109: 5687–5754
- 58 Korevaar PA, de Greef TFA, Meijer EW. *Chem Mater*, 2013, 26: 576–586

- 59 Wang Y, Fang H, Tranca I, Qu H, Wang X, Markvoort AJ, Tian Z, Cao X. *Nat Commun*, 2018, 9: 488–496
- 60 Qu H, Huang Z, Dong X, Wang X, Tang X, Li Z, Gao W, Liu H, Huang R, Zhao Z, Zhang H, Yang L, Tian Z, Cao X. *J Am Chem Soc*, 2020, 142: 16223–16228
- 61 Qu H, Wang Y, Li Z, Wang X, Fang H, Tian Z, Cao X. *J Am Chem Soc*, 2017, 139: 18142–18145
- 62 Fezoui Y, Teplow DB. *J Biol Chem*, 2002, 277: 36948–36954
- 63 Pallitto MM, Murphy RM. *Biophys J*, 2001, 81: 1805–1822
- 64 Anderson PW. *Science*, 1972, 177: 393–396
- 65 Toprak E, Yildiz A, Hoffman MT, Rosenfeld SS, Selvin PR. *Proc Natl Acad Sci USA*, 2009, 106: 12717–12722
- 66 Bressloff PC, Newby JM. *Rev Mod Phys*, 2013, 85: 135–196
- 67 Su NQ, Xu X. *Sci Sin-Chim*, 2013, 43: 1761–1779 (in Chinese) [苏乃强, 徐昕. 中国科学: 化学, 2013, 43: 1761–1779]
- 68 Zhang Y, Xu X, Goddard William AI. *Proc Natl Acad Sci USA*, 2009, 106: 4963–4968
- 69 Guo W, Wu A, Xu X. *Chem Phys Lett*, 2010, 498: 203–208
- 70 Guo W, Wu A, Zhang IY, Xu X. *J Comput Chem*, 2012, 33: 2142–2160
- 71 Liu Z, Chen Z, Xi J, Xu X. *Nat Sci Rev*, 2020, 7: 1036–1045
- 72 Srinivasan B, Kantae V, Robinson J. *Med Res Rev*, 2020, 40: 1776–1793
- 73 Daniel RM, Peterson ME, Danson MJ, Price NC, Kelly SM, Monk CR, Weinberg CS, Oudshoorn ML, Lee CK. *Biochem J*, 2009, 425: 353–360
- 74 Tantakitti F, Boekhoven J, Wang X, Kazantsev RV, Yu T, Li J, Zhuang E, Zandi R, Ortony JH, Newcomb CJ, Palmer LC, Shekhawat GS, de la Cruz MO, Schatz GC, Stupp SI. *Nat Mater*, 2016, 15: 469–476
- 75 Pauling L. *Proc Natl Acad Sci USA*, 1935, 21: 186–191
- 76 Ortega G, Mariottini D, Troina A, Dahlquist FW, Ricci F, Plaxco KW. *Proc Natl Acad Sci USA*, 2020, 117: 19136–19140
- 77 Gershberg J, Fennel F, Rehm TH, Lochbrunner S, Würthner F. *Chem Sci*, 2016, 7: 1729–1737
- 78 Whitesides GM, Grzybowski B. *Science*, 2002, 295: 2418–2421
- 79 Ingólfsson HI, Arnarez C, Periole X, Marrink SJ. *J Cell Sci*, 2016, 129: 258–268
- 80 Zhang J, Lei YK, Yang YI, Gao YQ. *J Chem Phys*, 2020, 153: 174115
- 81 Vetter T, Iggländ M, Ochsenbein DR, Hänseler FS, Mazzotti M. *Cryst Growth Des*, 2013, 13: 4890–4905
- 82 Cramer CJ, Truhlar DG. *Chem Rev*, 1999, 99: 2161–2200
- 83 Qi CS, Deng XW, Li W. *Journal of Beijing Institute of Petro-chemical Technology*, 2006, 14: 30–33 (in Chinese) [戚传松, 邓兴旺, 李巍. 北京石油化工学院学报, 2006, 14: 30–33]
- 84 Gao J, Thompson MA. *Combined Quantum Mechanical and Molecular Mechanical Models*. Washington DC: American Chemical Society, 1998
- 85 Wong MW, Wiberg KB, Frisch M. *J Chem Phys*, 1991, 95: 8991–8998
- 86 Dhar A, Samiotakis A, Ebbinghaus S, Nienhaus L, Homouz D, Gruebele M, Cheung MS. *Proc Natl Acad Sci USA*, 2010, 107: 17586–17591
- 87 Zimmerman SB, Minton AP. *Annu Rev Biophys Biomol Struct*, 1993, 22: 27–65
- 88 Baltierra-Jasso LE, Morten MJ, Laflör L, Quinn SD, Magennis SW. *J Am Chem Soc*, 2015, 137: 16020–16023
- 89 Chen Y, Krämer A, Charron NE, Husic BE, Clementi C, Noé F. *J Chem Phys*, 2021, 155: 084101
- 90 Basdogan Y, Groenenboom MC, Henderson E, De S, Rempe SB, Keith JA. *J Chem Theor Comput*, 2019, 16: 633–642
- 91 Alibakhshi A, Hartke B. *Nat Commun*, 2021, 12: 3584–3591
- 92 Lei ZC, Wang XC, Qu H, Zhou C, Li ZH, Yang LL, Cao XY, Tian ZQ. *Sci Sin-Chim*, 2020, 50: 1781–1800 (in Chinese) [雷志超, 王忻昌, 瞿航, 周岑, 李之豪, 杨柳林, 曹晓宇, 田中群. 中国科学: 化学, 2020, 50: 1781–1800]
- 93 Fick A. *Ann Phys Chem*, 1855, 170: 59–86
- 94 Höfling F, Franosch T. *Rep Prog Phys*, 2013, 76: 046602
- 95 Huang YR, Zheng Xu, Hu GQ. *Physics of Gases*, 2018, 3: 1–12 (in Chinese) [黄艺荣, 郑旭, 胡国庆. 气体物理, 2018, 3: 1–12]
- 96 Dix JA, Verkman AS. *Annu Rev Biophys*, 2008, 37: 247–263
- 97 Matthäus F, Jagodic M, Dobnikar J. *Biophys J*, 2009, 97: 946–957
- 98 Zheng X, Ten Hagen B, Kaiser A, Wu M, Cui H, Silber-Li Z, Löwen H. *Phys Rev E*, 2013, 88: 032304
- 99 Ochab-Marcinek A, Wiczorek SA, Ziębacz N, Hołyst R. *Soft Matter*, 2012, 8: 11173–11179
- 100 Yeung C, Shtrahman M, Wu X. *Biophys J*, 2007, 92: 2271–2280

- 101 Hyman AA, Weber CA, Jülicher F. *Annu Rev Cell Dev Biol*, 2014, 30: 39–58
- 102 Bergeron-Sandoval LP, Safaee N, Michnick SW. *Cell*, 2016, 165: 1067–1079
- 103 Banani SF, Lee HO, Hyman AA, Rosen MK. *Nat Rev Mol Cell Biol*, 2017, 18: 285–298
- 104 Chen P, Ma X, Hu K, Rong Y, Liu M. *Chem Eur J*, 2011, 17: 12108–12114
- 105 Sun J, Li Y, Yan F, Liu C, Sang Y, Tian F, Feng Q, Duan P, Zhang L, Shi X, Ding B, Liu M. *Nat Commun*, 2018, 9: 2599–2607
- 106 Ding H, Jiang H, Zhao N, Hou Z. *Phys Rev E*, 2017, 95: 012121
- 107 Brangwynne CP, Eckmann CR, Courson DS, Rybarska A, Hoege C, Gharakhani J, Jülicher F, Hyman AA. *Science*, 2009, 324: 1729–1732
- 108 Li P, Banjade S, Cheng HC, Kim S, Chen B, Guo L, Llaguno M, Hollingsworth JV, King DS, Banani SF, Russo PS, Jiang QX, Nixon BT, Rosen MK. *Nature*, 2012, 483: 336–340
- 109 Franzmann TM, Jahnel M, Pozniakovsky A, Mahamid J, Holehouse AS, Nüske E, Richter D, Baumeister W, Grill SW, Pappu RV, Hyman AA, Alberti S. *Science*, 2018, 359: eaao5654
- 110 Shin Y, Brangwynne CP. *Science*, 2017, 357: eaaf4382
- 111 Zwicker D, Decker M, Jaensch S, Hyman AA, Jülicher F. *Proc Natl Acad Sci USA*, 2014, 111: 2636–2645
- 112 Yuan C, Levin A, Chen W, Xing R, Zou Q, Herling TW, Challa PK, Knowles TPJ, Yan X. *Angew Chem Int Ed*, 2019, 58: 18116–18123
- 113 Mitrea DM, Mittasch M, Gomes BF, Klein IA, Murcko MA. *Nat Rev Drug Discov*, 2022, 21: 841–862
- 114 Wang B, Zhang L, Dai T, Qin Z, Lu H, Zhang L, Zhou F. *Sig Transduct Target Ther*, 2021, 6: 290
- 115 van Rossum SAP, Tena-Solsona M, van Esch JH, Eelkema R, Boekhoven J. *Chem Soc Rev*, 2017, 46: 5519–5535
- 116 Fang X, Wang J. *Annu Rev Biophys*, 2020, 49: 227–246
- 117 Li C, Wang J. *Proc Natl Acad Sci USA*, 2014, 111: 14130–14135
- 118 Lemus EH. *J Non-Equilibrium ThermoDyn*, 2009, 34: 371–394
- 119 Qian H. *J Phys Chem B*, 2006, 110: 15063–15074
- 120 Qian H, Roy S. *IEEE Transon NanoBiosci*, 2012, 11: 289–295
- 121 Kassem S, van Leeuwen T, Lubbe AS, Wilson MR, Feringa BL, Leigh DA. *Chem Soc Rev*, 2017, 46: 2592–2621
- 122 Wilson MR, Solà J, Carlone A, Goldup SM, Lebrasseur N, Leigh DA. *Nature*, 2016, 534: 235–240
- 123 Merindol R, Walther A. *Chem Soc Rev*, 2017, 46: 5588–5619
- 124 England JL. *Nat Nanotech*, 2015, 10: 919–923
- 125 Weng J, Yang M, Wang W, Xu X, Tian Z. *J Am Chem Soc*, 2020, 142: 21344–21352
- 126 Odrizola G, Jiménez-Angeles F, Lozada-Cassou M. *J Chem Phys*, 2008, 129: 111101
- 127 Whitesides GM, Simanek EE, Mathias JP, Seto CT, Chin D, Mammen M, Gordon DM. *Acc Chem Res*, 1995, 28: 37–44
- 128 Schwierz N, Frost CV, Geissler PL, Zacharias M. *J Am Chem Soc*, 2016, 138: 527–539
- 129 Dai XB, Zhang XY, Gao LJ, Yan LT. *Acta Polymerica Sinica*, 2021, 52: 1076–1099 (in Chinese) [戴晓彬, 张轩钰, 高丽娟, 燕立唐. 高分子学报, 2021, 52: 1076–1099]
- 130 Zhu G, Huang Z, Xu Z, Yan LT. *Acc Chem Res*, 2018, 51: 900–909
- 131 Ragazzon G, Prins LJ. *Nat Nanotech*, 2018, 13: 882–889
- 132 Schnakenberg J. *Rev Mod Phys*, 1976, 48: 571–585
- 133 Ge H, Qian M, Qian H. *Phys Rep*, 2012, 510: 87–118
- 134 Hess H, Ross JL. *Chem Soc Rev*, 2017, 46: 5570–5587
- 135 Murata T, Sonobe S, Baskin TI, Hyodo S, Hasezawa S, Nagata T, Horio T, Hasebe M. *Nat Cell Biol*, 2005, 7: 961–968
- 136 Hashimoto T. *Curr Opin Plant Biol*, 2013, 16: 698–703
- 137 Bowne-Anderson H, Zanic M, Kauer M, Howard J. *BioEssays*, 2013, 35: 452–461
- 138 Vilar JMG, Rubi JM. *Proc Natl Acad Sci USA*, 2001, 98: 11081–11084
- 139 Keizer J. *Statistical Thermodynamics of Nonequilibrium Processes*. New York: Springer-Verlag, 1987
- 140 Jou D, Lebo G. *Extended Irreversible Thermodynamics*. New York: Springer, 2010
- 141 Lebon G, Jou D, Casas-Vázquez J. *Understanding Non-equilibrium Thermodynamics: Foundations, Applications, Frontiers*. Berlin: Springer, 2008
- 142 Head DA, Briels WJ, Gompper G. *Phys Rev E*, 2014, 89: 032705

- 143 Grzybowski BA, Stone HA, Whitesides GM. *Nature*, 2000, 405: 1033–1036
- 144 Tretiakov KV, Szleifer I, Grzybowski BA. *Angew Chem Int Ed*, 2013, 52: 10304–10308
- 145 Bejan A, Lorente S. *J Appl Phys*, 2013, 113: 151301–151320
- 146 Hubler A. *Complexity*, 2014, 19: 6–7
- 147 Hubler A, Belkin A, Bezryadin A. *Complexity*, 2015, 20: 8–11
- 148 Arango-Restrepo A, Barragán D, Rubi JM. *Phys Chem Chem Phys*, 2019, 21: 17475–17493
- 149 Dudowicz J, Freed KF, Douglas JF. *Phys Rev Lett*, 2004, 92: 045502
- 150 Meakin P, Vicsek T, Family F. *Phys Rev B*, 1985, 31: 564–569
- 151 Serebrinsky SA. *Phys Rev E*, 2011, 83: 037701
- 152 Meng B, Weinberg WH. *J Chem Phys*, 1994, 100: 5280–5289
- 153 Slepoy A, Thompson AP, Plimpton SJ. *J Chem Phys*, 2008, 128: 205101
- 154 Romanczuk P, Bär M, Ebeling W, Lindner B, Schimansky-Geier L. *Eur Phys J Spec Top*, 2012, 202: 1–162
- 155 Vicsek T, Czirók A, Ben-Jacob E, Cohen I, Shochet O. *Phys Rev Lett*, 1995, 75: 1226–1229
- 156 Cross MC, Hohenberg PC. *Rev Mod Phys*, 1993, 65: 851–1112
- 157 Serna H, Muñozuri AP, Barragán D. *Phys Chem Chem Phys*, 2017, 19: 14401–14411
- 158 Stoop N, Lagrange R, Terwagne D, Reis PM, Dunkel J. *Nat Mater*, 2015, 14: 337–342
- 159 Arango-Restrepo A, Barragán D, Rubi JM. *Phys Chem Chem Phys*, 2018, 20: 4699–4707
- 160 Lagzi I. *Phys Chem Chem Phys*, 2002, 4: 1268–1270
- 161 Tretiakov KV, Bishop KJM, Grzybowski BA. *Soft Matter*, 2009, 5: 1279
- 162 Climent E, Yeo K, Maxey MR, Karniadakis GE. *J Fluids Eng*, 2007, 129: 379–387
- 163 Reguera D, Rubi JM, Vilar JMG. *J Phys Chem B*, 2005, 109: 21502–21515
- 164 Rubí JM, Lervik A, Bedeaux D, Kjelstrup S. *J Chem Phys*, 2017, 146: 185101
- 165 Reguera D, Rubi JM. *J Chem Phys*, 2003, 119: 9877–9887
- 166 Reguera D, Rubi JM. *J Chem Phys*, 2003, 119: 9888–9893
- 167 Amano S, Esposito M, Kreidt E, Leigh DA, Penocchio E, Roberts BMW. *Nat Chem*, 2022, 14: 530–537
- 168 Lehn JM. *Proc Natl Acad Sci USA*, 2002, 99: 4763–4768
- 169 Lehn JM. *Science*, 2002, 295: 2400–2403
- 170 Clausius R. *Ann Phys Chem*, 1865, 201: 353–400
- 171 Boltzmann L. *Vorlesungen über Gastheorie*. Leipzig: Johann Ambrosius Barth, 1896
- 172 Shannon CE. *Bell Syst Technical J*, 1948, 27: 379–423
- 173 Li HL. *College Phys*, 2004, 23: 37–40 (in Chinese) [李鹤龄. 大学物理, 2004, 23: 37–40]
- 174 Landauer R. *IBM J Res Dev*, 1961, 5: 183–191
- 175 Shannon CE. *Proc IRE*, 1949, 37: 10–21
- 176 Shannon CE, Weaver W. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press, 1949
- 177 Schneider TD. *Nucleic Acids Res*, 2010, 38: 5995–6006
- 178 Schneider TD. *J Theor Biol*, 1991, 148: 125–137
- 179 Schneider TD. *J Theor Biol*, 1991, 148: 83–123
- 180 Schneider TD. *Nano Communication Networks*, 2010, 1: 173–180
- 181 Rausbeck G. *Information Theory*. Cambridge: Massachusetts Institute of Technology, 1963
- 182 Heitman J, Zinder ND, Model P. *Proc Natl Acad Sci USA*, 1989, 86: 2281–2285
- 183 Dartnall HJA. *Vision Res*, 1968, 8: 339–358
- 184 Schneider G, Diller R, Stockburger M. *Chem Phys*, 1989, 131: 17–29
- 185 Meyer TE, Tollin G, Hazzard JH, Cusanovich MA. *Biophys J*, 1989, 56: 559–564
- 186 Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival Jr. B, Assad-Garcia N, Glass JI, Covert MW. *Cell*, 2012, 150: 389–401
- 187 Ma W, Trusina A, El-Samad H, Lim WA, Tang C. *Cell*, 2009, 138: 760–773
- 188 Li Z, Liu S, Yang Q. *Cell Syst*, 2017, 5: 72–81.e4

- 189 Brandman O, Ferrell Jr. JE, Li R, Meyer T. *Science*, 2005, 310: 496–498
- 190 Gregson N, Sparrowhawk K, Mauskopf J, Paul J. *Nat Rev Drug Discov*, 2005, 4: 121–130
- 191 Arkin AP, Hornung G, Barkai N. *PLoS Comput Biol*, 2008, 4: 0055–0061
- 192 Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. *Science*, 2002, 298: 824–827
- 193 Wagner A. *Proc Natl Acad Sci USA*, 2005, 102: 11775–11780
- 194 Klemm K, Bornholdt S. *Proc Natl Acad Sci USA*, 2005, 102: 18414–18419
- 195 Kobayashi T, Kageyama R. *Genes*, 2011, 2: 219–228
- 196 Barkai N, Leibler S. *Nature*, 1997, 387: 913–917
- 197 Yi TM, Huang Y, Simon MI, Doyle J. *Proc Natl Acad Sci USA*, 2000, 97: 4649–4653
- 198 Bai R, Wan R, Wang L, Xu K, Zhang Q, Lei J, Shi Y. *Science*, 2021, 371: eabg0879
- 199 Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Židek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. *Nature*, 2020, 577: 706–710
- 200 Zhu X, Huang G, Zeng C, Zhan X, Liang K, Xu Q, Zhao Y, Wang P, Wang Q, Zhou Q, Tao Q, Liu M, Lei J, Yan C, Shi Y. *Science*, 2022, 376: eab18280
- 201 Bley CJ, Nie S, Mobbs GW, Petrovic S, Gres AT, Liu X, Mukherjee S, Harvey S, Huber FM, Lin DH, Brown B, Tang AW, Rundlet EJ, Correia AR, Chen S, Regmi SG, Stevens TA, Jette CA, Dasso M, Patke A, Palazzo AF, Kossiakoff AA, Hoelz A. *Science*, 2022, 376: eabm9129
- 202 Fontana P, Dong Y, Pi X, Tong AB, Hecksel CW, Wang L, Fu TM, Bustamante C, Wu H. *Science*, 2022, 376: eabm9326
- 203 Mosalaganti S, Obarska-Kosinska A, Siggel M, Taniguchi R, Turoňová B, Zimmerli CE, Buczak K, Schmidt FH, Margiotta E, Mackmull MT, Hagen WJH, Hummer G, Kosinski J, Beck M. *Science*, 2022, 376: eabm9506
- 204 Petrovic S, Samanta D, Perriches T, Bley CJ, Thierbach K, Brown B, Nie S, Mobbs GW, Stevens TA, Liu X, Tomaleri GP, Schaus L, Hoelz A. *Science*, 2022, 376: eabm9798
- 205 Chen G, Su W, Ding S, Hong Y. *IEEE Trans Automat Contr*, 2020, 65: 3409–3424
- 206 Qian H. *Sci Sin-Vitae*, 2016, 47: 257–261 (in Chinese) [钱紘. 中国科学:生命科学, 2017, 47: 257–261]
- 207 Zhang X, Song B, Jiang L. *Acc Chem Res*, 2022, 55: 1195–1204
- 208 Eyring H. *J Chem Phys*, 1935, 3: 107–115
- 209 Prigogine I. *Introduction of Thermodynamics of Irreversible Processes*. Weinheim: John Wiley & Sons, Inc., 1967
- 210 Nicolis G, Prigogine I. *Selforganization in Nonequilibrium Systems*. Weinheim: John Wiley & Sons, Inc., 1977
- 211 Wang Y, Lin HX, Chen L, Ding SY, Lei ZC, Liu DY, Cao XY, Liang HJ, Jiang YB, Tian ZQ. *Chem Soc Rev*, 2014, 43: 399–411
- 212 Wang Y, Lin HX, Ding SY, Liu DY, Chen L, Lei ZC, Fan FR, Tian ZQ. *Sci Sin-Chim*, 2012, 42: 525–547 (in Chinese) [王宇, 林海昕, 丁松园, 刘德宇, 陈亮, 雷志超, 范凤茹, 田中群. 中国科学:化学, 2012, 42: 525–547]

Exploring the theoretical foundation of molecular assembly: current status and opportunities

Hang Qu¹, Tianyi Tong^{1,2}, Zhi-Chao Lei³, Peichen Shi¹, Liulin Yang¹, Xiaoyu Cao¹, Yiqin Gao⁴, Zhonghuai Hou⁵, Xin Xu², Zhong-Qun Tian^{1*}

¹ State Key Laboratory of Physical Chemistry of Solid Surfaces, Department of Chemistry, College of Chemistry and Chemical Engineering, and iChEM, Xiamen University, Xiamen 361005, China

² Department of Chemistry, Fudan University, Shanghai 200433, China

³ National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

⁴ Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

⁵ Department of Chemical Physics, Hefei National Center for Physical Sciences at the Microscale, University of Science and Technology of China, Hefei 230026, China

*Corresponding author (email: zqtian@xmu.edu.cn)

Abstract: Molecular assembly (MA) is an essential approach to create novel substances and new functional materials beyond molecules. Improving the controllability of MA processes and the functionality of assemblies is the core target in this field. However, due to the limitation of methods and theories for complex MA systems, most of current researches are based on the “black box” paradigm, which can mostly focus on the initial and final state of MA. Such a paradigm has become a bottleneck that restricts the development of MA. Here, we review the current theoretical methods and models of MA systems from the molecular to subcellular level. We try to measure the complexity of MA research from the dimension of matter, energy and information, and gradually explore the current status, opportunities and breakthrough of the theoretical research of MA: in the dimension of matter, we illuminate that the pathways of MA processes are multi-step, multi-pathway and multi-scale; in the integrated dimension of energy and matter, we demonstrate that MA systems are usually accompanied by the entropy-driven phenomena or the complementation of entropy and enthalpy. We further discuss how the MA systems far from the equilibrium can form an order spatiotemporal dissipated structure; in the integrated dimension of information, energy and matter, we demonstrate how the synergetic processes of positive or negative feedback facilitates the emergence of complex physiological functions of MA systems. To address these challenges in the development of MA theoretical researches, we probably need to establish a larger theoretical framework and synergistically develop the research methods of MA from the three dimensions of matter, energy and information. It seems to help us to explore its fundamental laws of MA comprehensively, then establish the new theory and promote the development of assisting methods with high efficiency, and finally elevate the complexity and functionality of MA system, which may provide new perspectives and methods for life science and soft matter science.

Keywords: molecular assembly, controllable assembly, cooperativity, feedback, non-equilibrium state, information theory

doi: [10.1360/SSC-2022-0186](https://doi.org/10.1360/SSC-2022-0186)