专题: 生物分子模拟中的机器学习

# 蛋白质 p$K_a$ 预测模型研究进展*

罗方芳 蔡志涛 黄艳东†

(集美大学计算机工程学院, 厦门 361021)

(2023 年 8 月 20 日收到; 2023 年 9 月 1 日收到修改稿)

pH 表征溶液的酸碱性, 是许多与人类重大疾病密切相关的生命活动的调控因子. p$K_a$ 决定可滴定基团在一定 pH 条件下的去质子化平衡, 是研究 pH 调控的生物化学过程的重要参量. 然而, 由于蛋白质结构的复杂性以及实验条件的限制, 蛋白质 p$K_a$ 通常需要借助理论预测. 近 30 年, 研究者们开发了各种基于先验知识的 p$K_a$ 预测模型. 随着近几年人工智能技术的快速发展, 人们开始尝试将人工智能算法应用于蛋白质 p$K_a$ 预测工具的开发. 本文介绍 p$K_a$ 理论预测近年来的一些重要研究进展, 主要包括恒定 pH 分子动力学以及基于泊松-玻尔兹曼方程、经验函数和机器学习的 p$K_a$ 预测模型. 在此基础上, 讨论蛋白质 p$K_a$ 预测模型的未来发展方向和应用前景.

**关键词:** 分子动力学, 泊松-玻尔兹曼方程, 机器学习, p$K_a$ 预测

**PACS:** 87.15.ap, 87.14.E–, 87.10.Vg, 87.15.A– **DOI:** 10.7498/aps.72.20231356

## 1 引 言

为保证正常的生命活动, 人体细胞的细胞质、细胞核以及各个细胞器需维持在特定的 pH 水平. 例如, 线粒体和溶酶体的 pH 分别是 8.0 和 4.7, 偏离细胞质的 7.2[1]. 其中, 用于表征溶液的酸碱度的 pH 为氢离子浓度的对数取负 (pH = –log[H+]), 其是人体中许多重要生物过程的调控因子, 例如物质跨膜转运[2]、酶催化[3]、蛋白质折叠[4]、多肽聚集[5]、脂质分子自组装[6]、病毒入侵细胞[7]和细胞能量代谢[8]. 从微观的角度, 以上生物过程均与关键可离子化基团的质子化 (protonation) 或去质子化反应 (deprotonation) 相关联. 可离子化基团的去质子化 (正反应) 和质子化反应 (逆反应): $\mathrm{AH} \rightleftharpoons \mathrm{A^-} + \mathrm{H^+}$, 其中, AH 是一种可离子化基团的质子化态, $\mathrm{A^-}$ 是去质子化态.

以 β 分泌酶 BACE1 为例阐述蛋白质功能和可离子化基团质子化/去质子化的关系. BACE1 的生物功能是裂解 β 淀粉样前体蛋白 APP. 它与神经退行性疾病阿尔茨海默症密切相关, 是典型的结构和功能依赖于 pH 的蛋白质. 该蛋白的催化中心含两个天冬氨酸 Asp32 和 Asp228 (图 1(a)). 实验指出, BACE1 仅在一个狭小的 pH 范围内具有活性[9]. 如图 1(b) 所示, 在最适 pH 条件下 (约等于 4.5), Asp32 处于质子化态, 扮演质子供体 (proton donor); Asp228 处于去质子化态, 扮演亲核试剂 (nucleophile). 然而, 当溶液 pH 偏离 4.5, 两个天冬氨酸同时质子化或去质子化, BACE1 无法行使其生物功能[10].

当一个可离子化基团的质子化和去质子化达到平衡, 可由以下公式计算解离常数 $K_a$:

$$K_a = \frac{[\mathrm{H^+}]\,[\mathrm{A^-}]}{[\mathrm{AH}]}, \tag{1}$$

其中, $[\mathrm{H^+}]$, $[\mathrm{A^-}]$ 和 [AH] 分别代表溶液中氢离子

* 国家自然科学基金 (批准号: 11804114, 62006096)、福建省自然科学基金 (批准号: 2023J01329, 2020J05146)、厦门市自然科学基金 (批准号: 3502Z20227205) 和集美大学校启动金 (批准号: ZQ2020027) 资助的课题.

† 通信作者. E-mail: yandonghuang@jmu.edu.cn

 

以及该基团去质子化和质子化态下的浓度. $K_a$ 代表一种酸 (如 AH) 离解氢离子的能力. 将方程 (1) 的两边对数取负, 可得到著名的 Henderson-Hasselbalch 方程:

$$\mathrm{pH} = \mathrm{p}K_a + \log\left(\frac{\mathrm{A}^-}{\mathrm{AH}}\right) \tag{2}$$

其中, $\mathrm{p}K_a$ 为解离常数 $K_a$ 的对数取负, 代表一种酸 (如 AH) 去质子化的难易程度. 例如, 溶液中天冬氨酸的 $\mathrm{p}K_a$ 测量值是 3.7[11]. 根据 (2) 式, 天冬氨酸在中性 (pH = 7.0) 水溶液中处于去质子化态 ($\mathrm{A}^-$); 在 pH 小于 3.7 的酸性溶液中, 天冬氨酸质子化 (AH); 当 pH 位于 $\mathrm{p}K_a$ 附近, 质子化和去质子化态共存. 如上所述, $\mathrm{p}K_a$ 决定了可离子化基团在任意 pH 条件下的质子化和去质子化反应平衡. 根据 $\mathrm{p}K_a$ 值, 可以推断不同 pH 条件下生物大分子质子化态的分布, 进而讨论结构和功能的关系. 因此, $\mathrm{p}K_a$ 是研究 pH 相关的生物化学过程的一个核心问题. 不仅如此, $\mathrm{p}K_a$ 与药物研发中长期存在的靶向性和抗药性问题以及蛋白质设计密切相关. 然而, 由于蛋白质结构的复杂性以及实验条件的限制, 人们难于通过实验获取蛋白质中可离子化氨基酸残基的 $\mathrm{p}K_a$, 需借助理论预测.

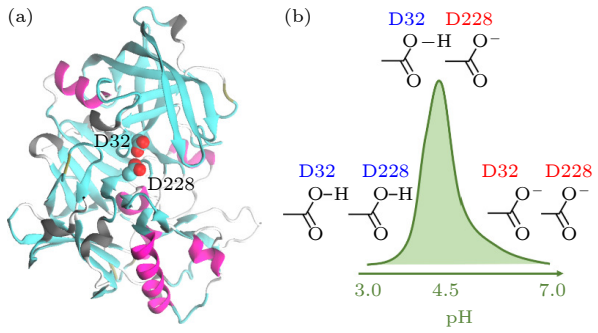


图 1 BACE1 催化中心质子化态和功能的关系 (a) BACE1 三维结构及其催化中心酸性二分体 D32 和 D228; (b) D32 和 D228 质子化态和蛋白质活性随 pH 的变化规律 (D 是 Asp 的缩写)

Fig. 1. Relationship between protonation state of BACE1 catalytic center and the function: (a) Crystal structure of BACE1 and the acidic dyad in the catalytic center; (b) protonation states of D32 and D228 and the activity as a function of pH (D is the abbreviation of Asp).

为此, 将以上 Henderson-Hasselbalch 方程转换为能量形式, 得到游离氨基酸关于 pH 和 $\mathrm{p}K_a$ 的去质子化自由能 $\Delta G^{\mathrm{mod}}$ 的表达式:

$$\Delta G^{\mathrm{mod}} = \ln 10 \times k_{\mathrm{B}}T\left(\mathrm{p}K_a^{\mathrm{mod}} - \mathrm{pH}\right), \tag{3}$$

其中, $k_{\mathrm{B}}$ 和 $T$ 分别是玻尔兹曼常数和温度; $\mathrm{p}K_a^{\mathrm{mod}}$ 为游离氨基酸的 $\mathrm{p}K_a$, 是可测量值. 去质子化自由能可分解为成键作用部分 $\Delta G_{\mathrm{Bond}}$ 和非键作用部分 $\Delta G_{\mathrm{NBond}}$. 其中, 成键作用部分描述共价键断裂的能量变化, 计算复杂度高, 不适用于生物大分子体系[12]. 值得一提的是, 当溶剂中的可离子化氨基酸参与蛋白质的合成, 蛋白质环境对成键作用部分的影响可忽略不计. 基于该假设, 我们只需考虑非键作用部分. 因此, 可离子化氨基酸从溶剂到蛋白质的去质子化自由能改变量 $\Delta G - \Delta G^{\mathrm{mod}}$ 可表示为

$$\Delta G - \Delta G^{\mathrm{mod}} = \Delta G_{\mathrm{NBond}} - \Delta G_{\mathrm{NBond}}^{\mathrm{mod}}. \tag{4}$$

根据 (3) 式, $\Delta G^{\mathrm{mod}}$ 为已知量. 因此, 求解蛋白质中氨基酸残基的去质子化自由能 $\Delta G$ 的问题简化为计算蛋白质环境对非键作用部分的自由能微扰 $\Delta G_{\mathrm{NBond}} - \Delta G_{\mathrm{NBond}}^{\mathrm{mod}}$.

基于以上框架, 人们发展了基于自由能计算的蛋白质 $\mathrm{p}K_a$ 预测模型, 例如恒定 pH 分子动力学 (constant pH molecular dynamics, CpHMD)[13]. 许多生物大分子含有不止一个功能构象, 并且构象的转变与质子化/去质子化反应相关联: 当活性位点质子化 ($\mathrm{pH} < \mathrm{p}K_a$), 蛋白处于构象 $C_1$; 去质子化 ($\mathrm{pH} > \mathrm{p}K_a$), 构象由 $C_1$ 转变到 $C_2$; 当 pH 取 $\mathrm{p}K_a$ 附近, 质子化和去质子化态共存, 构象 $C_1$ 与 $C_2$ 相互转变. 因此, 只有考虑了构象与质子化态耦合的理论模型, 才能得到和实验相一致的宏观 $\mathrm{p}K_a$ (macroscopic $\mathrm{p}K_a$)[14]. CpHMD 通过分子动力学模拟实现在不同构象下对质子化态空间进行采样. 在蛋白质 $\mathrm{p}K_a$ 预测精度方面, CpHMD 相对其他现有模型具有明显的优势[15]. CpHMD 的缺点是 $\mathrm{p}K_a$ 计算效率低. 例如, 完成一个蛋白质 $\mathrm{p}K_a$ 的计算通常需要进行几个小时甚至几天的分子动力学模拟, 因此难以满足工业界大批量计算的需求. 目前, CpHMD 多被应用于结构和功能依赖于 pH 的药物靶向蛋白的分子机制研究[16].

为了实现高通量的 $\mathrm{p}K_a$ 计算, 人们发展了基于泊松-玻尔兹曼 (Poisson-Boltzmann, PB) 方程的模型, 主要包括 MCCE[17], H++[18], APBS[19], DelPhiPKa[20] 和 PypKa[21]. 基于 PB 的模型能够在几分钟内完成一个蛋白质的 $\mathrm{p}K_a$ 计算, 极大地提高了计算效率. 然而, 基于 PB 的模型具有其理论局限性. 例如, 由于连续介质假设, PB 方程不适用于非水溶性的膜蛋白. 其次, 蛋白质结构的复杂性增加了

介电常数的不确定性, 因此即便是水溶性蛋白, 分子内部 (例如酶的催化反应中心) 的 p$K_a$ 计算对介电常数敏感[22].

除了以上基于能量的模型, 人们也可以用一个经验函数描述某可离子化氨基酸残基的蛋白质环境 (如疏水环境和氢键) 与其 p$K_a$ 偏移量的映射关系. 蛋白质某氨基酸残基 p$K_a$ 可表示为其游离状态下参考值 $\mathrm{p}K_a^{\mathrm{mod}}$ 和偏移量 $\Delta\mathrm{p}K_a$ 的和:

$$\mathrm{p}K_a = \mathrm{p}K_a^{\mathrm{mod}} + \Delta\mathrm{p}K_a. \tag{5}$$

2005 年, 基于前期的第一性原理计算工作[12], 哥本哈根大学 Jensen 课题组[23] 提出了一个计算蛋白质 p$K_a$ 的经验函数 PropKa. 该模型提出一组经验公式分别计算库仑力、去溶剂化效应和氢键等关键因素对 p$K_a$ 偏离参考值的贡献. PropKa 可在几秒内完成一个蛋白质的 p$K_a$ 计算, 计算效率明显比基于 PB 的模型高, 近 20 年得到了广泛的应用, 其最新版本 PropKa 3.0 发表于 2011 年[24].

直到 2021 年 12 月, 本课题组[25] 发表了首个人工智能 (artificial intelligence, AI) 驱动的蛋白质 p$K_a$ 预测模型 DeepKa. 随后, 美国卡内基·梅隆大学 Olexandr Lsayev、美国约翰斯·霍普金斯大学 Ana Damjanovic 和德国拜耳公司 Pedro Reis 研究小组陆续提出了基于机器学习的 p$K_a$ 预测模型 pKa-ANI[26], XGB-WMa[27] 和 PKAI/PKAI+[28]. 其中, DeepKa 和 PKAI/PKAI+主要依赖于数据集, 而为了在少样本情况下建立有效模型, pKa-ANI 和 XGB-WMa 需要一定程度的预训练或先验知识. 值得一提的是, 机器学习模型也能够在几秒内完成一个蛋白质的 p$K_a$ 计算.

上述的 CpHMD 以及基于 PB 方程、经验函数和机器学习的模型是目前 4 种主流的 p$K_a$ 预测方法. 最近, 本课题组[29] 采用 CpHMD 扩增了 p$K_a$ 数据集, 进一步提高了 DeepKa 的预测精度. 值得一提的是, DeepKa 已展现出类似物理模型 (如 CpHMD) 的高鲁棒性, 进一步证明了人工智能算法在蛋白质 p$K_a$ 预测领域的有效性. 下面将介绍这 4 种主流方法的理论基础及研究进展.

## 2 蛋白质 p$K_a$ 预测方法

### 2.1 CpHMD

根据质子化态采样方法的不同, 恒定 pH 分子动力学 CpHMD 分为随机采样 (discrete CpHMD, D-CpHMD)[30] 和 λ 动力学 (continuous CpHMD, C-CpHMD)[31]. 随机采样利用蒙特卡罗 (Monte Carlo, MC) 模拟在离散的质子化态空间 (反应坐标取 0 或 1) 进行采样[30]. λ 动力学则采用取值范围 0(质子化态) 到 1(去质子化态) 的连续变量 λ 作为反应坐标对可离子化基团的电荷或体系哈密顿量进行标度[31]. 如图 2 所示, 先使用以上基于 MC 或 λ 动力学的采样算法更新质子化态或者电荷. 基于更新后的电荷分布, 通过分子动力学模拟对构象进行采样. 更新位置坐标后, 进入下一轮质子化态的采样. 模拟结束后, 采用广义 Henderson-Hasselbalch 方程拟合 CpHMD 模拟产生的不同 pH 条件下某可离子化基团的去质子化概率 $S$, 进而获得其 p$K_a$ 值, 即 $S = 0.5$ 所对应的 pH[31].

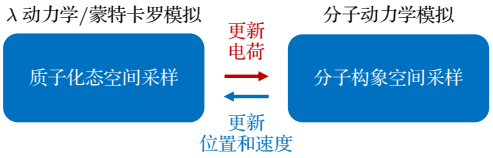


图 2 CpHMD 模拟框架

Fig. 2. Framework of a CpHMD simulation.

由于滴定动力学与构象动力学相关联, 提高质子化态和构象空间的采样是近 30 年 CpHMD 模型发展的主线. 下面将分别介绍 D-CpHMD 和 C-CpHMD.

#### 2.1.1 D-CpHMD

D-CpHMD 用一个反应坐标 λ 表示某可离子化位点的质子化态. λ 只能取 0 或 1. 其中, 0 和 1 分别表示质子化态和去质子化态. 经过一定长度的分子动力学 (molecular dynamics, MD) 模拟, 随机选取一个可离子化基团, 尝试改变其质子化态. 例如, 将其 λ 值从 0 改为 1. 然后, 计算 λ 值改变引起的能量变化 $\Delta E$. 将该能量变化代入 Metropolis 准则:

$$p = \begin{cases} 1, & \Delta E \leqslant 0, \\ \exp(-\Delta E/k_{\mathrm{B}}T), & \Delta E > 0. \end{cases} \tag{6}$$

如果能量差小于或等于 0, 接受 λ 值改变的概率为 1. 如果能量差大于 0, 则接受改变的概率 $p$ 小于 1. 在数值模拟中, 通常是随机生成一个取值范围为 [0, 1] 的数 $s$. 只有 $s$ 小于等于 $p$, 才接受 λ 值改变, 否则保留原值. 以上为一步的 MC, 和开始

的 MD 构成一个模拟周期. 因此, 在 MC 之后, 便是下一个周期的 MD 模拟. 显性溶剂下质子化或去质子化的能量变化较大, 导致较小的接受概率. 起初, 为了提高接受概率或质子化态的采样效率, MC 的能量计算使用隐性溶剂 (implicit solvent) 模型, 如广义玻恩 (generalized Born, GB)[32–34] 和引言提到的 PB 模型 [31,35,36]. 当 MC 和 MD 均采用隐性溶剂, 计算效率最高, 但是牺牲了精度 [32,33]. 为了提高构象方面的采样精度, MD 可替换成显性溶剂, 即杂化溶剂 [31,34,35]. 其中, 基于 GB 和 PB 的模型分别在分子模拟软件 Amber 和 GROMACS 中已被实现. 需要指出的是, 隐性溶剂难以描述活性位点附近与功能相关的水分子或盐离子对去质子化平衡的影响 [37].

为提高显性水溶剂下 MC 的接受概率, 2007 年 Stern[38] 提出了尝试改变 λ 值之后, 先进行一定长度的尝试性的分子动力学模拟, 再计算能量差. 该尝试性的 MD 使周围水溶剂构型得到调整, 可降低 λ 值改变前后的能量差. 然而, 以上尝试性 MD 的长度依赖于经验或不确定, 其应用可能受到限制. 尽管如此, 该模型为解决显性溶剂下质子化态空间的采样问题提供了一条新思路. 随着高性能计算的发展, 人们开始考虑将显性溶剂应用到蛋白质 D-CpHMD 的 MC 部分. 如无特别说明, 以下提到的显性溶剂均是分子动力学模拟中计算静电相互作用的标准算法 PME (particle mesh Ewald, PME)[39]. 2015 年芝加哥大学的 Roux 课题组 [40] 提出了显性溶剂下的非平衡 MD/MC 模拟. 例如, 对于某可滴定位点在 MC 阶段的去质子化 (λ 由 0 变为 1) 尝试, 该模型在 0 和 1 之间添加了 $m$ 个中间值. 对于每个 λ 值 ($m$ 个中间值和两个边界值 0 和 1), 执行一定长度的非平衡 MD, 令可离子化基团周围的环境根据 λ 值在构型上作出调整, 减缓了因 λ 值改变而导致的能量涨落. 结束 $\lambda=1$ 的非平衡 MD 后, 计算当 $\lambda=1$ 和 $\lambda=0$ 的能量差. 同样, 根据 Metropolis 准则, 如果接受该可滴定位点去质子化, 继续 $\lambda=1$ 的 MD. 否则, 退回到非平衡 MD 前的时刻, 继续 $\lambda=0$ 的 MD. 通过以上的非平衡模拟, 该模型提供了较合理的能量差的计算, 提高了总体接受概率. Roux 课题组 [40,41] 利用著名的 Jarzynski 方程将自由能变化与非平衡 MD 所做的功相关联, 使得以上非平衡 MD 的模拟时间可被量化. 值得一提的是, 该方法可被应用于生物大分子, 目前在分子模拟软件 NAMD 中已有实现. 然而, 可滴定氨基酸的固有 p$K_a$ (inherent p$K_a$) 是该模型的一个主要参量. 为了提高预测性能, 该模型要求固有 p$K_a$ 尽可能接近真实值 [41]. 因此, D-CpHMD 一个潜在的研究方向是消除上述模型对固有 p$K_a$ 的依赖.

### 2.1.2 C-CpHMD

本课题组统计了 4057 个蛋白质中可滴定氨基酸的个数 [29]. 这些蛋白质来自复旦大学王任小实验室 [42] 创建的蛋白质抑制剂复合物数据库 PDBbind 的精细集 v2016. 除了半胱氨酸 Cys, 蛋白质中其他可滴定氨基酸类型 (谷氨酸 Glu、天冬氨酸 Asp、赖氨酸 Lys、精氨酸 Arg、络氨酸 Tyr、组氨酸 His) 的平均个数不低于 10[29]. 理论上, 一个含有 $N$ 个可滴定氨基酸残基的蛋白质包含 $2^N$ 个质子化态. 然而, D-CpHMD 的 MC 每次只取一个可滴定位点来判断是否改变其质子化态, 采样效率较低 [34,43,44].

2004 年, 为了研究生物大分子体系 (如蛋白质, DNA 和 RNA) 的质子化和去质子化, 密西根大学 Brooks 课题组开发了首个 λ 动力学框架下 [45] 的恒定 pH 分子动力学 C-CpHMD[31]. 每个可滴定位点对应一个反应坐标 λ, 取值范围同样是 0—1. 和 D-CpHMD 不同的是, C-CpHMD 的反应坐标是连续的变量. 值得一提的是, C-CpHMD 同时更新所有可滴定位点的质子化态. 哈密顿量 $H$ 代表体系的总能量, 包括动能和势能. 除了真实的粒子, 如模拟体系中溶剂和溶质的原子, C-CpHMD 添加了虚粒子. 每个可滴定基团对应一个虚粒子. 这里用范围在 [0, 1] 的连续变量 λ 作为虚粒子的坐标. 为了模拟虚粒子的滴定动力学, 可将其质量设为 10 (单位是原子质量). 以下是修正后的总哈密顿量:

$$\begin{aligned} &H(\{\boldsymbol{r}_a\},\{\lambda_j\}) \\ &=\sum_a^{N_{\text{atom}}}\frac{1}{2}m_a\dot{\boldsymbol{r}}_a^2+U^{\text{bond}}(\{\boldsymbol{r}_a\})+U^{\text{nbond}}(\{\boldsymbol{r}_a\},\{\lambda_j\}) \\ &+\sum_j^{N_{\text{titr}}}\frac{1}{2}m_j\dot{\lambda}_j^2+U^*(\{\lambda_j\}), \end{aligned} \tag{7}$$

其中, $N_{\text{atom}}$ 是总粒子数, $\boldsymbol{r}$ 是原子的位置矢量, λ 是虚粒子的滴定坐标, $m_a$ 和 $m_j$ 是原子和虚粒子的质量. 第 1 和第 4 项的求和分别是原子和虚粒子的总动能. 第 2 项 $U^{\text{bond}}$ 是键相互作用能, 包括键伸缩能、键角弯折能和二面角扭转能. 这里假设键

相互作用与 $\lambda$ 无关. 第 3 项 $U^{\rm nbond}$ 是非键相互作用能, 包括静电 $U^{\rm elec}$ 和范德瓦耳斯 $U^{\rm vdW}$ 相互作用, 与 $\lambda$ 相关. 最后一项 $U^{*}$ 是偏置势, 利用经验势描述去质子化键断裂的能量变化, 只和 $\lambda$ 相关.

以下介绍如何利用 $\lambda$ 标度非键相互作用能和偏置势. 对于可滴定的氢原子和周围原子的范德瓦耳斯相互作用, 直接用 $1-\lambda_i$ 标度势能函数 (这里采用 6-12 勒让德琼斯势 $U^{\rm LJ}$ ):

$$U_{ij}^{\rm vdW}=(1-\lambda_i)\,U_{ij}^{\rm LJ}. \tag{8}$$

可见, 当 $\lambda=1$ 时, 残基 $i$ 去质子化, 残基 $i$ 的可滴定氢与 $j$ 无相互作用.

对于两个可滴定氢之间的范德瓦耳斯相互作用, 采用 $1-\lambda_i$ 和 $1-\lambda_j$ 进行标度:

$$U_{ij}^{\rm vdW}=(1-\lambda_i)\,(1-\lambda_j)\,U_{ij}^{\rm LJ}. \tag{9}$$

范德瓦耳斯力是近程非键相互作用力, 主导疏水基团间的相互作用. 然而, 由于原子半径的差异, 氢 (半径约 1 Å) 几乎被与之成键 (键长约 1 Å) 的重原子 (半径约 2 Å) 包围, 使其难以接触到其他原子. 因此, 质子化和去质子化对范德瓦耳斯相互作用影响不大, 相对长程静电相互作用可以忽略不计. 对于静电相互作用, $\lambda$ 标度的是原子电荷 [31]:

$$q_{{\rm a},j}=\lambda_j q_{a,j}^{\rm dep}+(1-\lambda_j)\,q_{a,j}^{\rm prot}, \tag{10}$$

其中, $q_{a,j}^{\rm dep}$ 和 $q_{a,j}^{\rm prot}$ 是氨基酸残基 $j$ 处于去质子化态和质子化态时原子 $a$ 所带电荷. 静电相互作用 $U^{\rm elec}$ 计算复杂度高, 在分子动力学模拟中占据大多数计算资源, 特别是和溶剂相关的部分. 因为静电势是 p$K_{\rm a}$ 计算的关键因子, 我们将详细介绍不同溶剂条件下的 C-CpHMD.

早期为了提高计算效率, Brooks 课题组 [31] 采用隐性溶剂模型计算溶剂对溶质的平均效应. 如此一来, 总静电能 $U^{\rm elec}$ 的溶质内静电相互作用仍采用库仑势 ((11) 式第 1 项), 而溶质与溶剂的静电相互作用 $U^{\rm solv}$ 采用 GB 势能函数 ((12) 式):

$$U^{\rm elec}=\sum_{a<b}^{N_{\rm atom}*}\frac{q_aq_b}{r_{ab}}+U^{\rm solv}, \tag{11}$$

$$\begin{aligned}U^{\rm solv}=&-\frac{1}{2}\sum_{a,b}^{N_{\rm atom}}\left(\frac{1}{\varepsilon_{\rm p}}-\frac{{\rm e}^{-\kappa r_{ab}}}{\varepsilon_{\rm w}}\right)\\&\times\frac{q_aq_b}{\sqrt{r_{ab}^2+\alpha_a\alpha_b{\rm e}^{-r_{ab}^2/4\alpha_a\alpha_b}}},\end{aligned} \tag{12}$$

$$\kappa^2=\frac{8\pi q^2 I}{ek_{\rm B}T}, \tag{13}$$

其中, 星号代表排除存在键相互作用的原子对; $r_{ab}$ 是电荷 $q_a$ 和 $q_b$ 的距离; $\varepsilon_{\rm p}$ 和 $\varepsilon_{\rm w}$ 是蛋白质和水的介电常数; $\kappa$ 是德拜长度取反 ((13) 式); $I$ 是盐离子强度; $q$ 是盐离子电荷; $e$ 是基本电荷; $k_{\rm B}$ 是玻尔兹曼常数; $T$ 是温度; $\alpha$ 是有效玻恩半径, 表征某原子埋在蛋白内部的程度, 为衡量 GB 模型精度的关键参数. 相对 PB 模型, GB 的计算复杂度较低, 并且是解析的, 适合需要对位置坐标求一阶导 (计算粒子所受合外力) 的分子动力学模拟. GB 模型的计算复杂度主要体现在有效玻恩半径的求解.

2004 和 2005 年 Brooks 课题组接连开发了 CHARMM 软件中基于隐性溶剂 GBMV[31] 和 GBSW[46] 的 C-CpHMD, 证明了基于 GB 的 C-CpHMD 在 p$K_{\rm a}$ 预测方面的有效性. 相对 GBSW/GBMV 溶剂模型, GBNeck2 可提供更优的构象采样 [47]. 于是, 马里兰大学 Shen 课题组 [48] 在 2018 年开发了 Amber 软件中基于隐性溶剂 GBNeck2 的 C-CpHMD. 值得一提的是, 对于实验科学家关心的酶催化中心 (如图 1 活性位点 Asp32 和 Asp228), 该方法也表现较好, 目前已被应用于共价抑制剂靶点的预测 [49–51], 蛋白质 p$K_{\rm a}$ 数据集的建立 [25,29], 以及依赖于 pH 的蛋白质分子机制研究 [52,53]. 目前, 基于 GBSW 和 GBNeck2 的 C-CpHMD 均已实现 GPU 加速, 这进一步扩展了模型的应用范畴 [54,55].

为了提高构象采样精度以及扩展 C-CpHMD 的应用范围, Shen 课题组 [56] 提出了杂化溶剂 C-CpHMD: 构象动力学使用显性溶剂; 而滴定动力学保留隐性溶剂. 为此, 构象动力学和滴定动力学采用不同的哈密顿量. 前者去掉方程 (7) 的最后两项, 第 3 项不再包含反应坐标 $\lambda$, 令方程 (7) 回归到常规分子动力学. 该方法不仅维持了质子化态空间采样效率, 而且提高了构象采样精度. 起初人们会担心隐性溶剂 GB 的理论局限性 (例如偏弱的疏水效应) 会影响 p$K_{\rm a}$ 预测精度. 然而, Shen 课题组 [56] 发现, 显性溶剂 PME 可导致偏高的疏水效应, 一定程度上抵消了隐性溶剂导致的偏弱的疏水效应. 相对隐性溶剂, 该杂化溶剂 C-CpHMD 获得了广泛的应用, 如钠离子质子交换蛋白 [37,57], 质子通道 [58], 类药物分子的膜渗透 [59], 芬太尼激活 G 耦联受体 [60], 糖苷水解酶 [61], 络氨酸激酶药物发现 [62], 以及上文提及的 β 分泌酶 [10].

为了描述和功能相关的水分子或其他辅助因子 (如金属离子和小分子) 对去质子化平衡的影响, 滴定动力学部分也需采用显性溶剂. 起初, Brooks 课题组和 Shen 课题组分别选择了较简单的基于截断的显性溶剂 FSh (force shifting, FSh)[63] 和 GRF (generalized reaction field, GRF)[64]. 然而, 由于截断, 这两个模型均低估了长程静电力对可滴定位点的影响[65]. 为此, Shen 课题组[66] 开发了基于显性溶剂 PME 的 C-CpHMD. 最近, 该模型在分子模拟软件 Amber 中实现了 GPU 加速[67]. 众所周知, PME 是满足周期性边界条件 (periodic boundary condition, PBC) 的分子模拟中计算静电相互作用的标准算法, 因此基于 PME 的 C-CpHMD 是 $\lambda$ 动力学框架下所能达到的最优版本. 理论上, 如果不考虑取样问题, 该模型的 p$K_{\rm a}$ 预测应该最接近实验. 对于一个满足 PBC 的分子动力学模拟体系, PME 的总静电能是 3 个能量项的加和:

$$U^{\rm elec}=U^{\rm dir}+U^{\rm rec}+U^{\rm corr}, \tag{14}$$

其中, $U^{\rm dir}$ 是实空间静电相互作用, 在库仑势基础上增加一个补偿函数, 负责截断距离以内的短程静电相互作用 ((15) 式). $U^{\rm rec}$ 最为耗时, 为倒格空间 (reciprocal space) 下求解的长程静电能, 负责截断以外的长程静电相互作用 ((16) 式). $U^{\rm corr}$ 是修正项 ((20) 式)[39].

$$U^{\rm dir}=\frac{1}{2}\sum_{\boldsymbol{n}}^{*}\sum_{a,b=1}^{N_{\rm atom}*}\frac{q_aq_b{\rm erf}\left(\beta\left|\boldsymbol{r}_b-\boldsymbol{r}_a+\boldsymbol{n}\right|\right)}{\left|\boldsymbol{r}_b-\boldsymbol{r}_a+\boldsymbol{n}\right|}, \tag{15}$$

其中, $\boldsymbol{r}_a$ 和 $\boldsymbol{r}_b$ 是中心元胞的位置矢量; $\boldsymbol{n}$ 是元胞的位置矢量, 其表达式为 $\boldsymbol{n}=n_1\boldsymbol{c}_1+n_2\boldsymbol{c}_2+n_3\boldsymbol{c}_3$, 其中 $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ 和 $\boldsymbol{c}_3$ 代表元胞的 3 个正交方向矢量; 星号代表被排除的原子对, 包括原子自身 ($a=b$), 形成化学键的原子对, 以及最近邻 ($n$ 的大小为 1) 以外的镜像; erf 是补偿误差函数; 参数 $\beta$ 决定 $U^{\rm dir}$ 和 $U^{\rm rec}$ 的相对收敛速度. 例如, $\beta$ 越大, $U^{\rm dir}$ 计算收敛越快, 而 $U^{\rm rec}$ 计算收敛会越慢.

$$U^{\rm rec}=\frac{1}{2\pi V}\sum_{m\neq0}\frac{\exp\left(-\pi^2\boldsymbol{m}^2/\beta^2\right)}{\boldsymbol{m}^2}S\left(\boldsymbol{m}\right)S\left(-\boldsymbol{m}\right), \tag{16}$$

式中 $\boldsymbol{m}$ 是倒格矢, 其表达式为 $\boldsymbol{m}=m_1\boldsymbol{c}_1^*+m_2\boldsymbol{c}_2^*+m_3\boldsymbol{c}_3^*$, 其中, $m_1$, $m_2$, $m_3$ 是非零整数; $\boldsymbol{c}_i^*$ 是以上 $\boldsymbol{c}_i$ ($i=1,2,3$) 的共轭倒格矢, 二者满足关系式 $c_i^*\cdot c_j=\delta_{ij}$, 这里 $i$ 和 $j$ 取 1, 2 和 3. 另外, $V=\boldsymbol{c}_1\cdot\boldsymbol{c}_2\times\boldsymbol{c}_3$, 是元胞的体积. $S\left(\boldsymbol{m}\right)$ 是结构因子:

$$S\left(\boldsymbol{m}\right)=\sum_{a=1}^{N_{\rm atom}}q_a\exp\left(2\pi{\rm i}\boldsymbol{m}\cdot\boldsymbol{r}_a\right). \tag{17}$$

该结构因子可近似表示为

$$\begin{aligned}S\left(\boldsymbol{m}\right)\approx&\sum_{k_1,k_2,k_3}Q\left(k_1,k_2,k_3\right)\\&\times\exp\left[2\pi{\rm i}\left(\frac{m_1k_1}{K_1}+\frac{m_2k_2}{K_2}+\frac{m_3k_3}{K_3}\right)\right]\\=&F\left(Q\right)\left(m_1,m_2,m_3\right),\end{aligned} \tag{18}$$

式中通过将元胞中的电荷分布 (B 样条) 插值到具有相同的 3 个维度 $k_1$, $k_2$, $k_3$ 的网格来构造三维矩阵 $\boldsymbol{Q}$; $k_i/K_i$ 是分数坐标, 其中, $k_i$ ($i=1,2,3$) 取值范围是 $(1,2,3,\cdots,K_i)$, 正整数常数 $K_i$ 代表元胞的尺寸; $F\left(Q\right)$ 是矩阵 $\boldsymbol{Q}$ 的三维快速傅里叶变换. 经过以上变换, $U^{\rm rec}$ 的表达式为

$$\begin{aligned}U^{\rm rec}=&\frac{1}{2\pi\nu}\sum_{m\neq0}\frac{\exp\left(\left[-\left(\pi\boldsymbol{m}/\beta\right)^2\right]\right)}{\boldsymbol{m}^2}\\&\times F\left(Q\right)\left(\boldsymbol{m}\right)F\left(Q\right)\left(-\boldsymbol{m}\right).\end{aligned} \tag{19}$$

值得一提的是, $U^{\rm rec}$ 线性依赖于格点电荷, 因此对 $\lambda$ 求一阶导和库仑势的一样简单.

$$\begin{aligned}U^{\rm corr}=&-\frac{1}{2}\sum_{(a,b)\in M}\frac{q_aq_b{\rm erf}\left(\beta\left|\boldsymbol{r}_b-\boldsymbol{r}_a\right|\right)}{\left|\boldsymbol{r}_b-\boldsymbol{r}_a\right|}\\&-\frac{\beta}{\sqrt{\pi}}\sum_{a=1}^{N}q_a^2-\frac{\pi}{2\beta^2V}\left(\sum_aq_a\right)^2.\end{aligned} \tag{20}$$

$U^{\rm rec}$ 考虑整体的电荷分布, 并未排除存在键相互作用的原子对, 因此需采用和 $U^{\rm dir}$ 相同的函数形式进行修正 ((20) 式第 1 项). 此外, $U^{\rm corr}$ 第 2 项的作用是排除点电荷自相互作用, 第 3 项则是中和体系净电荷的背景电荷 (background plasma). 其中, 后面两个修正只依赖于原子电荷.

为了避免元胞之间不真实的静电相互作用, 常规 MD 通过添加补偿盐离子使体系呈电中性. 然而, CpHMD 模拟中电荷是动态变化的. 为了解决该问题, Shen 课题组[64] 提出了将盐离子作为质子缓存器. 然而, 盐离子如果不带电会导致聚集, 于是改使用可滴定水分子[68]. 酸性氨基酸 (例如 Asp 和 Glu) 与水阴离子 (hydroxide, TIPU) 耦合 ($\rm AH+OH^-\rightleftharpoons A^-+H_2O$); 碱性氨基酸 (例如 Lys, Arg 和 His) 与水阳离子 (hydronium, TIPP) 耦合 ($\rm BH^++H_2O\rightleftharpoons H_3O^++B$). 该耦合令反应式两端的电荷守

恒. 电中性的另一个好处是消除 $U^{\rm corr}$ 中会导致反常 p$K_{\rm a}$ 偏移的背景电荷.

以上介绍了不同溶剂下静电能的具体求解. 下面介绍哈密顿量中只依赖于反应坐标 λ 的偏置势[31]:

$$U^{*}\left(\{\lambda_j\}\right)=\sum_{j}^{N_{\rm titr}}\left[-U^{\rm mod}\left(\lambda_j\right)+U^{\rm pH}\left(\lambda_j\right)+U^{\rm barr}\left(\lambda_j\right)\right], \tag{21}$$

其中, 第 1 项 ((22) 式) 和第 2 项 ((23) 式) 分别是游离可滴定氨基酸去质子化的非键相互作用能和总自由能. 对于单个可滴定位点的氨基酸 (如赖氨酸), $U^{\rm mod}$ 是一个关于 λ 的一元二次函数. $U^{\rm pH}$ 由 λ 线性标度 ((23) 式). $U^{\rm pH}-U^{\rm mod}$ 是化学能改变量的近似解. 为了减少 λ 处于不真实的中间态 (如 $\lambda=0.5$) 的概率, 另外添加了一个二次函数势垒 $U^{\rm barr}$ ((24) 式). $U^{\rm barr}$ 降低了 λ 的动力学, 对热力学统计没有影响. (23) 式和 (24) 式的参数为已知, 因此, C-CpHMD 的主要工作是确定 $U^{\rm mod}$ 的参数 (如 (22) 式中的 $A_j$ 和 $B_j$ ):

$$U^{\rm mod}\left(\lambda_j\right)=A_j(\lambda_j-B_j)^2, \tag{22}$$

$$U^{\rm pH}\left(\lambda_j\right)=\ln(10)\,k_{\rm B}T\left({\rm p}K_{\rm a}^{\rm mod}-{\rm pH}\right)\lambda_j, \tag{23}$$

$$U^{\rm barr}\left(\lambda_j\right)=4\eta(\lambda_j-0.5)^2, \tag{24}$$

其中, p$K_{\rm a}^{\rm mod}$ 是游离可滴定氨基酸的 p$K_{\rm a}$ 测量值, $\eta$ 决定势垒高度. 对于一个 C-CpHMD 模型, 需要通过平均力势 (potential of mean force, PMF) 模拟求 $U^{\rm mod}$ 函数中的系数. 这里可用单个可滴定位点的游离赖氨酸 (Lys) 为例. 固定 λ 值, 经过一定时间 (如 1 ns) 的 MD, 对作用在虚粒子上的力求时间平均, 即 $\langle {\rm d}U/{\rm d}\lambda\rangle$, 其中 λ 在 0—1 之间取离散的值. 基于线性响应理论, 用线性函数 $2A\left(\lambda-B\right)$ 拟合平均力, 确定模型参数 $A$ 和 $B$. 同时, 可利用以下热力学积分求 PMF, 计算去质子化自由能改变量:

$$U^{\rm mod}\left(\lambda\right)=\int_0^{\lambda}\left\langle\frac{\partial U\left(\lambda'\right)}{\partial\lambda'}\right\rangle_{\lambda'}{\rm d}\lambda'. \tag{25}$$

需要注意的是, 为了将 λ 约束在 [0, 1], 需定义另一个变量 $\theta$. λ 和 $\theta$ 的关系式为 $\lambda=\sin^2\theta$. 于是, 数值模拟中进行迭代的是 $\theta$, 而非反应坐标 λ.

对于含有两个可滴定位点的氨基酸, 需要定义反应坐标 $x$ 来描述处于去质子化 (His) 或质子化 (Glu 和 Asp) 态时质子所处的可滴定位点[46]. $x$ 同样是在 0 到 1 范围内的连续变量. 图 3 展示了 Asp 和 His 侧链 3 个质子化态对应的反应坐标值以及状态间的转化. 类似变量 λ, 可利用插值将 $x$ 加入哈密顿量的各个能量项. 例如, 以下分别是 Asp 和 His 电荷关于 λ 和 $x$ 的表达式:

$$q_{a,j}^{\rm D}=\lambda_j q_{a,j}^{\rm ASP}+\left(1-\lambda_j\right)\left[x_j q_{a,j}^{\rm ASP2}+\left(1-x_j\right)q_{a,j}^{\rm ASP1}\right], \tag{26}$$

$$q_{a,j}^{\rm H}=\lambda_j\left[x_j q_{a,j}^{\rm HSE}+\left(1-x_j\right)q_{a,j}^{\rm HSD}\right]+\left(1-\lambda_j\right)q_{a,j}^{\rm HSP}, \tag{27}$$

其中 $q_{a,j}^{\rm ASP2}$ 和 $q_{a,j}^{\rm ASP1}$ 分别是 Asp 侧链 $j$ 上原子 $a$ 在 $O_{\delta2}$ 和 $O_{\delta1}$ 质子化时所带的电荷, $q_{\rm a,j}^{\rm ASP}$ 是该侧链去质子化时原子 $a$ 所带电荷; $q_{\rm a,j}^{\rm HSE}$ 和 $q_{\rm a,j}^{\rm HSD}$ 分别是 His 侧链 $j$ 上原子 $a$ 在 $N_\delta$ 和 $N_\varepsilon$ 去质子化时所带的电荷, $q_{\rm a,j}^{\rm HSP}$ 是该侧链质子化时原子 $a$ 所带电荷. 具有双可滴定位点的 Glu/Asp 和 His 的 $U^{\rm mod}$ 是关于 λ 和 $x$ 的多项式, 需要取 λ 和 $x$ 值的不同组合计算平均力, 然后通过 Brooks 课题组提出的方法计算多项式系数[46].

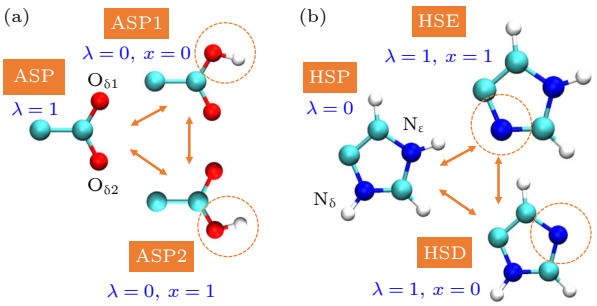


图 3 互变异构滴定模型的 3 个质子化态以及状态间的转化 (a) 天冬氨酸 Asp; (b) 组氨酸 His

Fig. 3. Three protonation states and their interconversion in the tautomeric titration model: (a) Aspartic acid; (b) histidine.

CpHMD 模拟同时对构象和质子化态采样. 根据设置的输出频率保存每个可离子化基团的滴定坐标 λ ( $\lambda\in[0,1]$)(图 4(a)). 统计处于质子化态 ( $0\leqslant\lambda\leqslant0.1$ ) 的次数 $N^{\rm prot}$ 以及去质子化态 ( $0.9\leqslant\lambda\leqslant1$ ) 的次数 $N^{\rm dep}$, 计算不同 pH 条件下的去质子化概率 $S$ (图 4(a))[31]:

$$S=\frac{N^{\rm dep}}{N^{\rm dep}+N^{\rm prot}}. \tag{28}$$

最后, 采用如下 Hill 函数 (广义 Henderson-Hasselbalch 函数) 拟合 $S$. p$K_{\rm a}$ 便是 $S=0.5$ 时所对应的 pH (图 4(b)):

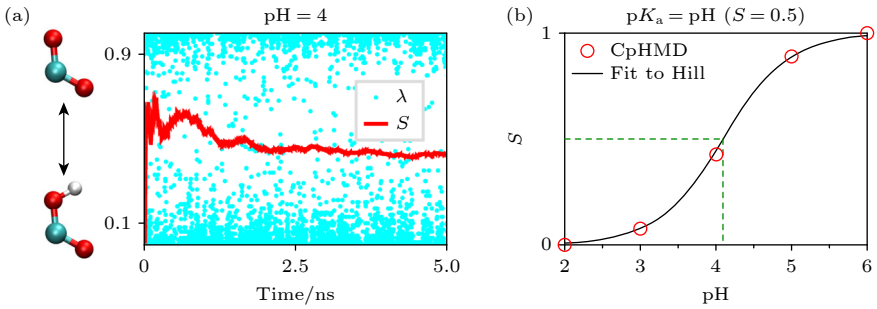


图 4 基于 C-CpHMD 的 $\mathrm{p}K_\mathrm{a}$ 计算 (a) 滴定坐标 $\lambda$ 和去质子化概率 $S$ 的轨迹; (b) 采用 Hill 函数拟合 $S$

Fig. 4. The $\mathrm{p}K_\mathrm{a}$ calculation based on C-CpHMD: (a) Trajectories of titration coordinate $\lambda$ and deprotonation fraction $S$; (b) fitting $S$ to Hill function.

$$S=\frac{1}{1+10^{h(\mathrm{p}K_\mathrm{a}-\mathrm{pH})}}, \tag{29}$$

其中 $h$ 是 Hill 系数, 表征一个可离子化基团与周围可滴定基团的滴定动力学是否存在耦合. $h=1$ 表示无耦合, 如位于分子表面的残基或游离氨基酸. $h<1$ 表示负耦合, 如形成盐桥键的去质子化的 Asp 和质子化的 Lys. $h>1$ 表示正耦合, 如酶活性位点距离相近的两个酸性氨基酸 (质子化的 Asp 或 Glu). $h$ 偏离 1 越多, 耦合越强 [69].

当两个氨基酸的滴定动力学存在耦合, 可将二者看作一个整体, 利用以下公式计算宏观 $\mathrm{p}K_1$ 和 $\mathrm{p}K_2$ (macroscopic sequential $\mathrm{p}K_\mathrm{a}$)[64,70]:

$$N=\frac{10^{(\mathrm{p}K_2-\mathrm{pH})}+2\times10^{(\mathrm{p}K_1+\mathrm{p}K_2-2\mathrm{pH})}}{1+10^{(\mathrm{p}K_2-\mathrm{pH})}+10^{(\mathrm{p}K_1+\mathrm{p}K_2-2\mathrm{pH})}}, \tag{30}$$

其中 $N$ 是一定 pH 条件下的平均质子数. 为获得 $\mathrm{p}K_1$ 和 $\mathrm{p}K_2$, 也可以采用以下非耦合模型 (31) 式 [71,72]:

$$S_1+S_2=\frac{1}{1+10^{(\mathrm{p}K_1-\mathrm{pH})}}+\frac{1}{1+10^{(\mathrm{p}K_2-\mathrm{pH})}}, \tag{31}$$

其中 $S_1$ 和 $S_2$ 分别是两个耦合的可滴定位点的去质子化概率.

当滴定动力学采用满足周期性边界条件的显性溶剂时, 需要考虑有限尺度效应 [73]. 由于采用耦合水离子实现了电中性, 有限尺度效应仅剩下和水分子模型相关的离散溶剂效应 (discrete solvent effect)[66]. 当某个可滴定氨基酸去质子化, 因离散溶剂效应引起的能量变化是

$$\Delta G^{\mathrm{offset}}=\frac{2\pi}{3}\kappa\gamma q\rho, \tag{32}$$

其中, $\kappa$ 是介电常数; $\rho$ 是水数量密度, 等于水分子数 $N$ 除以体积 $V$, 这里 $N$ 指的是和蛋白有相互作用的水分子数, $V$ 也是这些水包络范围内的体积; $q$ 是可滴定氨基酸的电荷, Asp/Glu 是–1$e$, His/Lys 为+1$e$; $\gamma$ 是显性溶剂模型范德瓦耳斯相互作用中心的电四极矩. 对于溶剂模型 TIP3P, $\gamma$ 的值为 0.764 $e\cdot\mathrm{Å}^2$. 为了估算该有限尺度效应导致的 $\mathrm{p}K_\mathrm{a}$ 偏移, 需要计算相对模型分子的能量变化 [66]:

$$\Delta\Delta G^{\mathrm{offset}}=\frac{2\pi}{3}\kappa\gamma q\left(\frac{N}{V}-\frac{N^{\mathrm{mod}}}{V^{\mathrm{mod}}}\right), \tag{33}$$

其中, $N$ 和 $N^{\mathrm{mod}}$ 分别是蛋白质和游离氨基酸模拟体系中与溶质有相互作用的水分子数; $V$ 和 $V^{\mathrm{mod}}$ 是相应的周期性元胞体积. 将以上表达式转化为 $\mathrm{p}K_\mathrm{a}$ 偏移量, 可得到 [66]

$$\Delta\mathrm{p}K_\mathrm{a}^{\mathrm{corr}}=\pm\frac{\Delta\Delta G^{\mathrm{offset}}}{\ln(10)\,RT}. \tag{34}$$

根据 $N$ 和 $V$ 的定义, 可以推断有限尺寸效应对 PME 影响较大. PME 考虑了周期性元胞内所有水分子, 蛋白质体积所占比例较小, 水数量密度 $\rho$ 较大; 另一方面, GRF 和 FSh 仅考虑截断以内的水, 蛋白质体积所占比例较大, 水数量密度可忽略不计. 对于膜蛋白体系, 可参考 Roux 课题组 [74] 提出的方法做相应的修正.

以上介绍的 C-CpHMD 属于对电荷插值, 实现电荷对反应坐标的线性响应. 实际上, 由于库仑势对电荷线性依赖, 库仑势和电荷两者的线性插值是等效的. 因为两种情况下, 关于插值变量 (反应坐标 $\lambda$) 负的一阶导数 (作用在虚粒子上的合外力) 是相等的. 然而, 并不是所有和静电势相关的能量项和电荷线性相关, 如 PME 算法中对点电荷自相互作用和净电荷的修正项 ((20) 式)[66]. 所以, 为了更好描述电荷变化对滴定动力学的影响, 基于截断的 GRF 和 FSh 较适合对静电势进行插值的 C-CpHMD, 因为它们的静电势保留了对电荷的线性依赖. 德国马克思普朗克研究所的 Grubmüller 课题组 [75] 在分子模拟软件 GROMACS 中开发的

C-CpHMD 便是对势函数进行插值. 最近, 芬兰的 Groenhof 课题组 [76,77] 基于该模型进行代码优化, 并实现基于 CHARMM 力场的 CpHMD 模拟. 然而, 该模型采用了显性溶剂 PME, 而不是基于截断的 GRF 或 FSh. 其次, 该模型没有像 Shen 课题组 [64] 一样考虑有限尺寸效应. 另一方面, 同样是对势能进行插值, Brooks 课题组 [63,71] 基于显性溶剂的 C-CpHMD 模型合理地采用了基于截断的 FSh. 除了以上正弦函数形式, Grubmüller 课题组和 Brooks 课题组提出了其他将 $\lambda$ 约束在区间 [0, 1] 的方法. 例如, Grubmüller 课题组 [75] 提出了余弦形式. Brooks 课题组 [78] 提出一个较复杂的指数形式. 对于显性溶剂 C-CpHMD, 体系电中性是一项重要的约束条件, Shen 课题组 [66] 和 Grubmüller 课题组 [79] 均采用了可滴定水分子实现体系净电荷恒等于 0. 然而, Brooks 课题组 [71] 的显性溶剂 C-CpHMD 还未考虑该约束. 因此, 为了避免溶质与其镜像的静电相互作用, 需对 FSh 静电势设置较小截断值.

从理论上看, Shen 课题组 [66] 开发的基于 PME 的 C-CpHMD 可应用于分子力场能描述的任何体系, 似乎没有改进的空间. 实际上, 一个酸性氨基酸残基的去质子化或一个碱性氨基酸残基的质子化可诱导周围可极化原子 (原子核外电子云的中心偏离原子核) 或基团 (组氨酸咪唑环上的电子离域) 形成偶极子 [80]. 偶极子与电荷相互吸引, 一定程度上加强了该氨基酸残基带电状态的稳定性. 然而, 传统力场下电荷分布是固定的, 不会因为滴定引起周围电场的变化而做出调整, 这可能导致可滴定氨基酸残基偏爱电中性, 特别是位于蛋白质内部的氨基酸残基 [66]. 基于以上考虑, 如果采用极化力场 (如 CHARMM 的 Drude[81]), C-CpHMD 的精度将得到进一步的提升. 其次, 大部分 CpHMD(包括该模型) 没有考虑质子化和去质子化对键相互作用的影响 [44].

随着显性溶剂 CpHMD 的快速发展, 急需解决质子化态和构象的采样问题. 2006 年 Brooks 课题组 [82] 率先将基于温度的副本交换 (replica exchange) 算法应用到 C-CpHMD, 即将副本以一定的概率交换到较高温度, 借助热涨落提高 CpHMD 模拟的采样. 受到哈密顿量副本交换算法的启发, 2011 年 Shen 课题组 [56] 提出了基于 pH 的副本交换算法: 将副本以一定的概率 $p$ 交换到较高的 pH, 提高去质子化态的采样; 或交换到较低的 pH, 提高质子化态的采样 ((35) 式). 因为实际进行交换的 pH 只存在于 $U^{\mathrm{pH}}$ ((23) 式), 交换前后总能量的变化 $\Delta/\beta$ 可简化为仅含 $U^{\mathrm{pH}}$ 的表达式 ((36) 式). 交换 pH 后, 两个副本将在新的 pH 条件下 (或新的 $U^{\mathrm{pH}}$) 进行采样. 该算法效率极高, 同时操作简单, 已被应用到其他 CpHMD 模型 [83–86]. 为了增强质子化态空间采样, 美国国立卫生研究院 NIH 的 Brooks 课题组 [87] 提出结合包络分布采样 (enveloping distribution sampling, EDS) 和哈密顿量副本交换 (Hamiltonian replica exchange, HREX). EDS 通过定义一个参数 $s$ 标度状态间的能垒. 较小的 $s$ 对应较平滑的能垒, 方便了状态间的转化. 然而, 能垒的消除促进了虚拟中间态的采样, 这将影响物理态的采样. 为了避免中间态的采样, 在 EDS 基础上利用 HREX 提高离散的质子化态空间的采样效率. 接着, 该课题组 [86] 加入以上基于 pH 的副本交换, 构成二维的副本交换. 从算法的角度, 该方法确实提高了采样效率, 但代价是产生大量的副本以及模拟过程中副本的频繁通讯, 对计算能力要求较高. 近期, 为了在有限 GPU 显卡数量的条件下实现基于 pH 的副本交换, Shen 课题组 [88] 提出了副本同步交换.

$$p=\begin{cases}1, & \Delta\leqslant 0,\\ \exp(-\Delta), & \Delta>0,\end{cases} \tag{35}$$

$$\begin{aligned}\Delta=\beta(&U^{\mathrm{pH}}\left(\{\lambda_j\};\mathrm{pH}'\right)+U^{\mathrm{pH}}\left(\{\lambda'_j\};\mathrm{pH}\right)\\&-U^{\mathrm{pH}}\left(\{\lambda_j\};\mathrm{pH}\right)-U^{\mathrm{pH}}\left(\{\lambda'_j\};\mathrm{pH}'\right)),\end{aligned} \tag{36}$$

其中, $p$ 是副本交换的概率; $U^{\mathrm{pH}}(\{\lambda_j\};\mathrm{pH})$ 和 $U^{\mathrm{pH}}(\{\lambda'_j\};\mathrm{pH}')$ 是两个副本交换前的 $U^{\mathrm{pH}}$. 将以上两项的 pH 和 pH′ 进行互换, 得到 $U^{\mathrm{pH}}(\{\lambda_j\};\mathrm{pH}')$ 和 $U^{\mathrm{pH}}(\{\lambda'_j\};\mathrm{pH})$.

除了副本交换, 另一种增强采样的方法是对生物大分子进行粗粒化 (coarse graining, CG), 减少模拟体系中粒子的数量, 从而降低了构象空间的自由度. 该方法通常被应用于具有较大空间和时间尺度的生物过程, 如蛋白质折叠、多肽聚集和物质跨膜转运等 [89]. 近几年, 研究者们开始将 CG 与 CpHMD 结合, 发展 CpHMD 的粗粒化模型 [90–93]. 值得一提的是, 提出 Martini 粗粒化力场的 Marrink 课题组 [92] 已在分子模拟软件 GROMACS 中实现了 CpHMD 的粗粒化模拟.

## 2.2 基于 PB 的 $pK_a$ 预测模型

实际上, 如果只考虑单个结构, 可以用 PB 方程计算相对去质子化自由能 $\Delta\Delta G = \Delta G - \Delta G^{\rm mod}$. 其中, $\Delta G^{\rm mod}$ 是某可离子化氨基酸 A 在游离状态下去质子化自由能改变量:

$$\Delta G^{\rm mod} = G^{\rm mod}\left(\mathrm{A}^-\right) - G^{\rm mod}\left(\mathrm{AH}\right), \tag{37}$$

式中 $G^{\rm mod}\left(\mathrm{A}^-\right)$ 和 $G^{\rm mod}\left(\mathrm{AH}\right)$ 分别是去质子化 ( $\mathrm{A}^-$ ) 和质子化 ( AH ) 状态的自由能. 同理, 当该氨基酸参与蛋白质的合成, 它在蛋白质中的去质子化自由能改变量 $\Delta G$ 表示为

$$\Delta G = G\left(\mathrm{A}^-\right) - G\left(\mathrm{AH}\right). \tag{38}$$

基于蛋白质环境不影响成键作用部分 $\Delta G_{\rm Bond}$ (见 (4) 式) 的假设, 以上两个自由能改变量的差可表示为

$$\begin{aligned}\Delta G - \Delta G^{\rm mod} = {} & \left(G_{\rm PB}\left(\mathrm{A}^-\right) - G_{\rm PB}\left(\mathrm{AH}\right)\right) \\ & - \left(G^{\rm mod}_{\rm PB}\left(\mathrm{A}^-\right) - G^{\rm mod}_{\rm PB}\left(\mathrm{AH}\right)\right),\end{aligned} \tag{39}$$

其中, 下标 PB 表示用 PB 方程分别计算等式右边 4 个状态下的静电能. 令 $\Delta G(\mathrm{AH}) = G_{\rm PB}\left(\mathrm{AH}\right) - G^{\rm mod}_{\rm PB}\left(\mathrm{AH}\right)$ 和 $\Delta G(\mathrm{A}^-) = G_{\rm PB}\left(\mathrm{A}^-\right) - G^{\rm mod}_{\rm PB}\left(\mathrm{A}^-\right)$, 可得到

$$\Delta G + \Delta G_{\rm PB}\left(\mathrm{AH}\right) = \Delta G^{\rm mod} + \Delta G_{\rm PB}\left(\mathrm{A}^-\right), \tag{40}$$

其中, $\Delta G_{\rm PB}\left(\mathrm{AH}\right)$ 和 $\Delta G_{\rm PB}\left(\mathrm{A}^-\right)$ 分别表示在水溶剂中将质子化 ( AH ) 和去质子化 ( $\mathrm{A}^-$ ) 的氨基酸放入蛋白质的静电能改变量. 基于该等式, 可以得到如图 5 所示的热力学循环 (thermodynamic cycle). 相对去质子化自由能 $\Delta\Delta G$ 可表示为

$$\Delta\Delta G = \Delta G_{\rm PB}\left(\mathrm{A}^-\right) - \Delta G_{\rm PB}\left(\mathrm{AH}\right), \tag{41}$$

接着, 将 $\Delta\Delta G$ 代入关系式 $\Delta \mathrm{p}K_{\rm a} = \Delta\Delta G/\left(k_{\rm B}T\ln 10\right)$ 计算 $\mathrm{p}K_{\rm a}$ 偏移量 $\Delta \mathrm{p}K_{\rm a}$. 最后, 利用 (5) 式计算 $\mathrm{p}K_{\rm a}$. 可见, 热力学循环 4 个状态的静电能计算决定了 $\mathrm{p}K_{\rm a}$ 的预测精度. 目前, 基于 PB 计算静电能并预测蛋白质 $\mathrm{p}K_{\rm a}$ 的方法包括 MCCE[17,94], H++[18], APBS[19], DelPhiPKa[20,95,96] 以及 PypKa[21]. 其中, MCCE 和 PypKa 利用 MC 对侧链二面角进行采样, 一定程度上提高了预测精度, 但总体精度仍低于 CpHMD, 说明了空间构象充分采样的重要性[15]. PB 方程的参数主要是介电常数, 原子的电荷和半径, 因此容易拓展到其他类型的体系. 例如, 除了蛋白质, DelPhiPKa 也适用于 DNA 和 RNA. 除了蛋白质单体, H++也考虑了含有配体的复合物.

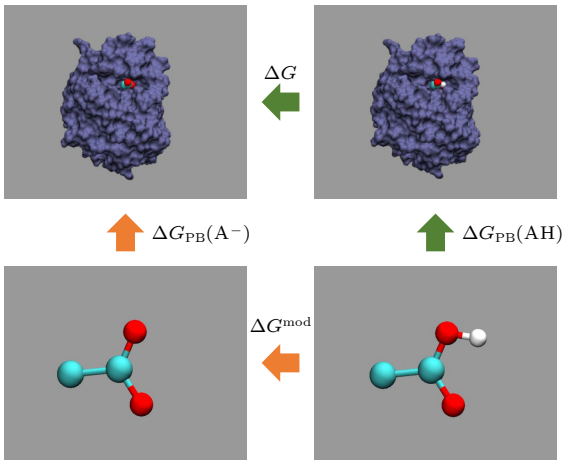


图 5 相对去质子化自由能计算的热力学循环

Fig. 5. Thermodynamic cycle of relative deprotonation free energy calculation.

## 2.3 基于经验函数的 $pK_a$ 预测模型

以上物理模型 (CpHMD 和基于 PB 的模型) 需要计算体系的静电能, 计算复杂度较高. 为了进一步提高 $\mathrm{p}K_{\rm a}$ 计算的效率 (例如将单个蛋白的 $\mathrm{p}K_{\rm a}$ 计算时长缩短到秒量级), 2005 年哥本哈根大学的 Jensen 课题组[23] 提出了一组经验函数 PropKa 分别描述点电荷相互作用 (Coulomb force)、去溶剂化效应 (desolvation) 和氢键相互作用 (hydrogen bonding) 对 $\mathrm{p}K_{\rm a}$ 偏移量的贡献:

$$\Delta \mathrm{p}K_{\rm a} = \Delta \mathrm{p}K_{\rm a}^{\rm Columb} + \Delta \mathrm{p}K_{\rm a}^{\rm Desolv} + \Delta \mathrm{p}K_{\rm a}^{\rm HBond}. \tag{42}$$

以上 3 项的函数均采用分段的一次函数, 计算复杂度低, 已被应用到蛋白质单体[23], 蛋白质和小分子配体的复合物[97]. 然而, 该版本的 PropKa 没区分可滴定氨基酸残基是处于蛋白质的表面还是内部.

为此, 2011 年 Jensen 课题组[24] 提出了改进的 PropKa 3.0. 新版本考虑了相同的 $\Delta \mathrm{p}K_{\rm a}$ 决定因子, 将 (42) 式的氢键相互作用导致的 $\Delta \mathrm{p}K_{\rm a}^{\rm HBond}$ 和去溶剂化效应导致的 $\Delta \mathrm{p}K_{\rm a}^{\rm Desolv}$ 归为自能 $\Delta \mathrm{p}K_{\rm a}^{\rm Self}$. 不同的是, PropKa 3.0 采取了一个折中的方案, 即部分使用能量公式. 例如, 点电荷相互作用采用经典的库仑势. 去溶剂化效应采用了和 GB 模型中求解有效波恩半径的倒数 ( $1/\alpha$ ) 类似的原子体积 ( $V$ ) 除以原子间距离的四次方 ( $r^4$ ). 此外, 蛋白质表面和内部被赋予不同的介电常数. 对于氢键相互作用, 则保留了一次函数形式. 该模型的参数化基于谷氨酸和天冬氨酸的 $\mathrm{p}K_{\rm a}$ 实验值, 对酸性氨基酸的预测能力接近 CpHMD[98]. 然而, 该模型对碱性

氨基酸 (如 Lys 和 His) 的预测效果较差[25].

## 2.4 基于机器学习的 p$K_a$ 预测模型

上述 PropKa 经验函数的提出较大程度依赖于科学家的先验知识. 理论上, 如果有足够多的 p$K_a$ 实验测量值, 可以结合数据和机器学习算法训练出一个经验函数, 而不需要依靠已有的知识. 2018 年 波兰华沙大学 Siedlecki 课题组[99] 提出首个基于深度学习的蛋白质配体结合亲和性 (binding affinity) 预测模型. 这里的配体通常指具有几何结构的小分子. 我们知道, p$K_a$ 表征某可滴定基团去质子的难易程度. 换一种表达, p$K_a$ 代表蛋白质和质子的结合亲和性. 可见, 蛋白质配体结合亲和性预测方法对 p$K_a$ 预测具有参考价值[25].

由于实验条件的限制, 迄今为止蛋白质可滴定氨基酸残基的 p$K_a$ 实验测量值不到两千个[100,101]. 于是, 本课题组采用基于隐性溶剂 GBNeck2 的 CpHMD[48] 建立了一个蛋白质 p$K_a$ 数据集 (包含 12809 个 p$K_a$)[25]. 2021 年 12 月, 本课题组提出了国际上首个基于机器学习的蛋白质 p$K_a$ 预测模型 DeepKa, 证明了引入人工智能方法解决蛋白质 p$K_a$ 预测问题的可行性[25]. 本课题组对现有的 p$K_a$ 数据库 PKAD[100](包含 1350 个蛋白质 p$K_a$ 实验测量值) 进行数据清洗, 得到了测试集 EXP67S. 首先, 根据氨基酸序列相似性比对排除了冗余数据. 剩下的 67 个蛋白质的 470 个 Asp, Glu, Lys 或 His 的 p$K_a$ 构成数据集 EXP67. 接着, 对 EXP67 进行欠采样, 使得不同 $\Delta$p$K_a$ 区域分布均匀. 最后剩下的 167 个 p$K_a$ 为该模型的测试集 EXP67S. 该测试集的优势将在下文的多模型比对体现出来 (图 6). 模型的大部分输入特征以及三维卷积神经网络 (convolutional neural network, CNN) 框架均借鉴 Siedlecki 课题组[99] 提出的 Pafnucy 模型. 值得一提的是, 为了解决截断导致的边界问题, DeepKa 采用格点电荷 (Siedlecki 课题组[99] 采用原子电荷) 描述对 p$K_a$ 预测精度起决定性作用的静电环境[25]. 虽然 DeepKa 第一版本的预测精度高于 PropKa 3.0, 但是和 CpHMD 还存在一定差距[25]. 此外, 该工作只测试了 DeepKa 的总体性能, 并未对特定的问题 (如酶催化中心或无序蛋白) 进行讨论.

2022 年 1 月, 美国卡内基·梅隆大学 Lsayev 课题组[26] 开发了基于神经网络势 ANI-2X 和原子环境矢量 AVE 的深度学习模型 pKa-ANI. 然而, 该模型将所有的实验数据用于模型的训练, 不利于对其性能进行客观的评价. 另外, 该模型对结构敏感, 需要在预处理阶段对初始结构进行能量最小化, 否则将得到不合理的预测结果[26]. 2022 年 3 月, 美国约翰斯·霍普金斯大学 Damjanovic 课题组[27] 测试了 4 种基于树的机器学习算法. 其中, XGB-WMa 表现最好. 该小组同样采用有限的实验数据来训练和测试模型. 为了建立有效的模型, 他们在特征描述上加入了较多的经验知识: 首先, 统计可滴定基团参与的氢键数量; 其次, 计算可滴定基团的溶剂可及表面积 (solvent accessible surface area, SASA); 最后, 根据是否带电或亲水对可滴定基团附近氨基酸残基进行分类. 显然, 以上特征基本上覆盖了 PropKa 模型中影响 p$K_a$ 偏移量的 3 个关键因素: 氢键相互作用、去溶剂化效应和点电荷相互作用. 2022 年 7 月, Reis 课题组[102] 利用基于 PB 的 PypKa 建立了包含 1200 万个 p$K_a$ 值的数据集, 并基于该数据集开发了深度学习模型 PKAI[28]. 为了提高精度, 在 PKAI 基础上对损失函数进行正则化处理, 从而得到 PKAI+. 然而, PKAI+在其他测试集 (如 EXP67S) 的表现与 PKAI 相似, 说明上述的正则化处理缺乏普适性[29]. 因此, 如果没有特别说明, 下文只讨论 PKAI.

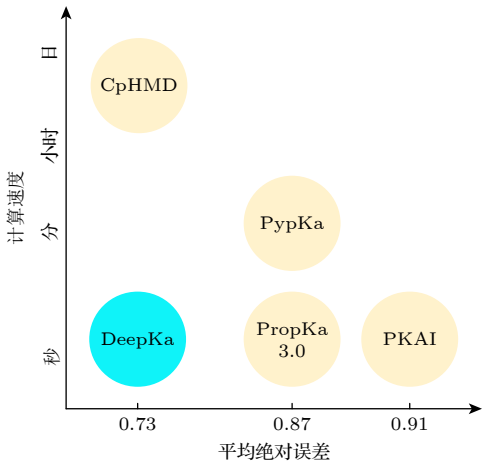


图 6 p$K_a$ 预测模型性能对比

Fig. 6. Comparison of existing p$K_a$ predictors.

2023 年 5 月, 本课题组发布了 DeepKa 的最新版本[29]. 该版本的输入特征和模型框架与旧版本相同, 仅仅是增加了训练和验证集的 p$K_a$ 样本量. 这些样本出自 549 个蛋白质的 26552 个 Asp, Glu, Lys 和 His. 相对旧版本, 该版本预测性能更接近 CpHMD. 此外, 在这个工作中特定的蛋白质体系

被用于进一步评估 DeepKa 的可靠性. 例如, 酶催化中心具有复杂的静电环境, 是 p$K_a$ 预测的一个重要挑战. 新版本通过 p$K_a$ 计算准确预测了 5 个酶催化中心的质子供体. 除了具有稳定三维结构的蛋白, 该模型也可被应用于无序蛋白. 理论预测 p$K_a$ 偏移量较小的滴定位点往往容易做到预测精确, 但难以做到预测相关, 而即使在 p$K_a$ 偏移量小于 1.0 的情况下, 理论和实验仍然表现出较高的相关性, 证明了该模型的高鲁棒性[29]. 如无特别说明, 下文的 DeepKa 代表该新版本.

上述基于 AI 的模型均采用 PKAD 中的实验数据来训练或测试模型. 然而, pKa-ANI, XGB-WMa 和 PKAI 忽略了存在于 PKAD 的冗余数据 (例如一个蛋白质有两组相同的 p$K_a$ 值), 这可能导致过拟合. 其次, PKAD 中大多数 p$K_a$ 处于参考值 p$K_a^{mod}$ 附近, 因此测试结果并不能反应模型真实的预测能力[25]. 值得一提的是, 本课题组创建的测试集 EXP67S 不存在以上两个问题, 可较为客观地对模型进行评价[25]. 研究发现, 除了在实验和理论相关性方面仍旧低于 CpHMD, DeepKa 的预测精度明显高于其他主流 p$K_a$ 预测模型, 包括 PypKa, PropKa, PKAI 和 pKa-ANI[29]. 其中, PypKa 代表基于 PB 的模型, PropKa 代表基于经验函数的模型, PKAI 和 pKa-ANI 代表其他 AI 模型. 基于树的 XGB-WMa 没有开放源代码, 所以无法利用 EXP67S 对其进行测试. 因此, XGB-WMa 不参与下面的模型讨论. 同时考查精度和速度, 图 6 展示了 5 个模型的预测性能. 其中, 平均绝对误差用于表征模型的精度. 显而易见, 如果以 PropKa 的速度和 CpHMD 的精度作为参照, 目前只有 DeepKa 能提供准确的高通量 p$K_a$ 计算[29]. 最近, 加拿大国家研究委员会 Sulea 课题组[103] 比较了现有的 7 种高通量 p$K_a$ 预测模型, 包括基于经验函数的 PropKa 3.0[24], 基于深度学习的 DeepKa[29]、PKAI 和 PKAI+[28] 以及基于 PB 方程的 DelPhiPKa[95]、MCCE2[94] 和 H++[18]. 该研究指出在以上高通量模型中 DeepKa 的精度最高, 与图 6 的结论一致.

## 3 结 论

pH 与温度、压强一样是基本的环境参量. 传统的分子动力学假设溶剂是中性水 (pH=7.0), 不考虑其他 pH 条件; 此外, 传统分子动力学假设电荷是固定的, 不受溶质静电场的影响. 以上两个假设限制了传统分子动力学进一步探究细胞中许多与 pH 相关的生物过程, 而可靠的 p$K_a$ 计算将有助于解决该难题. 本综述主要介绍了 4 类主流的 p$K_a$ 预测方法. 显然, 对于不同理论的 p$K_a$ 预测模型, 其适用范围也存在差异. 首先, 不论何种特定的问题, 如果不要求高通量计算, 可采用预测精度较高但计算效率较低的 CpHMD. 当涉及非水溶性蛋白 (如膜蛋白) 的 p$K_a$ 计算, 目前理论上可行的模型为基于杂化溶剂[37,56] 或显性溶剂[66,67] 的 CpHMD. 另一方面, 需要开发高通量的 p$K_a$ 预测模型, 从而满足工业界批量的 p$K_a$ 计算需求. 由于隐性溶剂的理论局限性和实验条件的限制, 上述的高通量模型仅适用于水溶性蛋白. 对于水溶性蛋白质单体的 p$K_a$ 计算, 在所有高通量模型中 DeepKa 无疑是最优的选择[29,103]. 若只关心酸性氨基酸残基 (如 Asp 和 Glu) 的质子化态, 也可考虑 PropKa 3.0[24]. 而对于主要的 4 种可离子化氨基酸残基 (Asp, Glu, Lys 和 His) 以外的可滴定基团 (如 Cys 和 Tyr), 可考虑基于 PB 的模型 (如 H++[18] 和 PypKa[21]).

随着计算机软件和硬件的快速发展, 国际著名的美国药物设计公司薛定谔 (Schrödinger) 开始尝试利用自由能微扰 (free energy perturbation, FEP) 方法计算 p$K_a$, 说明蛋白质 p$K_a$ 理论计算开始引起工业界的关注[104]. 值得一提的是, 基于机器学习的 p$K_a$ 预测模型虽处于起步的阶段 (2021 年至今), 却已表现出和物理模型同水平的预测精度, 例如本课题组开发的 DeepKa. 我们相信: AI 模型有可能突破先验知识, 在不久的将来提供更为高效的预测; 利用物理模型 CpHMD 建立的 p$K_a$ 数据集 PHMD549 和基于 p$K_a$ 数据库 PKAD 建立的测试集 EXP67S 将为基于机器学习的 p$K_a$ 预测工具的研发奠定基础[29]. 最近, 基于 DeepKa 本课题组开发了国内首个蛋白质 p$K_a$ 在线计算平台 (http://www.computbiophys.com/DeepKa/main), 这对未来参与到人工智能驱动的新药研发产业具有重要意义[105,106].

## 参考文献

[1] Casey J R, Grinstein S, Orlowski J 2010 *Nat. Rev. Mol. Cell Biol.* **11** 50
[2] Qian H, Wu X L, Du X M, Yao X, Zhao X, Lee J, Yang H Y, Yan N 2020 *Cell* **182** 98
[3] Yang G H, Zhou R, Zhou Q, Guo X F, Yan C Y, Ke M, Lei J L, Shi Y G 2019 *Nature* **565** 192

[4] Chung H S, Piana-Agostinetti S, Shaw D E, Eaton W A 2015 *Science* **349** 1504

[5] Nasica-Labouze J, Nguyen P H, Sterpone F, Berthoumieu O, Buchete N, Cote S, Simone A D, Doig A J, Faller P, Garcia A, Laio A, Li M S, Melchionna S, Mousseau N, Mu Y, Paravastu A, Pasquali S, Rosenman D J, Strodel B, Tarus B, Viles J H, Zhang T, Wang C, Derreumaux P 2015 *Chem. Rev.* **115** 3518

[6] Morrow B H, Payne G F, Shen J 2015 *J. Am. Chem. Soc.* **137** 13024

[7] Kumar A, Hossain R A, Yost S A, Bu W, Wang Y, Dearborn A D, Grakoui A, Cohen J I, Marcotrigiano J 2021 *Nature* **598** 521

[8] Singharoy A, Maffeo C, Delgado-Magnero K H, Swainsbury D J K, Sener M, Kleinekathofer U, Vant J W, Nguyen J, Hitchcock A, Israelewitz B, Teo I, Chandler D E, Stone J E, Phillips J C, Pogorelov T V, Mallus M I, Chipot C, Luthey-Schulten Z, Tieleman D P, Hunter C N, Schulten K 2019 *Cell* **179** 1098

[9] Shimizu H, Tosaki A, Kaneko K, Hisano T, Sakurai T, Nukina N 2008 *Mol. Cell Biol.* **28** 3663

[10] Ellis C R, Shen J 2015 *J. Am. Chem. Soc.* **137** 9543

[11] Thurlkill R L, Grimsley G R, Scholtz J M, Pace C N 2006 *Protein Sci.* **15** 1214

[12] Jensen J H, Li H, Robertson A D, Molina P A 2005 *J. Phys. Chem. A* **109** 6634

[13] Baptista A M, Martel P J, Petersen S B 1997 *Proteins* **27** 523

[14] Shi C, Wallace J A, Shen J K 2012 *Biophys. J.* **102** 1590

[15] Qing R, Hao S L, Smorodina E, Jin D, Zalevsky A, Zhang S G 2022 *Chem. Rev.* **122** 14085

[16] Henderson J A, Liu R, Harris J A, Huang Y D, de Oliveria V M, Shen J D 2022 *Liv. J. Comput. Mol.* **4** 1563

[17] Georgescu R E, Alexov E G, Gunner M R 2002 *Biophys. J.* **83** 1731

[18] Anandakrishnan R, Aguilar B, Onufriev A V 2012 *Nucleic Acids Res.* **40** W537

[19] Dolinsky T J, Nielsen J E, McCammon J A, Baker N A 2004 *Nucleic Acids Res.* **32** 665

[20] Wang L, Li L, Alexov E 2015 *Proteins.* **83** 2186

[21] Reis Pedro B P S, Vila-Viçosa D, Rocchia W, Machuqueiro M 2020 *J. Chem. Inf. Model.* **60** 4442

[22] Huang Y D, Yue Z, Tsai C C, Henderson J A, Shen J 2018 *J. Phys. Chem. Lett.* **9** 1179

[23] Li H, Robertson A D, Jensen J H 2005 *Proteins* **61** 704

[24] Olsson Mats H M, Søndergaard C R, Rostkowski M, Jensen J H 2011 *J. Chem. Theory Comput.* **7** 525

[25] Cai Z T, Luo F F, Wang Y X, Li E L, Huang Y D 2021 *ACS Omega* **6** 34823

[26] Gokcan H, Lsayev O 2022 *Chem. Sci.* **13** 2462

[27] Chen A Y, Lee J, Damjanovic Ana, Brooks B R 2022 *J. Chem. Theory Comput.* **184** 2673

[28] Reis Pedro B P S, Bertolini M, Montanari F, Rocchia W, Machuqueiro M, Clevert D A 2022 *J. Chem. Theory Comput.* **18** 5068

[29] Cai Z T, Liu T Z, Lin Q L, He J H, Lei X W, Luo F F, Huang Y D 2023 *J. Chem. Inf. Model* **63** 2936

[30] Baptista A M, Teixeira V H, Soares C M 2002 *J. Chem. Phys.* **117** 4184

[31] Lee M S, Salsbury F R, Brooks Ⅲ C L 2004 *Proteins* **56** 738

[32] Mongan J, Case D A, McCammon J A 2004 *J. Comput. Chem.* **25** 2038

[33] Meng Y, Roitberg A E 2010 *J. Chem. Theory Comput.* **6** 1401

[34] Swails J M, York D M, Roitberg A E 2014 *J. Chem. Theory Comput.* **10** 1341

[35] Machuqueiro M, Baptista A M 2006 *J. Phys. Chem. B* **110** 2927

[36] Sequeira J G N, Rodrigues F E P, Silva T G D, Reis Pedro B P S, Machuqueiro M 2022 *J. Phys. Chem. B.* **126** 7870

[37] Huang Y D, Chen W, Dotson D L, Beckstein O, Shen J 2016 *Nat. Commun.* **7** 12940

[38] Stern H A 2007 *J. Chem. Phys.* **126** 164112

[39] Essmann U, Perera L, Berkowitz M L, Darden T, Lee H, Pedersen L G 1995 *J. Chem. Phys.* **103** 8577

[40] Chen Y, Roux B 2015 *J. Chem. Theory Comput.* **11** 3919

[41] Radak B K, Chipot C, Suh D, Jo S, Jiang W, Philips J C, Schulten K, Roux B 2017 *J. Chem. Theory Comput.* **13** 5933

[42] Wang R X, Fang X L, Lu Y P, Yang C Y, Wang S M 2005 *J. Med. Chem.* **48** 4111

[43] Pieri E, Ledentu V, Sahlin M, Dehez F, Olivucci M, Ferre N 2019 *J. Chem. Theory Comput.* **15** 4535

[44] de Oliveria V M, Liu R, Shen J 2022 *Curr. Opin. Struct. Biol.* **77** 102498

[45] Kong X, Brooks III C L 1996 *J. Chem. Phys.* **105** 2414

[46] Khandogin J, Brooks Ⅲ C L 2005 *Biophys. J.* **89** 141

[47] Nguyen H, Maier J, Huang H, Perrone V, Simmerling C 2014 *J. Am. Chem. Soc.* **136** 13959

[48] Huang Y D, Harris R C, Shen J 2018 *J. Chem. Inf. Model.* **58** 1372

[49] Liu R, Yue Z, Tsai C C, Shen J 2019 *J. Am. Chem. Soc.* **141** 6553

[50] Harris R C, Liu R, Shen, J 2020 *J. Chem. Theory Comput.* **16** 3689

[51] Liu R, Zhan S, Che Y, Shen J 2022 *J. Med. Chem.* **65** 1525

[52] Yao X, Chen C, Wang Y, Dong S, Liu Y, Li Y, Cui Z, Gong W, Perrett S, Yao L, Lamed R, Bayer E A, Cui Q, Feng Y 2020 *Sci. Adv.* **6** eabd7182

[53] Verma N, Henderson J A, Shen J 2020 *J. Am. Chem Soc.* **142** 21883

[54] Arthur E J, Brooks III C L 2016 *J. Comput. Chem.* **37** 2171

[55] Harris R C, Shen J 2019 *J. Chem. Inf. Model.* **59** 4821

[56] Wallace J A, Shen J K 2011 *J. Chem. Theory Comput.* **7** 2617

[57] Henderson J A, Huang Y D, Beckstein O, Shen J 2020 *Proc. Natl. Acad. Sci. U. S. A.* **117** 25517

[58] Chen W, Huang Y D, Shen J 2016 *J. Phys. Chem. Lett.* **7** 3961

[59] Yue Z, Li C, Voth G A, Swanson J M J 2019 *J. Am. Chem. Soc.* **141** 13421

[60] Vo Q N, Mahinthichaichan P, Shen J, Ellis C R 2021 *Nat. Commun.* **12** 984

[61] Li Z, Zhang X, Wang Q, Li C, Zhang N, Zhang X, Xu B, Ma B, Schrader T E, Coates L, Kovalevsky A, Huang Y D, Wan Q 2018 *ACS Catal.* **8** 8058

[62] Tsai C C, Yue Z, Shen J 2019 *J. Am. Chem. Soc.* **141** 15092

[63] Goh G B, Knight J L, Brooks III C L 2012 *J. Chem. Theory Comput.* **8** 36

[64] Wallace J A, Shen J K 2012 *J. Chem. Phys.* **137** 184105

[65] Chen W, Shen J K 2014 *J. Comput. Chem.* **35** 1986

[66] Huang Y D, Chen W, Wallace J A, Shen J 2016 *J. Chem. Theory Comput.* **12** 5411

[67] Harris J A, Liu R, de Oliveira V M, Vázquez-Montelongo E A, Henderson J A, Shen J 2022 *J. Chem. Theory Comput.* **18** 7510

[68] Chen W, Wallace J A, Yue Z, Shen J K 2013 *Biophys. J.*

**105** L15
[69] Wallace J A, Shen J K 2009 *Methods Enzymol.* **466** 455
[70] Ullmann G M 2003 *J. Phys. Chem. B* **107** 1263
[71] Goh G B, Hulbert B S, Zhou H, Brooks Ⅲ C L 2014 *Proteins* **82** 1319
[72] Webb H, Tynan-Connolly B M, Lee G M, Farrell D, O' Meara F, Sondergaard C R, Teilum K, Hewage C, Mclntosh L P, Nielsen J E 2010 *Proteins* **79** 685-702
[73] Rocklin G J, Mobley D L, Dill K A, Hunenberger P H 2013 *J. Chem. Phys.* **139** 184103
[74] Bignucolo O, Chipot C, Kellenberger S, Roux B 2022 *J. Phys. Chem. B.* **126** 6868
[75] Donnini S, Tegeler F, Groenhof G, Grubmüller H 2011 *J. Chem. Theory Comput.* **7** 1962
[76] Aho N, Buslaev P, Jansen A, Bauer P, Groenhof G, Hess B 2022 *J. Chem. Theory Comput.* **18** 6148
[77] Buslaev P, Aho N, Jansen A, Bauer P, Hess B, Groenhof G 2022 *J. Chem. Theory Comput.* **18** 6134
[78] Knight J L, Brooks Ⅲ C L 2011 *J. Comput. Chem.* **32** 3423
[79] Donnini S, Ullmann R T, Groenhof G, Grubmüller H 2016 *J. Chem. Theory Comput.* **12** 1040
[80] Huang Y D, Shuai J 2013 *J. Phys. Chem. B* **117** 6138
[81] Lemkul J A, Huang J, Roux B, MacKerell A D 2016 *Chem. Rev.* **116** 4983
[82] Khandogin J, Brooks Ⅲ C L 2006 *Biochemistry* **45** 9363
[83] Itoh S G, Damjanović A, Brooks B R 2011 *Proteins* **79** 3420
[84] Dashti D S, Meng Y, Roitberg A E 2012 *J. Phys. Chem. B.* **116** 8805
[85] Swails J M, Roitberg A E 2012 *J. Chem. Theory Comput.* **8** 4393
[86] Lee J, Miller B T, Damjanovic A, Brooks B R 2015 *J. Chem. Theory Comput.* **11** 2560
[87] Lee J, Miller B T, Damjanovic A, Brooks B R 2014 *J. Chem. Theory Comput.* **10** 2738
[88] Henderson J A, Verma N, Harris R, Shen J 2020 *J. Chem. Phys.* **153** 115101
[89] Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid A E, Kolinski A 2016 *Chem. Rev.* **116** 7898
[90] Bennett W D, Chen A W, Donnini S, Groenhof G, Tieleman D P 2013 *Can. J. Chem.* **91** 839
[91] da Silva F L B, Sterpone F, Derreumaux P 2019 *J. Chem. Theory Comput.* **15** 3875
[92] Crünewald F, Souza P C T, Abdizadeh H, Barnoud J, de Vries A H, Marrink S J 2020 *J. Chem. Phys.* **153** 024118
[93] Reilley D J, Wang J, Dokholyan N V, Alexandrova A N 2021 *J. Chem. Theory Comput.* **17** 4583
[94] Song Y, Mao J, Gunner M R 2009 *J. Comput. Chem.* **30** 2231
[95] Wang L, Zhang M, Alexov E 2016 *Bioinformatics* **32** 614
[96] Pahari S, Sun L, Basu S, Alexov E 2018 *Proteins* **86** 1277
[97] Bas D C, Rogers D M, Jensen J H 2008 *Proteins* **73** 765
[98] Sun Z, Wang X, Song J 2017 *J. Chem Inf. Model.* **57** 1621
[99] Stepniewska-Dziubinska M M, Zielenkiewicz P, Siedlecki P 2018 *Bioinformatics* **34** 3666
[100] Pahari S, Sun L, Alexov E 2019 *Database* **2019** baz024
[101] Ancona N, Bastola A, Alexov E 2023 *J. Comput. Biophys. Chem.* **22** 515
[102] Reis Pedro B P S, Clevert D A, Machuqueiro M 2022 *Bioinformatics* **38** 297
[103] Wei W, Hogues H, Sulea T 2023 *J. Chem. Inf. Model.* **63** 5169
[104] Coskun D, Chen W, Clark A J, Lu C, Hardr E D, Wang L, Friesner R A, Miller E B 2022 *J. Chem. Theory Comput.* **18** 7193
[105] Hagg A, Kirschner K N 2023 *J. Chem. Inf. Model.* **63** 4505
[106] Bueschbell B, Caniceiro A B, Suzano P M S, Machuqueiro M, Rosário-Ferreira N, Moreira I S 2022 *Drug Resist. Updat.* **60** 100811

SPECIAL TOPIC—Machine learning in biomolecular simulations

# Progress in protein p$K_a$ prediction*

Luo Fang-Fang Cai Zhi-Tao Huang Yan-Dong †

(*College of Computer Engineering, Jimei University, Xiamen* 361021, *China*)

( Received 20 August 2023; revised manuscript received 1 September 2023 )

**Abstract**

The pH value represents the acidity of the solution and plays a key role in many life events linked to human diseases. For instance, the β-site amyloid precursor protein cleavage enzyme, BACE1, which is a major therapeutic target of treating Alzheimer's disease, functions within a narrow pH region around 4.5. In addition, the sodium-proton antiporter NhaA from *Escherichia coli* is activated only when the cytoplasmic pH is higher than 6.5 and the activity reaches a maximum value around pH 8.8. To explore the molecular mechanism of a protein regulated by pH, it is important to measure, typically by nuclear magnetic resonance, the binding affinities of protons to ionizable key residues, namely p$K_a$ values, which determine the deprotonation equilibria under a pH condition. However, wet-lab experiments are often expensive and time consuming. In some cases, owing to the structural complexity of a protein, p$K_a$ measurements become difficult, making theoretical p$K_a$ predictions in a dry laboratory more advantageous. In the past thirty years, many efforts have been made to accurately and fast predict protein p$K_a$ with physics-based methods. Theoretically, constant pH molecular dynamics (CpHMD) method that takes conformational fluctuations into account gives the most accurate predictions, especially the explicit-solvent CpHMD model proposed by Huang and coworkers (2016 *J. Chem. Theory Comput.* **12** 5411) which in principle is applicable to any system that can be described by a force field. However, lengthy molecular simulations are usually necessary for the extensive sampling of conformation. In particular, the computational complexity increases significantly if water molecules are included explicitly in the simulation system. Thus, CpHMD is not suitable for high-throughout computing requested in industry circle. To accelerate p$K_a$ prediction, Poisson-Boltzmann (PB) or empirical equation-based schemes, such as H++ and PropKa, have been developed and widely used where p$K_a$ values are obtained via one-structure calculations. Recently, artificial intelligence (AI) is applied to the area of protein p$K_a$ prediction, which leads to the development of DeepKa by Huang laboratory (2021 *ACS Omega* **6** 34823), the first AI-driven p$K_a$ predictor. In this paper, we review the advances in protein p$K_a$ prediction contributed mainly by CpHMD methods, PB or empirical equation-based schemes, and AI models. Notably, the modeling hypotheses explained in the review would shed light on future development of more powerful protein p$K_a$ predictors.

**Keywords:** molecular dynamics, Poisson-Boltzmann equation, machine learning, p$K_a$ prediction

**PACS:** 87.15.ap, 87.14.E–, 87.10.Vg, 87.15.A– **DOI:** 10.7498/aps.72.20231356

* Project supported by the National Natural Science Foundation of China (Grant Nos. 11804114, 62006096), the Natural Science Foundation of Fujian Province, China (Grant Nos. 2023J01329, 2020J05146), the Natural Science Foundation of Xiamen, China (Grant No. 3502Z20227205), and the Scientific Starting Research Foundation of Jimei University, China (Grant No. ZQ2020027).

† Corresponding author. E-mail: yandonghuang@jmu.edu.cn

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

# 蛋白质p$K_a$预测模型研究进展

罗方芳　蔡志涛　黄艳东

# Progress in protein p$K_a$ prediction

Luo Fang-Fang　Cai Zhi-Tao　Huang Yan-Dong

引用信息 Citation: Acta Physica Sinica, 72, 248704 (2023)　DOI: 10.7498/aps.72.20231356
在线阅读 View online: https://doi.org/10.7498/aps.72.20231356
当期内容 View table of contents: http://wulixb.iphy.ac.cn

---