arXiv:2611.30527v1 [cond-mat.stat-mech] 23 Nov 2026

CRISPR-Cas9系统中R环形成的非平衡动力学

Yu Liu 和 Bing Miao*Center of Materials Science and Optoelectronics Engineering,
College of Materials Science and Opto-Electronic Technology,
University of Chinese Academy of Sciences (UCAS), Beijing 100049, China.

CRISPR-Cas9的结合特异性由R环的逐步形成所决定：在这一过程中，sgRNA逐步置换非靶DNA链，并与靶链形成碱基对。本文提出一个非平衡动力学框架，将R环形成建模为跨越二十个碱基对的连续时间马尔可夫生灭过程。每个转变速率均根据Eyring过渡态理论和Bell模型推导，并结合三种物理贡献：解链前沿处由B型到A型螺旋构象转换产生的扭转应变、耦合相邻碱基对的碱基堆叠修正，以及Cas9构象变化提供的变构功。我们进一步推导了该过程的熵产生，并将其分解为活化项、堆叠项、扭转项和变构项。所得预测模型无需外部热力学数据库，在两个独立的、于体内测量的失活Cas9-sgRNA错配变体文库中，分别达到0.84和0.82的斯皮尔曼相关系数。错配类型依赖性由每种类型对应的一个扭转偏差和一个配对权重体现。我们的模型再现了种子区敏感性、PAM近端活性梯度，以及dDrG错配的中等代价，同时指出高活性双重错配尾部是当前自由能映射不足以描述的区域。该框架定量预测了DNA超螺旋和序列背景如何调节结合亲和力，并为sgRNA设计提供了机制依据。

I. 引言

CRISPR-Cas9系统中，单引导 RNA（sgRNA）引导对互补 DNA 序列的切割 [1], ，已成为基因组编辑与调控的通用工具 [2]。然而，其特异性并不完全，因为 sgRNA 能够容忍错配，因此 Cas9 可能在非预期的基因组位点结合或切割 [3, 4]。这种脱靶活性仍是治疗应用面临的主要障碍，而控制该活性需要对错配如何调节 Cas9 的结合与切割进行定量的物理描述。

Cas9 的结合特异性由 R环的逐步形成所决定。在这一过程中，sgRNA 排开非靶DNA链，并与靶DNA链杂交，由此形成的 RNA–DNA 异源双链被容纳在蛋白质的通道中 [5, 6]。被排开的非靶DNA链固定在蛋白质表面的沟槽中，因此，引导序列与靶序列的匹配情况会沿着异源双链逐个碱基地被读取；与侧翼的 B型 DNA 相比，该异源双链采用 A型构象。单分子实验表明，R环并非通过一次协调事件形成。它从 PAM近端向远端沿方向性以离散的碱基对步进方式增长，并且 DNA 解旋，而非 sgRNA–DNA 杂交，是限速步骤 [7–9]。这些步骤会受到引导序列–靶序列错配和 DNA超螺旋的影响，因此扭转应力与 R环传播在机械上相耦合 [10]。从 PAM近端开始、跨度约为十个碱基对的种子区，决定了 CRISPR 的特异性干扰 [11], ，而高通量结合实验已经绘制出脱靶结合和解离的序列决定因素 [12]。切割与结合的错配耐受性特征在定性上存在差异，这意味着不能仅凭结合亲和力推断切割，并且结合与切割是同一连续过程中的两个不同阶段。R环形成具有逐步、可逆且由机械驱动的特性，因此需要采用主方程进行描述；在该方程中，未解旋碱基对数量的概率分布通过连续时间马尔可夫生灭过程不断演化。

CRISPR-Cas9靶向的理论研究沿着两条路线展开。第一条路线从动力学角度描述引导序列-靶序列杂交：杂交体形成动力学模型再现了 Cas 核酸酶的脱靶规则 [13], ，其扩展模型预测 SpCas9 的结合与切割活性，并且比基于序列的评分更好地对脱靶位点进行分类 [14], ；而 Feng 等人拟合的可逆马尔可夫链也属于同一类模型 [15]。第二条路线为每个错配分配一个平衡自由能，如用于对脱靶位点评分的双链体能量模型 [16]。在这两种路线中，特异性均依据杂交体的平衡占据率来判断；此外，平衡路线还受到已列出参数的错配类型数量较少的限制。

然而，判别本身并不是一种平衡性质。Cas9通过一个构象检查点对R环进行校对，只有在感知到完全互补时，才会促使酶进入切割过程 [17], ，因此，错配靶标是通过动力学而非结合亲和力被拒绝的。动力学校对解释了这种检查点如何提高准确性，但这需要付出代价，因为额外的判别需要一个不可逆的

* bmiao@ucas.ac.cn

由化学能驱动的步骤 [18]。误差校正的热力学明确揭示了这一代价：判别器的保真度受到其必须维持的熵产生所限制 [19], ，而随机热力学则提供了计算给定反应网络中这一代价的工具 [20–22]。

然而，R环形成的熵产生仍未得到量化。在此，我们推导了连续R环过程的熵产生，并将其分解为活化、堆叠、扭转和变构贡献，从而将 Cas9 的保真度与其反应路径上的耗散联系起来。

我们将这一框架与 Feng等人 [15], 的结合数据进行比较。他们在活细菌细胞中，利用两个独立文库测量了 dCas9-sgRNA 的 3400 多种饱和单错配和双错配变体。我们的出发点是，R环形成是一个连续过程，而非平衡结合问题，因为引导序列从 PAM 近端开始，每次侵入双链体一个碱基对。因此，我们将其建模为状态空间 $\{0, 1, \ldots, 20\}$ 上的连续时间马尔可夫生灭过程，其中状态表示连续解开的碱基对数量，每次转变涉及一个碱基对，包括双链体解旋，随后是 sgRNA 与靶链结合。速率根据 Eyring过渡态理论 [23] 和 Bell模型 [24], 得出，并综合了三种物理贡献：解链前沿处 B型向A型转变产生的扭转应变、耦合相邻碱基对的碱基堆叠修正，以及 Cas9 提供的变构功。根据同一过程，我们推导出轨迹的精确熵产生，它分为来自二十次转变的介质部分和来自边界状态的系统部分。

由此得到的物理约束模型不使用外部热力学数据库。在七个独立分区上，该模型达到 R^2 =0.68 ± 0.06 和 0.64 ± 0.05，斯皮尔曼相关系数分别为0.84和0.82。该模型再现了种子区敏感性、PAM近端活性梯度以及 dDrG错配的中等代价，其中错配类型依赖性由每种类型的一个扭转偏差和一个配对权重承载；同时，它还确定了高活性双重错配尾部是当前自由能映射不适用的状态区域。

该框架将Cas9的特异性与一组少量物理量联系起来。扭转参数和堆叠参数定量预测超螺旋和序列背景如何改变结合亲和力；此外，无需完整的错配数据库，即可在新的文库上重新训练该参数结构，从而为改进现有的 sgRNA设计工具提供一条机制性途径 [25, 26]。

本文结构安排如下。第II节阐述R环形成的序贯模型及其所蕴含的转移速率。第III节分析单个碱基对动力学。第IV节构建形成效率的热力学模型。第V节推导熵产生及其分解。第VI节在两个文库上验证受物理约束的预测模型。最后，我们在第VII节给出结论。详细推导见附录A和B。

II. R环形成的顺序模型

R环形成是一个DNA双链逐步解链的过程，始于PAM近端，并逐个碱基对向远端推进。由于Cas9识别一个20 bp的间隔区，完整的R环形成包含20个顺序步骤，从PAM近端到远 PAM 端依次标记为 $i = 1$、2、……、20，如图1所示。每个步骤对应一个碱基对从闭合状态转变为解链状态，并包含两个严格有序的微观子过程：DNA解链和sgRNA结合。只有在位置 $i-1$ 解链后，位置 i的DNA双链才能解链；只有在位置 i 打开后，sgRNA才能与该位置的靶链配对。由于实验中尚未观察到跳过中间步骤的情况，我们排除远端碱基对直接打开，以及近端步骤完成前启动远端步骤这两种可能性。

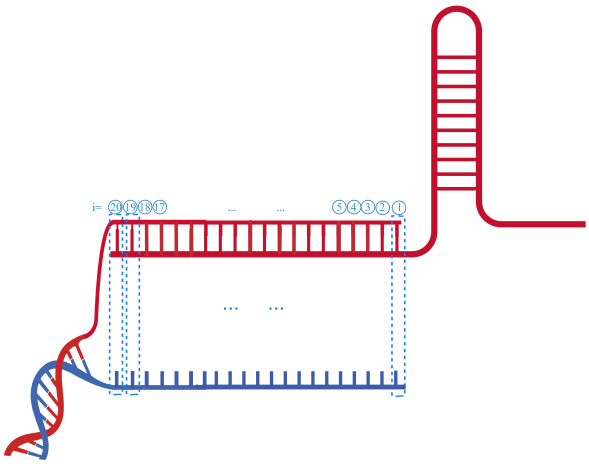


图 1. 20 bp 间隔区沿线碱基配对示意图。带圈数字标出间隔区上的位置，每个数字都通过一条引线连接到其所标示的碱基对。双链体在右侧延伸至间隔区之外。R环形成按顺序以 20 个碱基对为步长进行，从 PAM 近端到远 PAM 端依次编号为 $i = 1$、……、20。

根据此图，我们将状态变量$i(t)$定义为时刻 t连续解开的碱基对数量。这里的“连续”是指，从PAM近端开始计数的第1至 i个碱基对均已解开，而碱基对 $i+1$ 仍保持闭合。该变量取值为0，1、……、20，其中 $i = 0$ 是所有20个碱基对均闭合的初始状态，而 $i = 20$ 是最终

R环完全形成的状态。

令 $Q_i(t)$ 表示系统在时间 t处于状态 i 的概率。由于系统在任一给定时刻必须且只能处于一个状态，因此归一化条件自动成立：

$$\sum_{i=0}^{20} Q_i(t) = 1, \quad \forall\, t \geq 0. \tag{1}$$

在 Cas9-sgRNA复合物首次与含 PAM 的靶DNA结合的初始时刻，尚无碱基对解旋。因此

$$Q_0(0) = 1, \quad Q_i(0) = 0 \quad (i \geq 1). \tag{2}$$

该过程受两个基本物理约束支配。首先，状态空间具有顺序性。系统必须依次经过状态 0、1、2、.....、20，不得跳过。从状态 i只能转移到 $i{+}1$ 或$i{-}1$ 。禁止直接跳转到 $i \pm 2$ 或更远的状态，因此该过程具有局部动力学。其次，局部相互作用约束要求任一时刻最多只有一个碱基对发生状态改变。每次向前或向后的转移都恰好涉及一步。总之，这些约束将该系统定义为状态空间上的连续时间马尔可夫生灭过程。

在生灭框架下，$Q_i(t)$的演化由主方程支配。

$$\begin{aligned} \frac{dQ_0}{dt} &= -\lambda_{1,0}\, Q_0 + \lambda_{0,1}\, Q_1, \\ \frac{dQ_i}{dt} &= \lambda_{i,i-1}\, Q_{i-1} - (\lambda_{i-1,i} + \lambda_{i+1,i})\, Q_i \\ &\quad + \lambda_{i,i+1}\, Q_{i+1}, \qquad (1 \leq i \leq 19), \\ \frac{dQ_{20}}{dt} &= \lambda_{20,19}\, Q_{19} - \lambda_{19,20}\, Q_{20}. \end{aligned} \tag{3}$$

此处，$\lambda_{i,\,j}$ 表示从状态 j 到状态 i的转移速率矩阵元素。概率守恒和稳态递归结构可直接由方程（3）的数学形式得出，而与$\lambda_{i,\,i-1}$ 和 $\lambda_{i-1,\,i}$的具体取值无关。最终状态 $i = 20$表示已完成的 R环。它并非严格意义上的吸收态，因为方程（3）中保留了反向转移$20 \to 19$ 。切割通过因子 $1-\mu_{\text{cleavage}}$引入，该因子乘以反向转移的速率，因此已被切割的复合物无法再回退，其中 μ_{cleavage} 的定义见下文。

III. 单碱基对动力学

在主方程框架中，只有当系统处于状态 $i{-}1$时，正向速率$\lambda_{i,\,i-1}$ 才会被激活，这意味着前 $i{-}1$ 个碱基对已经解旋。因此，单个碱基对动力学的自然表述基于一个条件概率方程，其条件是所有先前步骤均已完成。在 20 个反应位点中，每个位点都遵循相同的单步动力学方案。DNA 解旋和第 i 步的 sgRNA结合子过程可写为以下化学反应

$$\text{T-DNA} + \text{NT-DNA} \underset{k^i_{\text{off},1}}{\overset{k^i_{\text{on},1}}{\rightleftharpoons}} \text{dDNA}, \tag{4}$$

$$\text{sgRNA} + \text{T-DNA} \underset{k^i_{\text{off},2}}{\overset{k^i_{\text{on},2}}{\rightleftharpoons}} \text{RNA–DNA}, \tag{5}$$

其中，dDNA 表示完整的双链DNA，RNA–DNA 表示RNA–DNA杂交双链，T-DNA 表示靶DNA链，NT-DNA 表示非靶DNA链。速率常数 $k^i_{\text{on},\ 1}$ 和 $k^i_{\text{off},\ 1}$ 控制步骤 i中 DNA双链的闭合与解旋，而$k^i_{\text{on},\ 2}$ 和 $k^i_{\text{off},\ 2}$ 控制 sgRNA 与暴露的靶链的结合与解离。下标 1 和 2 用于区分这两个子过程。

由于这些反应涉及同一 DNA双链或 RNA–DNA 杂交体上互补片段的运动，因此它们属于分子内过程。任一子过程的速率方程都具有统一形式

$$\frac{d[C]}{dt} = k^i_{\text{on}}\,[O] - k^i_{\text{off}}\,[C], \tag{6}$$

其中，$[O]$ 和 $[C]$ 分别表示分子内开放位点和闭合位点的比例。因此，k^i_{on}和 k^i_{off} 的量纲为 s^{-1}。

现在，我们根据上述反应方案推导解链概率的演化方程。定义$q_i(t)$为条件概率：已知前 $i{-}1$个碱基对已经解链，在时间 t时第 i个碱基对已解链的概率：

$$q_i(t) = \frac{P_1^i(t)}{P_1^{i-1}(t)}, \tag{7}$$

其中，$P_1^i(t)$表示在时间 t时至少前 i 个碱基对已解链的概率。主方程给出概率分布 $Q_m(t)$，即恰有 m 个碱基对解链时处于该状态的概率。相比之下，P_1^i 是前 i 个碱基对全部解链的累积概率，而不论第 $i{+}1$个及其后续碱基对的状态如何。这两种描述通过以下恒等式相关联：

$$P_1^i(t) = \sum_{m=i}^{20} Q_m(t). \tag{8}$$

这一关系的物理基础在于解旋过程具有顺序性。由于该过程从靠近 PAM 的一端开始，并且逐步进行、不跳过任何位置，因此，至少有 i 个碱基对解旋，就等价于系统处于某个状态 $m \geq i$。由式（8）还可进一步得到

$$Q_i = P_1^i - P_1^{i+1}, \quad (1 \leq i \leq 19), \tag{9}$$

$$Q_{20} = P_1^{20}. \tag{10}$$

此外，P_1^0 表示PAM识别的概率。由于我们关注PAM结合后的R环形成，因此设定 $P_1^0 = 1$。

结合物理图像与反应方案，DNA解链子过程的 q_i 动力学为

$$\dot{q}_i = k^i_{\mathrm{off},1}\,(1-q_i) - (1-\mu_{\mathrm{cleavage}}\delta_{i,20})k^i_{\mathrm{on},1}\,q_i\,(1-q_{i+1}), \tag{11}$$

其中， μ_{cleavage} 是 Cas9 复合物的切割效率，$0 \le \mu_{\mathrm{cleavage}} < 1$，且 $\delta_{i,\,20}$ 是克罗内克 delta。正向项 $k^i_{\mathrm{off},\,1}(1-q_i)$ 表示在第 1 至 $i-1$ 个碱基对已打开的情况下，碱基对 i 仍处于闭合状态的概率（$1-q_i$）。这使碱基对 i 能够解旋。反向项$k^i_{\mathrm{on},\,1}q_i(1-q_{i+1})$ 表示顺序性约束。只有当碱基对 i 处于打开状态且碱基对 $i{+}1$ 处于闭合状态时，碱基对 i 才能重新缠绕。在此约束下，步骤 $i{-}1$ 的动力学远慢于步骤 i。观察步骤 i时，前 $i{-}1$个碱基对的状态似乎被冻结，因此 $\dot{P}_1^{i-1} \approx 0$。于是，式（11）可改写为

$$\frac{\dot{P}_1^i}{P_1^{i-1}} = k^i_{\mathrm{off},1}\,(1-q_i) - (1-\mu_{\mathrm{cleavage}}\delta_{i,20})k^i_{\mathrm{on},1}\,q_i\,(1-q_{i+1}). \tag{12}$$

对于 sgRNA 结合子过程，我们假设只有在解旋使该位点的靶DNA链暴露后，才会触发该位点 i 的结合。定义 $P_2^i(t)$ 为在时间 t，参与步骤 i 的 sgRNA片段与暴露的靶链结合的概率。结合概率 P_2^i 满足

$$\frac{dP_2^i}{dt} = k^i_{\mathrm{on},2}\,(1-P_2^i) - k^i_{\mathrm{off},2}\,P_2^i, \tag{13}$$

其中 $k^i_{\mathrm{on},\,2}(1-P_2^i)$ 为结合速率，与未结合位点的比例成正比；$k^i_{\mathrm{off},\,2}P_2^i$ 为解离速率，与已结合位点的比例成正比。

现在，我们用微观速率常数表示主方程的跃迁速率 $\lambda_{i,\,i-1}$ 和 $\lambda_{i-1,\,i}$。回顾以下关系式

$$Q_i = P_1^i - P_1^{i+1}, \quad (1 \le i \le 19), \tag{14}$$

$$Q_{20} = P_1^{20}, \tag{15}$$

经过严格推导（见附录 A），可得精确表达式

$$\lambda_{i,i-1} = k^i_{\mathrm{off},1}, \tag{16}$$

$$\lambda_{i-1,i} = (1-\mu_{\mathrm{cleavage}}\delta_{i,20})k^i_{\mathrm{on},1}. \tag{17}$$

IV. R环形成效率的热力学模型

在本节中，我们将上文建立的动力学与热力学联系起来。在 20 个子步骤中的每一步，首先发生的微观过程都是 DNA 双链体的解旋，其中涉及多种因素：单导 RNA 碱基的吸引力、促进解旋的 Cas9 构象变化，以及一个常被忽视的因素——A型 RNA–DNA 杂交体与 B-型双链体交界处积累的扭转应变；每个碱基对从 B型转变为 A型时，都必须克服这种应变。

要量化这一扭转贡献，需要进行力学分析。sgRNA 和非靶DNA链均与靶DNA链结合，在靶链上形成一个边界：边界一侧由 sgRNA 占据，另一侧则由非靶链占据。在此边界的两侧，螺旋采取不同的构象。sgRNA 与靶链形成的 RNA–DNA 杂交体通常呈 A型螺旋，而 DNA 双链体则保持 B型结构。这两种螺旋构象的每个碱基对的扭转角不同，A型约为 32.73°，B型约为 34.29°。沿螺旋轴前进一个碱基对的距离，即为一步使连接点前移的距离，为 $\Delta x \approx 0.34$ nm。因此，在靶链上的边界处，A型和 B型区段分别施加扭矩 τ_A 和 τ_B，其中

$$\tau_A = C\,\frac{32.73^\circ}{\Delta x}, \qquad \tau_B = C\,\frac{34.29^\circ}{\Delta x}, \tag{18}$$

其中 $d\theta/ds \approx \Delta\theta_0/\Delta x$ 为 C 扭转刚度。两条螺旋由蛋白质固定，因此它们之间的扭转差异由连接点来调节。我们将这一残余错配取为 $\Delta\theta_0 \approx 1.56^\circ$，即 A型与 B型扭转角之差，这与 Cas9-sgRNA-DNA复合物冷冻电镜结构中观察到的近乎笔直的异源双链一致 [5,6]。如果边界以速度 v从 B型一侧向 A型一侧推进，扭矩瞬时提供的功率为

$$P(t) = \tau_{\mathrm{ex}}\,\partial_t\theta = (\tau_A-\tau_B)(v\,\partial_s\theta) = \frac{v\,(\tau_A-\tau_B)^2}{C}. \tag{19}$$

边界穿过位点 i 的转变，对应于在位置 $i{-}1$ 形成 A型结构，随后在位置 i解旋 DNA双链，即 B→A 转变。因此，推动连接点前进一个步长所需提供的扭转能为

$$\Delta W_0 = \int_0^{\Delta x/v} P(t)\,dt = \frac{v\,(\tau_A-\tau_B)^2}{C}\,\frac{\Delta x}{v} = \frac{C\,(\Delta\theta_0)^2}{\Delta x}. \tag{20}$$

式（20）在理想条件下成立，且 $C/\Delta x$可解释为有效局部扭转刚度。式（20）中的扭转能是每步的代价：它提高连接点处打开一个碱基对的能垒，并降低将其闭合的能垒；下文的速率常数中也以此符号计入该能量。在完全匹配的杂交体中，sgRNA 与靶链以标准的沃森–克里克几何构型配对，即腺嘌呤与胸腺嘧啶配对，鸟嘌呤与胞嘧啶配对。然而，如果 sgRNA 与靶DNA序列并非完全互补，就可能出现错配，甚至完整的非互补位点。在双链核酸中，错配碱基对通常会引起局部螺旋畸变，从而改变相邻碱基对之间的扭转角。这些角度扰动的幅度和方向取决于

错配类型和局部序列环境。此外，错配的影响并不局限于当前碱基对。它可以传播至一个或两个相邻碱基对，尽管其幅度会逐渐减小。

为考虑错配，我们将 ΔW_i 推广为

$$\Delta W_i = \frac{C\,(\Delta\theta_i)^2}{\Delta x}, \tag{21}$$

其中， $\Delta\theta_i$ 表示位点 i处两种螺旋构象之间的扭转角差:

$$\Delta\theta_i = \Delta\theta_0 + \delta\theta_i^{\mathrm{mis}} + \sum_{k=1}^{2} \alpha_k\, \delta\theta_{i-k}^{\mathrm{mis}}. \tag{22}$$

这里， $\delta\theta_i^{\mathrm{mis}}$是位点 i处错配导致的角偏差，而求和项则考虑了两个相邻位置处错配的影响，并按α_k加权。

应用Eyring过渡态理论和Bell模型，我们得到双链闭合与打开速率常数的以下表达式:

$$k_{\mathrm{on},1}^{i} = \kappa\,\frac{k_B T}{h}\,\exp\Big[-\Big(\Delta G_{1,i}^{\neq} - \Delta G_{\mathrm{sf1}}^{i} - \Delta W_i N_A + P_2^i\, W_{\mathrm{cas},i}^{\mathrm{on}}\Big)/RT\Big], \tag{23}$$

$$k_{\mathrm{off},1}^{i} = \kappa\,\frac{k_B T}{h}\,\exp\Big[-\Big(\Delta G_{2,i}^{\neq} + \Delta G_{\mathrm{sf1}}^{i} + \Delta W_i N_A - P_2^i\, W_{\mathrm{cas},i}^{\mathrm{off}}\Big)/RT\Big], \tag{24}$$

其中， $\Delta G_{1,\,i}^{\neq}$ 和 $\Delta G_{2,\,i}^{\neq}$ 分别是碱基对闭合和打开的活化自由能。 $\Delta G_{\mathrm{sf1}}^{i}$是过渡态能垒中的碱基堆叠贡献。 $W_{\mathrm{cas},\,i}^{\mathrm{on}}$和 $W_{\mathrm{offcas}\,i}$分别是Cas9构象变化在正向和反向R环传播过程中所做的功， N_A 是阿伏伽德罗常数， h 是普朗克常数， κ 是透射系数。需要指出的是，式（23）和式（24）涉及隐式粗粒化。我们假设 Cas9 在 R 环形成过程中所做的实际功与结合概率 P_2^i成正比。

现在我们分析 sgRNA结合子过程。如前所述， P_2^i 是在靶DNA链的位点 i 已经暴露的条件下定义的。因此，忽略非靶链对结合的影响。速率常数为

$$k_{\mathrm{on},2}^{i} = \kappa_{\mathrm{on}}\,\frac{k_B T}{h}\,\exp\left(-\frac{\Delta G_{3,i}^{\neq} - \Delta G_{\mathrm{sf2}}^{i}}{RT}\right), \tag{25}$$

$$k_{\mathrm{off},2}^{i} = \kappa_{\mathrm{off}}\,\frac{k_B T}{h}\,\exp\left(-\frac{\Delta G_{4,i}^{\neq} + \Delta G_{\mathrm{sf2}}^{i}}{RT}\right), \tag{26}$$

其中， $\Delta G_{3\ ,\ i}^{\neq}$ 和$\Delta G_{4\ ,\ i}^{\neq}$ 分别是结合与解离的活化自由能，而 $\Delta G_{\mathrm{sf2}}^{i}$ 是此子过程的碱基堆叠势垒修正项。

结合步骤的平衡常数为

$$K_2^i = \frac{k_{\mathrm{on},2}^{i}}{k_{\mathrm{off},2}^{i}} = \exp\left(\frac{\Delta G_{4,i}^{\neq} - \Delta G_{3,i}^{\neq} + 2\,\Delta G_{\mathrm{sf2}}^{i}}{RT}\right), \tag{27}$$

且相应的平衡自由能变化满足 $\Delta G_{i,\ \mathrm{rd}}^{\circ} = -RT\ln K_2^i$, 从而得到

$$\Delta G_{i,\mathrm{rd}}^{\circ} = \Delta G_{3,i}^{\neq} - \Delta G_{4,i}^{\neq} - 2\,\Delta G_{\mathrm{sf2}}^{i}. \tag{28}$$

现在我们考虑 Cas9-sgRNA复合物在靶DNA上形成R环的连续动力学过程。将 PAM近端标记为 $i=1$ ，将远端标记为 $i=20$，每个子步骤 i 都涉及两个微观过程：（i）DNA双链体的解链与重新闭合；（ii）sgRNA与互补靶链之间的碱基配对。我们将片段 i 处双链闭合和打开的基本速率常数分别表示为 $k_{\mathrm{on},\ 1}^{i}$ 和$k_{\mathrm{off},\ 1}^{i}$，将 sgRNA结合概率表示为 P_2^i，将 Cas9构象变化的有效功表示为 $W_{\mathrm{cas},\ i}^{\mathrm{on}}$和 $W_{\mathrm{cas},\ i}^{\mathrm{off}}$，并将闭合和打开的活化自由能表示为 $\Delta G_{1\ ,\ i}^{\neq}$ 和$\Delta G_{2\,i,\ }^{\neq}$。

消去正向和反向过渡态表达式中的公共前因子，可得到片段 i 处速率比的热力学形式:

$$\ln\frac{k_{\mathrm{off},1}^{i}}{k_{\mathrm{on},1}^{i}} = \frac{\Delta G_{1,i}^{\neq} - \Delta G_{2,i}^{\neq} - 2\,\Delta G_{\mathrm{sf1}}^{i} - 2\,\Delta W_i N_A + P_2^i\big(W_{\mathrm{cas},i}^{\mathrm{on}} + W_{\mathrm{cas},i}^{\mathrm{off}}\big)}{RT}, \tag{29}$$

其中， $\Delta G_{\mathrm{sf1}}^{i}$ 为碱基堆叠势垒修正， ΔW_i 为每个子步的扭转能， N_A 为阿伏伽德罗常数。为简洁起见，我们引入

$$\Delta\Delta G_i^{\neq} \equiv \Delta G_{1,i}^{\neq} - \Delta G_{2,i}^{\neq}, \qquad \mathcal{W}_i^{\mathrm{cas}} \equiv W_{\mathrm{cas},i}^{\mathrm{on}} + W_{\mathrm{cas},i}^{\mathrm{off}}, \tag{30}$$

因此，方程（29）变为

$$\ln\frac{k_{\mathrm{off},1}^{i}}{k_{\mathrm{on},1}^{i}} = \frac{\Delta\Delta G_i^{\neq} - 2\,\Delta G_{\mathrm{sf1}}^{i} - 2\,\Delta W_i N_A + P_2^i\,\mathcal{W}_i^{\mathrm{cas}}}{RT}. \tag{31}$$

因子 P_2^i 反映了 Cas9 变构激活与 DNA结合之间的协同性。

将单个位点处的净驱动功定义为

$$W_i^{\mathrm{drive}} \equiv P_2^i\,\mathcal{W}_i^{\mathrm{cas}} - 2\,\Delta W_i N_A, \tag{32}$$

公式（31）表明，速率比由能量差和外功共同决定。当位点未结合（$P_2^i = 0$）时，构象项消失，该比值仅由活化能差、堆叠修正和

扭转贡献决定。

由于DNA解链控制着R环形成，我们将单个位点处双链体正向与逆向速率系数之比作为局部R环形成效率的度量， η_i。该度量不考虑切割。考虑到透射系数 κ 在正向和逆向之间存在差异，我们得到

$$\eta_i \propto \exp\Big[(RT)^{-1}\Big(\Delta\Delta G_i^{\neq} - 2\,\Delta G_{\mathrm{sf1}}^{i} + W_i^{\mathrm{drive}}\Big)\Big]. \quad (33)$$

当靶DNA与sgRNA存在错配时，三种效应会抑制形成效率。首先，错配会破坏氢键网络并削弱堆叠作用，从而提高局部能垒 $\Delta G_{\mathrm{sf1}}^{i}$。其次，扭转扰动会增加扭转应变能 ΔW_i，因此必须针对连接处额外提供更多功。第三，配对的不稳定性会降低结合概率 P_2^i 以及Cas9贡献的变构功。三者都会降低 η_i ，从而抑制R环形成。

sgRNA结合概率 P_2^i 可以在稳态下独立确定。根据公式（13）的稳态条件，

$$\begin{aligned} P_{2,\mathrm{ss}}^{i} &= \frac{k_{\mathrm{on},2}^{i}}{k_{\mathrm{on},2}^{i} + k_{\mathrm{off},2}^{i}} \\ &= \left[1 + \exp\!\left(\frac{\Delta G_{i,\mathrm{rd}}^{\circ}}{RT}\right)\right]^{-1}, \end{aligned} \quad (34)$$

其中， $\Delta G_{i,\ \mathrm{rd}}^{\circ} = \Delta G_{3\ ,\ i}^{\neq} - \Delta G_{4\ ,\ i}^{\neq} - 2\,\Delta G_{\mathrm{sf2}}^{i}$ 是结合过程的平衡自由能变化。该概率直接调节 Cas9 在片段 i 处提供的驱动功 $\mathcal{W}_i^{\mathrm{cas}}$，从而形成反馈回路。更强的结合会提供更多驱动功，进而促进 DNA 进一步解链。

上述模型决定了DNA解链和RNA–DNA结合的动力学，并为正向和逆向过程设定了不同的速率常数。速率常数的不对称性会产生熵产-生，这是表征非平衡系统不可逆性的一个关键量。因此，量化R环形成过程中的熵产- 生，可以揭示驱动力的来源以及该过程固有的不可逆程度。

V. 熵产生与不可逆性

为了量化 R环形成的不可逆性，我们使用沿着一条从完全缠绕状态到完全解链状态的轨迹产生的熵 σ 。对于正向路径 Γ^{+} 及其时间反演路径 Γ^{-}，

$$\hat{\sigma} = k_B \ln\left(\frac{P_{+}[\Gamma_{+}]}{P_{-}[\Gamma_{-}]}\right) \quad (35)$$

由于我们考虑的是一条完整形成 R环的轨迹，路径从 Q_0 运行到 Q_{20}。两条路径概率中的驻留时间相互抵消，因此只剩下沿序列的速率比。对于熵产生，系统熵变化 $k_B \ln[Q_0(0)/Q_{20}(T)]$ 的贡献与介质熵变化相比相对较小；后者等于 k_B 乘以前向和反向跃迁速率之比的对数。推导过程见附录 B，最终得到

$$\hat{\sigma} = k_B \sum_{k=1}^{20} \ln \frac{k_{\mathrm{off},1}^{k}}{\left(1 - \mu_{\mathrm{cleavage}}\delta_{k,20}\right) k_{\mathrm{on},1}^{k}}, \quad (36)$$

以便

$$\hat{\sigma} = k_B \sum_{k=1}^{19} \ln \eta_k + k_B \ln \frac{\eta_{20}}{1 - \mu_{\mathrm{cleavage}}}. \quad (37)$$

利用速率比的表达式，即式 (33)，

$$\ln \frac{k_{\mathrm{off},1}^{k}}{k_{\mathrm{on},1}^{k}} = (RT)^{-1}\Big[\Delta\Delta G_k^{\neq} - 2\Delta G_{\mathrm{sf1}}^{k} + W_k^{\mathrm{drive}}\Big], \quad (38)$$

并代入式 (36)，我们得到

$$\hat{\sigma} = \frac{k_B}{RT} \sum_{k=1}^{20} \big(\Delta\Delta G_k^{\neq} - 2\Delta G_{\mathrm{sf1}}^{k} + W_k^{\mathrm{drive}}\big) - k_B \ln(1 - \mu_{\mathrm{cleavage}}). \quad (39)$$

该表达式将熵产生、局部活化自由能差、碱基堆叠以及 Cas9 变构做功的四项贡献分离开来，其中后两项通过 W_k^{drive} 体现。错配通过提高 $\Delta G_{\mathrm{sf1}}^{k}$ 和 ΔW_k 并降低 P_2^k 来影响这些项，因此每个错配都会使 R环通路变得不利。由此，熵产生框架为编辑特异性提供了一种随机热力学描述。

VI. 物理约束活性预测模型

前几节为 dCas9-sgRNA 靶标结合提供了热力学和动力学框架。现在，我们用实验对其进行检验：所有能量尺度要么取自测定的结构，要么由一个校准常数确定，且不使用外部热力学数据库。

我们使用 Feng等人 [15], 的数据，他们通过$sacB$ 反向筛选系统，将失活 Cas9 的结合与细菌生长关联起来，

并通过对筛选文库和对照文库进行高通量测序，将与生长关联的保留转化为每个变体的富集程度。r0 文库的重复测量与 Pearson $r = 0.91$ [15]一致，表明测得的活性具有足够的可重复性，可作为定量目标进行分析。研究分析了两个文库，即 r0 和 r349，数据分别来自原始工作簿 Binding activity r0.xlsx 和 Binding activity r349.xlsx。r0 文库包含野生型以及 1704 个突变体，其中 60 个带有单错配，1644 个带有双错配。其野生型 sgRNA 为 UACACUUGAACUACCGCGAG，其 DNA 靶链为 ATGTGAACTTGATGGCGCTC。r349 文库包含野生型以及 1718 个突变体，其中 60 个带有单错配，1658 个带有双错配。r0 的活性范围为 -8.494 至 5。420，均值为 -2.758，标准差为 1。952；r349 的活性范围为 -9.764 至 4。380，均值为 -3.728 ，标准差为 2。182。Feng等人将活性定义为相对于野生型的对数$_2$ 富集度，

$$\mathrm{Fitness}_i = \log_2 \frac{\text{Read count}_{i,\text{selective}}}{\text{Read count}_{i,\text{control}}},$$
$$\text{Binding activity}_i = \mathrm{Fitness}_i - \mathrm{Fitness}_{\mathrm{WT}}, \tag{40}$$

工作簿中的分数列直接记录了该量，野生型的基线值为零，因此两个文库分别包含 1705 个（r0）和 1719 个（r349）变体。本节中，位置按从 PAM远端到 PAM近端的顺序编号为 $p = 1$、.....、20，因此 $p = 20$ 紧邻 PAM。这是数据集采用的编号方式，与第 II–V 节中的索引 i 方向相反。错配类型遵循 Feng等人的记法，即 dXrY，其中 X 为靶 DNA 碱基，Y 为 sgRNA 碱基。

富集是各位置因子的乘积。由于失活 Cas9 以完整 R环形成的概率保留变体 i ，且 R环传播具有方向性 [7–9], ，保留因子可分解为二十个位置上的因子，

$$\frac{N_{i,\text{selective}}}{N_{i,\text{control}}} \propto \prod_{p=1}^{20} \eta_p, \qquad \mathrm{score}_i \propto \sum_{p=1}^{20} \log_2 \eta_p, \tag{41}$$

其中 η_p 是位置 p完成其 RNA–DNA 杂交体形成的局部概率。除一个可加的野生型常数外，测量活动因此等于各位置对数形成效率之和，因而可以用自由能总和替代。

每解开一个碱基对，周围双链体的扭转也会改变，变化量为 B 型 DNA 每个碱基对的螺旋扭转$\approx 34.3^\circ$ ，[10, 27], 因此，随着完成杂交体的 位置 增多，累积的扭转应力也会增大。因此，我们认为错配的自由能代价由局部扭转偏差承担，而非仅由碱基对氢键承担，

$$\mathrm{score} \propto \Delta G_{\text{total}} = \sum_{p=1}^{20} \Delta G_p, \tag{42}$$

我们将此视为该模型的核心假设。这也是该模型有别于 Feng 等人 [15],之处：他们将相同的活性分解为最近邻碱基配对热力学，其参数取自 DNA:RNA 文献 [28–30], , 30], 并加入全局 Cas9 补偿项，数值为 -3000 J/mol/碱基。该热力学模型仅适用于 $\approx$ 9.2% 的突变体，因为只有这些突变体的错配类型具有表列参数，因此 Feng 等人 还训练了一个涵盖所有突变体的卷积神经网络。本研究的构建保留了自由能方法的可解释性，同时仅用少量物理参数便涵盖所有突变体。

变体 i 的活性为

$$S_i = b_{\mathrm{rd}} + c \sum_{p=1}^{20} \big[w_{\mathrm{cas},p}\, g_p - \Delta W_p \big], \tag{43}$$

其中， b_{rd}为截距， c $\gamma_p W_{\mathrm{Cas}}^{\mathrm{init}}$ W $= 0.055$ g_p $w_{\mathrm{cas}\,p}$ $=$ 为全局尺度，. 其中 $_{\mathrm{initCas}}$为配对权重， ΔW_p 为扭转代价。配对权重在沃森–克里克匹配的位置处为 1，在错配处等于$[1+\exp(-w_{\mathrm{rd}})]^{-1}$，其中 w_{rd}为取决于错配类型的权重。扭转代价为

$$\Delta W_p = \frac{C\,(\Delta\theta_p)^2}{\Delta x},$$
$$\Delta\theta_p = \Delta\theta_0 + \delta\theta_p^{\mathrm{mis}} + \alpha_1\,\delta\theta_{p-1}^{\mathrm{mis}} + \alpha_2\,\delta\theta_{p-2}^{\mathrm{mis}}, \tag{44}$$

其中 $C = 1.4169 \times 10^{-6}$、 $\Delta x = 1$和 $\Delta\theta_0 = 1.56^\circ$ =为 27。20 毫弧度。能量以 $k_B T$ 表示，角度以毫弧度表示，长度以碱基对表示。可训练参数为 w_{rd} 的 12 个值、 $\delta\theta$ 的 12 个值、 γ_p 的 20 个值，以及 b_{rd}、 c、 α_1、 α_2。

式（44）中的能量是弹性杆扭转能， C 是以（毫弧度）$^{-2}$ 为单位的扭转模量，其数值在模型中固定，而非根据活性数据拟合得到。因此，我们将 C 视为该粗粒化模型中的能量尺度校准，而不是独立测得的 DNA 常数。偏移量$\Delta\theta_0 = 1.56^\circ$ 表示 A 型 RNA–DNA 杂交体与两侧 B 型 DNA 之间的扭转差异：前者每圈含 11.0 个碱基对（每个碱基对为 32.7° ），后者每圈含 10.5 个碱基对（每个碱基对为 34.3° ） [28]。系数 α_1, α_2描述扭转形变如何传播至相邻位置，而非仅局限于局部 [31]。尺度因子$W_{\mathrm{initCas}} = 0.055$用于确定变构项的大小，使拟合得到的每个位置的功$c\,\gamma_p W_{\mathrm{initCas}}$保持在 Feng等人 [15]测得的每碱基对 $-1.2\,k_B T$ 补偿范围内，并与单分子实验测得的 R环形成每碱基对 2–6 $k_B T$相符 [8, 10]。训练流程、数据划分、初始化以及施加于参数的约束详见补充材料 S1 节。

在留出测试集上，模型对 r0 达到 R^2 =0.68 $\pm$ 0.06 和斯皮尔曼相关系数 $= 0.84 \pm 0.03$ ，对 r349 达到$R^2 = 0.64 \pm 0.05$ 和斯皮尔曼相关系数 $= 0.82 \pm 0.03$ ，结果为七个独立随机划分（随机种子 0 至 6）的平均值。对于所使用的代表性分区，见下-

文（随机种子 0），r0 达到 $R^2 = 0.6994$，RMSE 为 $= 1.047$，斯皮尔曼相关系数为 $= 0.8237$，皮尔逊相关系数为 $= 0.8364$；r349 达到$R^2 = 0.6869$，RMSE 为 $= 1.236$，斯皮尔曼相关系数为 $= 0.8339$，皮尔逊相关系数为 $= 0.8325$ （图2和图3）。两种秩相关均超过 Feng 等人 为其热力学模型报告的 Spearman相关系数 0.64 [15]。这一差距部分是物理上的，因为本模型将位置与错配类型解析为各自独立的参数集；部分是方法上的，因为 Feng 等人 采用随机 80/20 划分，而本文采用固定覆盖划分，因此该比较只具指示性，而非受控基准。七个分区上测试指标的变化，以及从拟合中排除最活跃变体的影响，见补充材料第 S2 节和图S1。

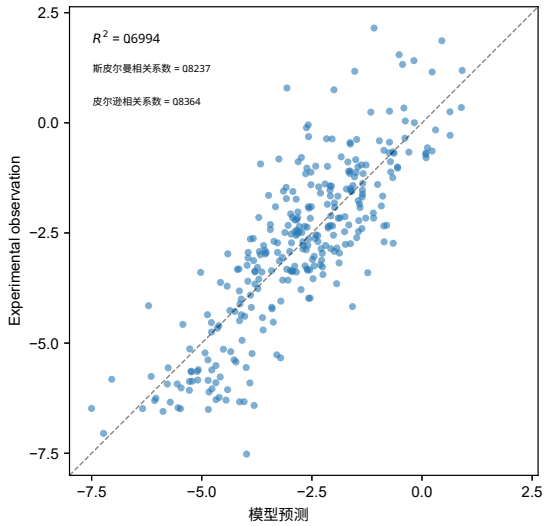


图 2. r0测试集上的预测结合活性与实测结合活性对比（$n = 302$）。虚线为 $y = x$。

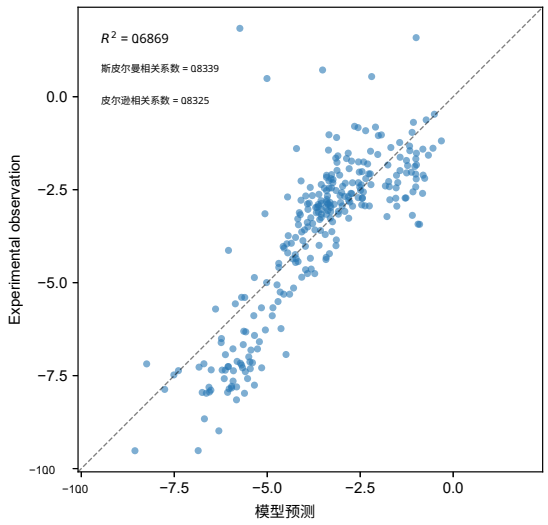


图 3. r349测试集上的预测结合活性与实测结合活性对比（$n = 275$）。虚线为 $y = x$。

残差呈现结构性而非均匀分布，且这种结构具有信息价值。在七个划分中，两个文库的测试残差都带有厚左尾：r0 的超额峰度介于 0.9 和 17.6 之间，r349 的则介于 1.7 和 15.0 之间；被标记的变体几乎全是测量活性较高、但模型低估了其活性的双重错配（第 S3 节及补充材料中的图 S2–S9）。在非正活性子集上使用同一模型进行训练后，残差恢复为类高斯分布；该子集包含 r0 中 94% 和 r349 中 99% 的变体，而正活性的双重错配尾部仍被低估（补充材料中的图 S10–S13）。因此，这个尾部标示出当前自由能映射不适用的区域，原因有两点。首先，两个相近的错配并非彼此独立；而逐位效率相乘（等价于逐位自由能相加）的形式，无法表示这类变体维持其活性的协同作用。系数 α_k 只能吸收部分这种耦合。其次，双重错配造成的部分代价可以在局部得到缓解。Cas9 [5]，中 Arg71–鸟嘌呤接触，加上含 G 的非经典碱基对具有较高稳定性 [32]，，提供了一种结构途径，使涉及 G 的错配所造成的代价低于其表列热力学参数所暗示的程度；这一点也与

方向，如下方针对 dDrG 类拟合得到的小幅扭转偏差所示。

拟合参数再现了实验现象学，并揭示了文库特异性结构。对于代表性分区，系数 γ_p 在位置 1 至 4 处为零，在位置 5 至 14 之间逐渐升高，并在 PAM近端附近达到最大值，其中 r0 的 $\gamma_{15} = 3.79$、$\gamma_{19} = 2.56$和 $\gamma_{20} = 4.69$ 如图 4 所示。由于按照这种编号方式，位置 20 是 PAM近端，因此最大的惩罚位于种子区和 PAM区域，这与具有方向性的 R环相一致：在早期遇到的错配会阻断传播，其代价高于在晚期遇到的错配；这也与其他研究中报告的种子区敏感性和远端耐受性一致 [3, 4, 7, 15, 33]。

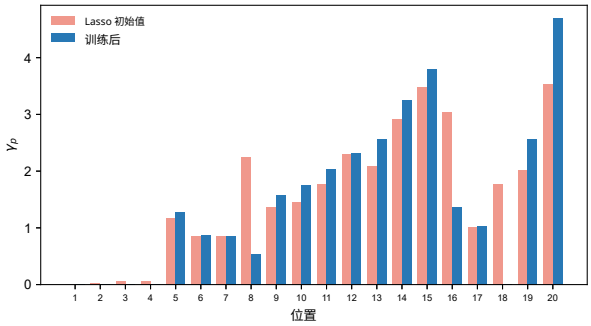


图 4. 初始位置系数和拟合位置系数 γ_p （r0）。

对于代表性分区，dDrG 类的扭转偏差在大多数情况下都很小：r0 中为 dArG=−0.32、dGrG= 3.58 和 dTrG= 3.14 毫弧度，而

r349 中为 dArG= 3.92、dGrG= −0.13 和 dTrG = −0.56 毫弧度，这与 dDrG错配导致的中等程度活性损失相呼应 [15]（图5）。最大的偏差则具有文库特异性：在 r0 中，dCrA= 11.42、dCrC= 10.96 和 dCrU = 10.95 毫弧度占主导；而在 r349 中，dGrA 占主导，达到 10.04 毫弧度，dCr 值则仅为中等水平，分别为 1.38、4.81 和 6.69 毫弧度。因此，dDrG 的扭转偏差普遍较小是一种稳健特征，而其余类型的排序在不同文库间并不一致。在模型中，较小或为负的 $\delta\theta$ 意味着这些错配几乎不会扰动杂交体的扭转，且形变代价较小；负号表示偏差方向与 $\Delta\theta_0$相反，即部分补偿杂交双链的扭转错配，这与 Arg71–鸟嘌呤接触 [5]相符。由于 dDrG 类的这些量级约为 1 毫弧度，接近拟合分辨率，因此我们将符号视为定性结果。

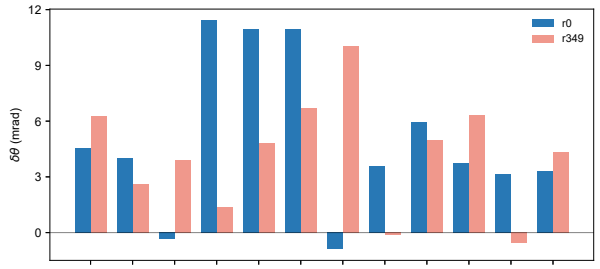


图5. 每种错配类型的拟合扭转偏差 $\delta\theta$ ， r0 和 r349。

最后，扭转传播表现出很强的不对称性：对于 r0，$\alpha_1 = 22.48$ 和 $\alpha_2 = 5.7 \times 10^{-4}$；对于 r0， $c = 19.47$ 和 $b_{\mathrm{rd}} = -31.35$ ；对于 r349， $c = 17.70$和 $b_{\mathrm{rd}} = -33.65$。较大的 α_1 与消失的 α_2 意味着形变基本只传播一个位置，而不会传播两个位置，这为 Feng 等人 [15], 观察到的随机种子协同作用提供了一种机制；而可忽略的 α_2 则表明，在拟合精度范围内，这种耦合的作用距离很短。

综上所述，这些结果表明，无需任何外部热力学数据库，物理约束的扭转模型就能在七个独立分区上以 $R^2 = 0.68 \pm 0.06$ 和 0.64 ± 0.05 重现两个文库，恢复随机种子和 PAM 梯度以及 dDrG错配的中等代价，并将高活性尾部确定为当前自由能映射不足以描述的区域。

VII. 结论

本研究基于主方程与热力学分析相结合的方法，建立了一个非平衡解析框架，用于研究 CRISPR-Cas9系统中的R环形成，并通过熵产生量化该过程的不可逆性。在此基础上我们构建了一个基于实验数据训练的预测模型，该模型不使用外部热力学数据库，并以秩相关系数0.84和0.82再现了两个文库的结合活性。这一比较为该框架提供了初步验证，也使模型参数具有了可解释的物理意义。

值得指出的是，错配耐受性的位置依赖性源于R环传播的顺序结构，其中每一步都取决于前一步是否完成，而最大的惩罚出现在随机种子和PAM区域。两种文库中，dDrG错配的拟合扭转偏差都很小，这与实验中观察到的此类错配导致的中等程度活性损失一致。残差在这样的情况下最大：序列位置相近的双重错配仍能保持其结合活性，因此逐位乘积形式无法充分捕捉邻近错配之间的相互作用。部分残差也可能反映了有限样本量的影响，因为它限制了模型参数的确定精度。

由此可沿两个方向进一步开展研究。首先，本研究假设每个位置对结合概率的贡献彼此独立，而扭转传播系数 α_1 和α_2 仅针对传播至两个最近邻位置的扭转形变修正这一假设。该修正无法捕捉相隔两个以上位置的双重错配，这可能是 r349 文库中残留重尾的来源。采用更一般的耦合结构，例如依赖于位置的相互作用范围或多体项，可以解决这一问题。其次，主方程表述可推广至其他 CRISPR效应复合物。例如，Cas12a 也会从 PAM近端开始，通过顺序解链形成 R环，但其核酸酶激活机制和 PAM 要求与 Cas9 不同。将这一框架应用于 Cas12a，便可直接比较两个系统拟合得到的扭转、堆叠和变构参数，从而揭示 Cas9 与 Cas12a 在错配耐受性上的差异是否源于这些特定物理量的不同。

附录 A：跃迁速率的推导

本文表明，引入累积概率 P_1^i 后，主方程式（3）可化简为式（12），并可由此读出跃迁速率。

由式 (8) 和 (3),

$$\begin{aligned}\dot{P}_1^i &= \sum_{m=i}^{20} \dot{Q}_m \\ &= \sum_{m=i}^{19} \Big[\lambda_{m,m-1}\, Q_{m-1} - (\lambda_{m-1,m} + \lambda_{m+1,m})\, Q_m \\ &\qquad + \lambda_{m,m+1}\, Q_{m+1}\Big] \\ &\qquad + \lambda_{20,19}\, Q_{19} - \lambda_{19,20}\, Q_{20}.\end{aligned} \tag{A1}$$

展开求和,

$$\begin{aligned}\dot{P}_1^i &= \lambda_{i,i-1}\, Q_{i-1} - (\lambda_{i-1,i} + \lambda_{i+1,i})\, Q_i + \lambda_{i,i+1}\, Q_{i+1} \\ &\quad + \lambda_{i+1,i}\, Q_i - (\lambda_{i,i+1} + \lambda_{i+2,i+1})\, Q_{i+1} \\ &\quad + \lambda_{i+1,i+2}\, Q_{i+2} \\ &\quad + \cdots \\ &\quad + \lambda_{20,19}\, Q_{19} - \lambda_{19,20}\, Q_{20}.\end{aligned} \tag{A2}$$

相邻项两两抵消后得到

$$\begin{aligned}&\dot{P}_1^i = 0, \quad (i = 0), \\ &\dot{P}_1^i = \lambda_{i,i-1}\, Q_{i-1} - \lambda_{i-1,i}\, Q_i, \quad (1 \le i \le 20).\end{aligned} \tag{A3}$$

将式 (9) 代入式 (A3)，并将等式两边除以 P_1^{i-1},

$$\frac{\dot{P}_1^i}{P_1^{i-1}} = \lambda_{i,i-1}\, \frac{P_1^{i-1} - P_1^i}{P_1^{i-1}} - \lambda_{i-1,i}\, \frac{P_1^i - P_1^{i+1}}{P_1^{i-1}}. \tag{A4}$$

代入 $q_i = P_1^i / P_1^{i-1}$ 和 $q_{i+1} = P_1^{i+1}/P_1^i$，得到

$$\frac{\dot{P}_1^i}{P_1^{i-1}} = \lambda_{i,i-1}\,(1 - q_i) - \lambda_{i-1,i}\, q_i\,(1 - q_{i+1}). \tag{A5}$$

将系数与式 (12) 比较可得一致性条件，由此我们得到

$$\lambda_{i,i-1} = k^i_{\mathrm{off},1}, \qquad \lambda_{i-1,i} = (1 - \mu_{\mathrm{cleavage}}\delta_{i,20}) k^i_{\mathrm{on},1}. \tag{A6}$$

附录 B：连续R环过程的熵产生

在随机热力学框架下，R环形成的不可逆性通过连续过程 $0 \to 1 \to \cdots \to 20$ 的一条轨迹上的总熵产生来量化,

$$\hat{\sigma} = k_B \ln \frac{P_+[\Gamma^+]}{P_-[\Gamma^-]}, \tag{B1}$$

其中，P_+ 和 P_- 分别是正向轨迹 Γ^+ 及其时间反演轨迹 Γ^- 的概率密度,

分别。

考虑一条从状态 0 出发、依次访问各状态，并因切割而在状态 20 终止的正向轨迹。令 $\tau_i (i = 0,$ 1、.....、19) 表示在状态 i 中、跃迁 $i \to i{+}1$ 发生前的等待时间。在状态 i 中的停留时间贡献指数因子 $e^{-r_i \tau_i}$，其中r_i 是状态 i 的总离开速率，而跃迁 $i \to i{+}1$ 贡献速率$\lambda_{i+1,\, i}$。因此

$$P_+(\Gamma^+) = \left(\prod_{i=1}^{20} \lambda_{i,i-1}\right) \exp\left(-\sum_{i=0}^{19} r_i \tau_i\right), \tag{B2}$$

其中离开速率为

$$r_0 = \lambda_{1,0}, \qquad r_i = \lambda_{i-1,i} + \lambda_{i+1,i} \quad (i = 1, 2, \ldots, 19). \tag{B3}$$

每个状态 $i \ge 1$ 都有两个逃逸通道：正向跃迁 $i \to i{+}1$ 和反向跃迁 $i \to i{-}1$，两者均计入式 (B3)。

在时间反演轨迹上，各跳变以反向速率发生$\lambda_{i-1,\, i}$，而时间反演保持驻留时间序列和离开速率不变。因此，反向概率密度中的等待时间因子与式（B2）中的完全相同。令反向轨迹从状态20及其占据概率Q_{20}开始,

$$P_-(\Gamma^-) = Q_{20} \left(\prod_{i=1}^{20} \lambda_{i-1,i}\right) \exp\left(-\sum_{i=0}^{19} r_i \tau_i\right). \tag{B4}$$

将式（B2）和式（B4）代入式（B1）后，等待时间因子完全抵消，熵产生分解为系统项和介质项,

$$\hat{\sigma} = k_B \ln \frac{Q_0(0)}{Q_{20}(T)} + k_B \sum_{i=1}^{20} \ln \frac{\lambda_{i,i-1}}{\lambda_{i-1,i}}. \tag{B5}$$

此处，$Q_0(0)$表示初始时刻状态0的占据概率，$Q_{20}(T)$表示终止时刻 T状态20的占据概率。

对于具有催化活性的Cas9复合物，其切割效率很高，因此已完成的R环会通过切割而非解离被清除；对于本文所考虑的错配靶标，R环一旦形成，就会受到强烈的正向驱动。经过从 $Q_0(0)$开始的很长时间 T后，系统因此处于 $Q_{20}(T)$，系统项 $k_B \ln[Q_0(0)/Q_{20}(T)]$ 与介质项相比可以忽略。

代入适用于 $1 \le i \le 20$ 的方程和（17）中的速率对应关系，可得 (16)

$$\ln \frac{\lambda_{i,i-1}}{\lambda_{i-1,i}} = \ln \frac{k^i_{\mathrm{off},1}}{(1 - \mu_{\mathrm{cleavage}}\, \delta_{i,20})\, k^i_{\mathrm{on},1}}, \tag{B6}$$

对 i 求和可得

$$\hat{\sigma} = k_B \sum_{i=1}^{20} \ln \frac{k^i_{\mathrm{off},1}}{(1 - \mu_{\mathrm{cleavage}}\, \delta_{i,20})\, k^i_{\mathrm{on},1}}. \tag{B7}$$

由于克罗内克符号选出 $i = 20$，求和分为十九个可逆跃迁和终

止跃迁，切割在此发生，

$$\hat{\sigma} = k_B \sum_{i=1}^{20} \ln \frac{k^i_{\text{off},1}}{k^i_{\text{on},1}} - k_B \ln(1 - \mu_{\text{cleavage}}). \qquad \text{(B8)}$$

致谢

本研究得到中国国家自然科学基金（NSFC）资助，项目编号为 12575045。

数据可用性

支持本文研究结果的数据未公开提供。如有合理请求，可向作者获取这些数据。

[1]M. Jinek、K. Chylinski、I. Fonfara、M. Hauer、J. A. Doudna 和 E. Charpentier，《适应性细菌免疫中的可编程双 RNA 引导 DNA 内切核酸酶》，《科学》**337**，816（2012）。[2] J. A. Doudna 和 E. Charpentier，《Cas9 及其同类的新生物学》，《科学》**346**，1258096（2014）。[3] Y. Fu、J. A. Foden、C. Khayter、M. L. Maeder、D. Reyon、J. K. Joung 和 J. D. Sander，《CRISPR-Cas 核酸酶在人类细胞中诱导的高频脱靶诱变》，《自然-生物技术》**31**，822（2013）。[4] P. D. Hsu、D. A. Scott、J. A. Weinstein、F. A. Ran、S. Konermann、V. Agarwala、Y. Li、E. J. Fine、X. Wu、O. Shalem、T. J. Cradick、L. A. Marraffini、G. Bao 和 F. Zhang，《RNA 引导的 Cas9 核酸酶的 DNA 靶向特异性》，《自然-生物技术》**31**，827（2013）。[5] C. Anders、O. Niewoehner、A. Duerst 和 M. Jinek，《Cas9 内切核酸酶依赖 PAM 识别靶DNA 的结构基础》，《自然》**513**，569（2014）。[6] H. Nishimasu、F. A. Ran、P. D. Hsu、S. Konermann、S. I. Shehata、N. Dohmae、R. Ishitani、F. Zhang 和 O. Nureki，《Cas9 与引导 RNA 和靶DNA 复合物的晶体结构》，《细胞》**156**，935（2014）。[7] S. H. Sternberg、S. Redding、M. Jinek、E. C. Greene 和 J. A. Doudna，《CRISPR RNA 引导的内切核酸酶 Cas9 对 DNA 的探查》，《自然》**507**，62（2014）。[8] M. D. Szczelkun、M. S. Tikhomirova、T. Sinkunas、G. Gasiunas、T. Karvelis、P. Pschera、V. Siksnys 和 R. Seidel，《直接观察单个 RNA 引导的 Cas9 和 Cascade 效应复合物形成 R环》，《美国国家科学院院刊》**111**，9798（2014）。[9] S. Gong、H. H. Yu、K. A. Johnson 和 D. W. Taylor，《DNA 解链是 CRISPR-Cas9 活性的首要决定因素》，《细胞报告》**22**，359（2018）。[10] I. E. Ivanov、A. V. Wright、J. C. Cofsky、K. D. Palacio Aris、J. A. Doudna 和 Z. Bryant，《Cas9 以离散步骤探查 DNA，且该过程受错配和超螺旋调节》，《美国国家科学院院刊》**117**，5853（2020）。[11] E. Semenova、M. M. Jore、K. A. Datsenko、A. Semenova, E. R. Westra, B. Wanner, J. van der Oost, S. J. J. Brouns, and K. Severinov，《美国国家科学院院刊》**108**，10098（2011），干扰由成簇规律间隔短回文重复序列（CRISPR）RNA介导，并受种子序列控制。[12] E. A. Boyle, J. O. L. Andreasson, L. M. Chircus, S. H. Sternberg, M. J. Wu, C. K. Guegler, J. A. Doudna, and W. J. Greenleaf，《美国国家科学院院刊》**114**，5461（2017），高通量生化分析揭示了失活 Cas9脱靶结合与解离的序列决定因素。[13] M. Klein, B. Eslami-Mossallam, D. G. Arroyo, and M. Depken，《细胞报告》**22**，1413（2018），杂交动力学解释了CRISPR-Cas的脱靶规律。[14] B. Eslami-Mossallam, M. Klein, C. V. D. Smagt, K. V. D. Sanden, S. K. Jones, J. A. Hawkins, I. J. Finkelstein, and M. Depken，《自然·通讯》**13**，1367（2022），动力学模型可预测SpCas9活性、改进脱靶分类，并揭示靶向保真度的物理基础。[15] H. Feng, J. Guo, T. Wang, C. Zhang, and X.-H. Xing，《核酸研究》**49**，1263（2021），引导序列-靶标错配对活细菌细胞中失活Cas9–sgRNA结合活性的影响。[16] F. Alkan, A. Wenzel, C. Anthon, J. H. Havgaard, and J. Gorodkin，《基因组生物学》**19**，177（2018），利用核酸双链能量参数评估CRISPR-Cas9脱靶效应。[17] J. S. Chen, Y. S. Dagdas, B. P. Kleinstiver, M. M. Welch, A. A. Sousa, L. B. Harrington, S. H. Sternberg, J. K. Joung, A. Yildiz, and J. A. Doudna，《自然》**550**，407（2017），增强的校对机制决定了CRISPR-Cas9靶向准确性。[18] J. J. Hopfield，《美国国家科学院院刊》**71**，4135（1974），动力学校对：一种降低需要高特异性生物合成过程中错误的新机制。[19] P. Sartori and S. Pigolotti，《物理评论 X》**5**，041039（2015），错误校正的热力学。[20] U. Seifert，《物理学进展报告》**75**，126001（2012），随机热力学、涨落定理与分子机器。[21] P. Gaspard，《化学物理学杂志》**120**，8898（2004），非平衡反应的涨落定理。[22] R. Rao and M. Esposito，《物理评论 X》**6**，041064（2016），化学反应网络的非平衡热力学：随机热力学带来的启示。[23] H. Eyring，《化学物理学杂志》**3**，107（1935），化学反应中的活化络合物。[24] G. I. Bell，《科学》**200**，618（1978），细胞间特异性黏附模型。[25] J. G. Doench, N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, K. F. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, H. W. Virgin, X. Liang, D. K. Bhatt, D. E. Root, and W. C. Hahn，《自然-生物技术》**34**，184（2016），优化sgRNA 设计，以最大限度提高CRISPR-Cas9活性并尽量减少脱靶效应。[26] M. A. Moreno-Mateos, C. E. Vejnar, J.-D. Beaudoin, J. P. Fernandez, E. K. Mis, M. K. Khokha, and A. J. Gi-raldez，《自然·方法》（2015），CRISPRscan：设计高效sgRNA以在体内靶向CRISPR-Cas9**12**，982。

[27]J. D. Watson and F. H. C. Crick, 《自然》**171**，737（1953），核酸的分子结构：脱氧核糖核酸的结构。[28] N. Sugimoto, S.-i. Nakano, M. Katoh, A. Matsumura, H. Nakamuta, T. Ohmichi, M. Yoneyama, and M. Sasaki, 《生物化学》**34**，11211（1995），用于预测RNA/DNA杂交双链稳定性的热力学参数。[29] N. E. Watkins, W. J. Kennelly, M.-J. Tsay, A. Tuin, L. Swenson, H.-R. Lee, S. Morosyuk, K.-M. Weng, and J. SantaLucia, 《核酸研究》**39**，1894（2011），RNA/DNA双链中单个内部rA·dA、rC·dC、rG·dG和rU·dT错配的热力学贡献。[30] E. A. Lesnik and S. M. Freier, 《生物化学》**34**，10807（1995），DNA、RNA和DNA:RNA杂交双链的相对热力学稳定性。[31] P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii, 《核酸研究》**34**，564（2006），碱基堆积和碱基配对对DNA双螺旋热稳定性的贡献。[32] H. T. Allawi and J. SantaLucia, 《生物化学》**36**，10581（1997），DNA中内部G·T错配的热力学与核磁共振研究。[33] V. Pattanayak, S. Lin, J. P. Guilinger, E. Ma, J. A. Doudna, and D. R. Liu, 《自然-生物技术》**31**，839（2013），脱靶DNA切割的高通量分析揭示了RNA编程Cas9核酸酶的特异性。

《CRISPR-Cas9系统中R环形成的非平衡动力学》补充材料

S1. 数据划分、优化与参数约束

两个文库分别独立拟合。两个文库使用相同的模型、优化器和训练轮数预算。下文所讨论的变体之间，只有训练子集不同。划分并非简单的随机划分：每个分区首先分配至少一个包含每种错配类型和每个位置的变体到训练集，以确保训练集覆盖整个参数集；随后将训练集扩充至文库样本数的 60% 至 70%（均匀抽取），剩余变体则平均分配到留出集和测试集。使用 Adam 优化参数，学习率为 10^{-3}，权重衰减为 10^{-4}，梯度范数裁剪值为 1；以训练集均方误差 $\mathcal{L}=\mathrm{MSE}_{\mathrm{train}}$ 为目标，训练 90000 轮。留出子集不参与目标函数计算，但用于驱动 ReduceLROnPlateau 调度器；测试集仅在训练完成后评估一次，因此这两个集合都不参与梯度计算。结构约束，即 $c>0$、$0<\alpha_2<\alpha_1$ 以及非负的位置系数，通过软惩罚施加，并在每一步之后进行硬投影。初始化遵循模型的物理机制，其中 γ_p 取自训练子集的 Lasso 系数绝对值，$\delta\theta$ 取自 1.56°、$\alpha_1=\alpha_2=0.1$、$c=1.0$和 $b_{\mathrm{rd}}=-4.0$。正文中使用的代表性分区里，r0 的训练集、留出集和测试集分别包含 1101、302 和 302 个样本；r349 则分别包含 1169、275 和 275 个样本。

S2. 划分间变异及极端变体的排除

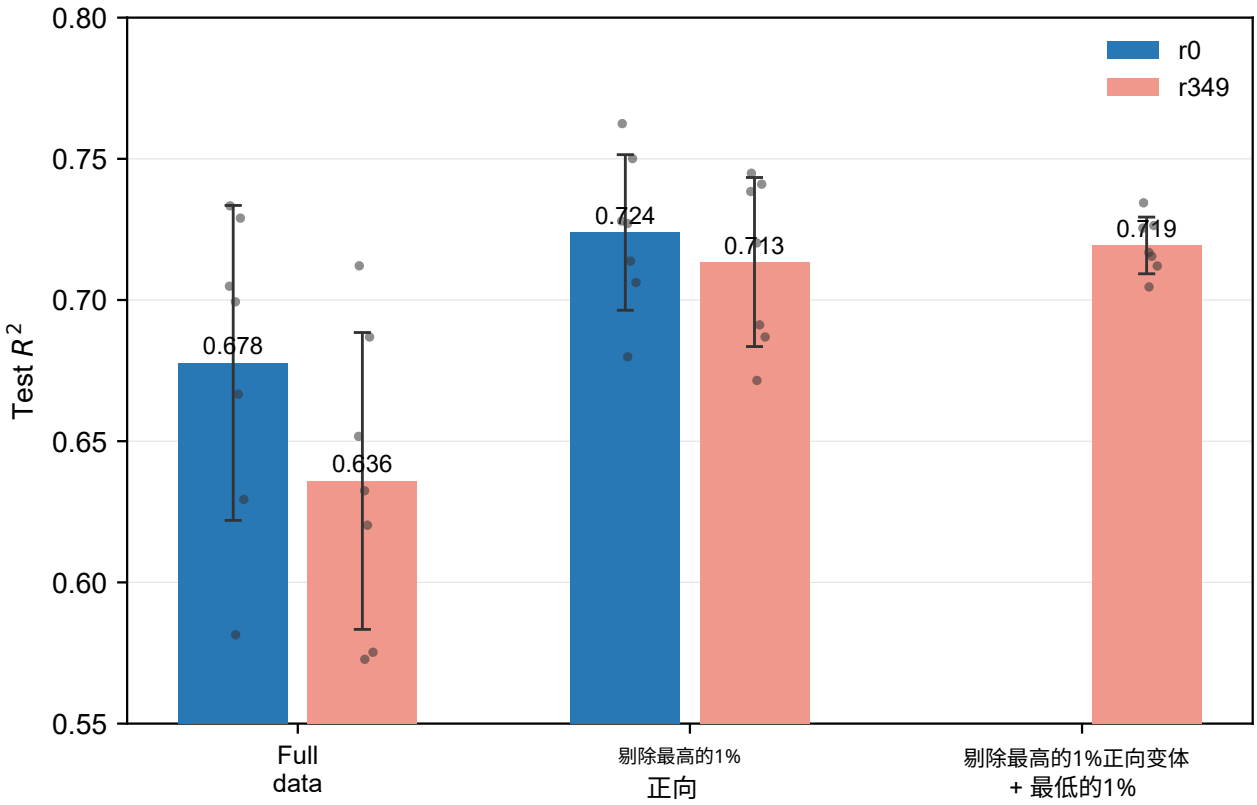
在留出测试集上，对七个独立随机划分（随机种子为 0 至 6）取平均后，模型在 r0 上达到 $R^2=0.68\pm0.06$，斯皮尔曼相关系数为 $=0.84\pm0.03$；在 r349 上达到 $R^2=0.64\pm0.05$，斯皮尔曼相关系数为 $=0.82\pm0.03$。秩相关不受划分选择的影响。图S1比较了三种训练变体：全数据拟合、移除活性最高的前 1% 正向变体后的拟合，以及仅针对 r349，移除正向变体中活性最高的前 1% 和活性分布中活性最低的后 1% 后的拟合。对于 r0，第一种变体已足够（第S3节），因此未运行其他变体。

移除同时为正向且活性位于前 1% 的 18个变体（约占每个文库的 1%）后，r0 留下 1687 个样本，r349 留下 1701 个样本。使用相同的七个分区重新拟合相同的模型后，两个文库的测试集性能均有所提升：r0 达到 $R^2=0.724\pm0.028$，r349 达到 0.713 ± 0.030；完整数据拟合的对应结果分别为 0.678 ± 0.056 和 0.636 ± 0.053，且分区间离散度大约减半。

S3. 残差诊断

生物测量包含噪声，因此我们检验拟合是否已吸收数据中的确定性部分。我们根据残差 $r_i=\hat{}y_i-y_i$ 的前四阶矩来判断噪声是否呈类高斯分布，因为拟合应满足的要求是：残差没有系统性偏移、没有偏度，也没有超额峰度。我们用 Jarque–Bera 统计量概括后两项；为提高透明度，还报告两种更严格的整体形状检验：参数自助法 Kolmogorov–Smirnov 检验（进行 9999 次重抽样，每次均重新估计均值和标准差）以及 Shapiro–Wilk 检验。这两种检验更容易拒绝原假设，因为它们还会考察与偏度和峰度无关的形状特征。在七个分区中，两个文库的测试残差都呈现厚左尾，其程度随分区而变化：r0 的超额峰度范围为 0.9 至 17.6（中位数为 6.4），r349 的范围为 1.7 至 15.0（中位数为 6.7），因此 Jarque–Bera 检验在每个分区中都拒绝高斯性（图 S2 和 S3）。正文中采用的代表性分区——随机种子 0——是 r0 中情况最温和的，因此并不具有代表性。保留该分区仅用于参数讨论。两个测试集各自按分区绘制的 Q-Q 图见图 S4 和 S5。

去除每个测试集中的极端残差（占比为几个百分点）后，每个分区中的 r0 以及七个分区中的两个分区的 r349 均恢复了高斯性。被去除的样本几乎全部是双重错配：汇总七个分区后，r349 离群值的 100% 和 r0 离群值的 93% 均携带两个错配，而单错配仅在一个变体中被标记，且只出现在七个分区中的四个分区。被标记的变体也位于活性分布的两端：汇总各分区后，r0 离群值的 88% 和 r349 离群值的 90% 落在其测试集活性的极端五分位组中，并且主要集中在高活性端（分别为 63% 和 74%）。因此，在双重错配类别中，偏差集中出现在两个错配位于序列中彼此接近的位置，且测量活动为正的情况下。对于 r349，所有正活性变体中有 95% 被标记为离群值，其残差系统性为负，也就是说，模型低估了这些变体。将残差与测量活动作图后，这一点表现得十分直接



图S1. 三种训练变体的检验 R^2，从左到右依次为：完整数据、剔除活性最高的1%正向变体后的拟合，以及同时剔除活性分布中正向变体最高的1%和最低的1%后的拟合。柱形表示七次随机划分的均值，误差线表示标准差。圆点表示各个分区。第三种变体仅针对 r349 运行，因为对于 r0，剔除最高的1%后已恢复类高斯残差（图S6），因此无需进一步测试其他变体。全程使用相同的模型和优化器，仅训练子集有所变化。

（图 S8 和 S9）：对于评分 ≤ 0，其以零为中心；而对于评分 > 0，其显著转为负值，被标记的点位于正活性区域。

这种偏离有三个来源，应加以区分；现有数据只能部分将它们区分开来。首先，样本具有对抗性：两个相邻错配并非相互独立，而将逐位效率相乘（等价于将逐位自由能相加）的模型，无法体现使这类变体得以保留活性的协同作用，系数 α_k 只能吸收其中一部分。其次，局限可能在于模型，而非数据，因为结合是一个多步可逆过程，其结果未必是平衡自由能 [7, 8]的单调函数。错配也可能缓解局部应变；Cas9 [5] 中的 Arg71–鸟嘌呤接触，或许会减轻 dDrG错配带来的代价，同时 G 所含非经典碱基对的稳定性也很高 [32]。因此，高活性尾部恰恰是自由能映射不完整之处。第三，实验噪声也有影响，因为低丰度变体的计数波动以及非线性生长响应会抬高个体活性 [15]，；而这种抬高对罕见活性变体而言是向上的，方向上正好会产生观察到的左尾。均值接近零本身并不能证明噪声是高斯白噪声，因为残差分布混合了模型偏差、测量波动以及可能存在的异方差性；因此，没有重复实验就无法定量区分这三种贡献。

为检验是否由少数极端变体所致，我们将它们剔除后重新拟合。我们剔除了同时为正且活性位于前 1% 的 18个变体，r0 剩余 1687 个样本，r349 剩余 1701 个样本，并在相同的七个分区上重新拟合相同的模型。两个文库的测试集性能均有所提升，分区间离散度也大致减半（图 S6 和 S7）。随后，重新拟合模型的残差在 r0 上呈类高斯分布（测试样本的超额峰度为 0.89，评分为 ≤ 0），但在 r349 上仍呈重尾分布（超额峰度为 3.51）；残差-活性图说明了原因（图 S8 和 S9）：r0 中的标记点已消失，而 r349 中的低估型双重错配变体仍分布在整个活性范围内。因此，重尾只能部分归因于最极端的变体。对于 r349，重尾也反映了双错配耦合

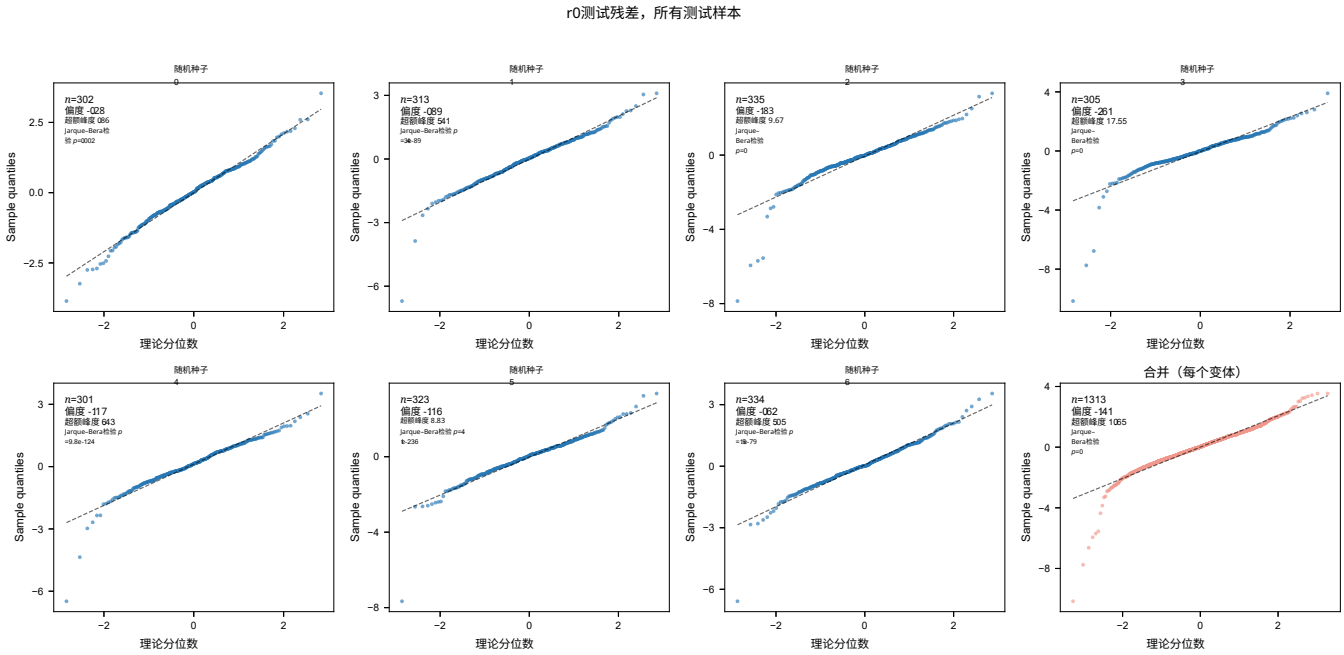


图 S2. 移除任何样本之前，r0测试集残差的 Q-Q 图。面板 1–7 分别对应七个随机划分（随机种子 0–6）。面板 8 将这些划分合并，为每个变体指定一个残差，即该变体落入测试集的各个划分中的残差均值，因此在多个测试集中出现的变体仅计数一次。每个划分都存在厚左尾，且随机种子 2–4 对应的左尾最明显。每条注释均给出样本量 n、偏度、超额峰度和 Jarque–Bera p。

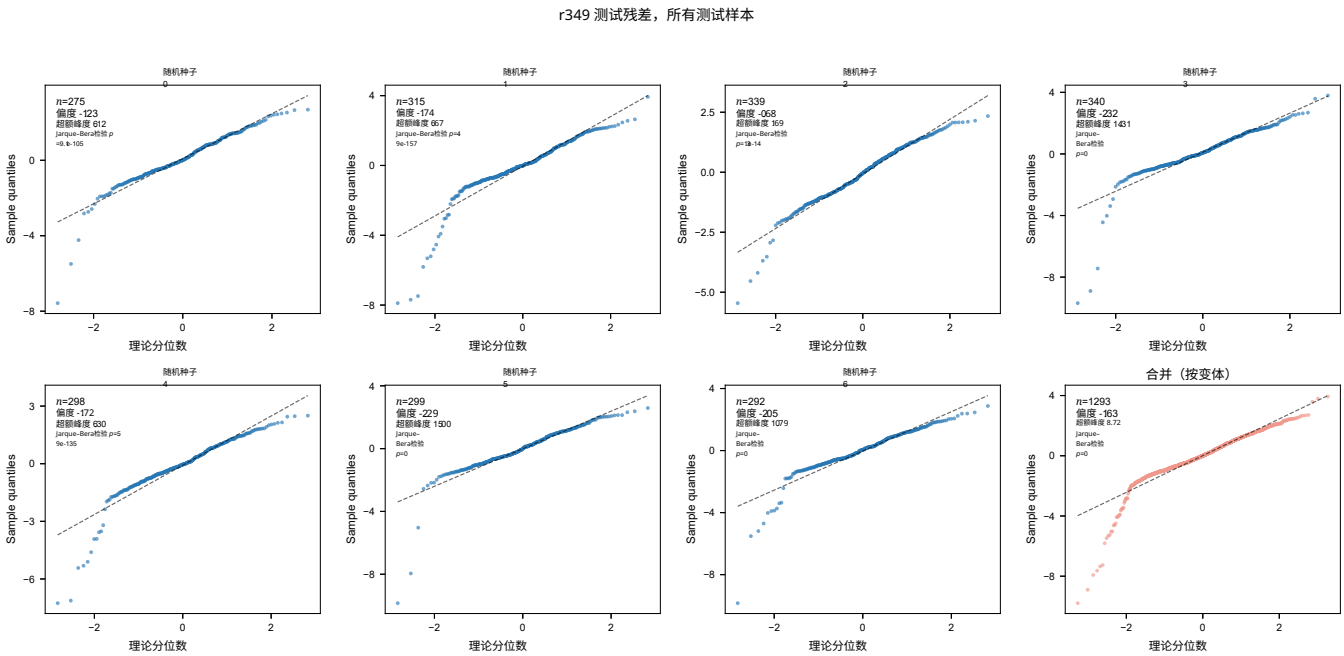


图 S3. 与图 S2 相同，但展示的是 r349测试集残差。两个文库偏离高斯性的程度相当，因此从各个划分的整体情况来看，没有任何一个文库特别突出。

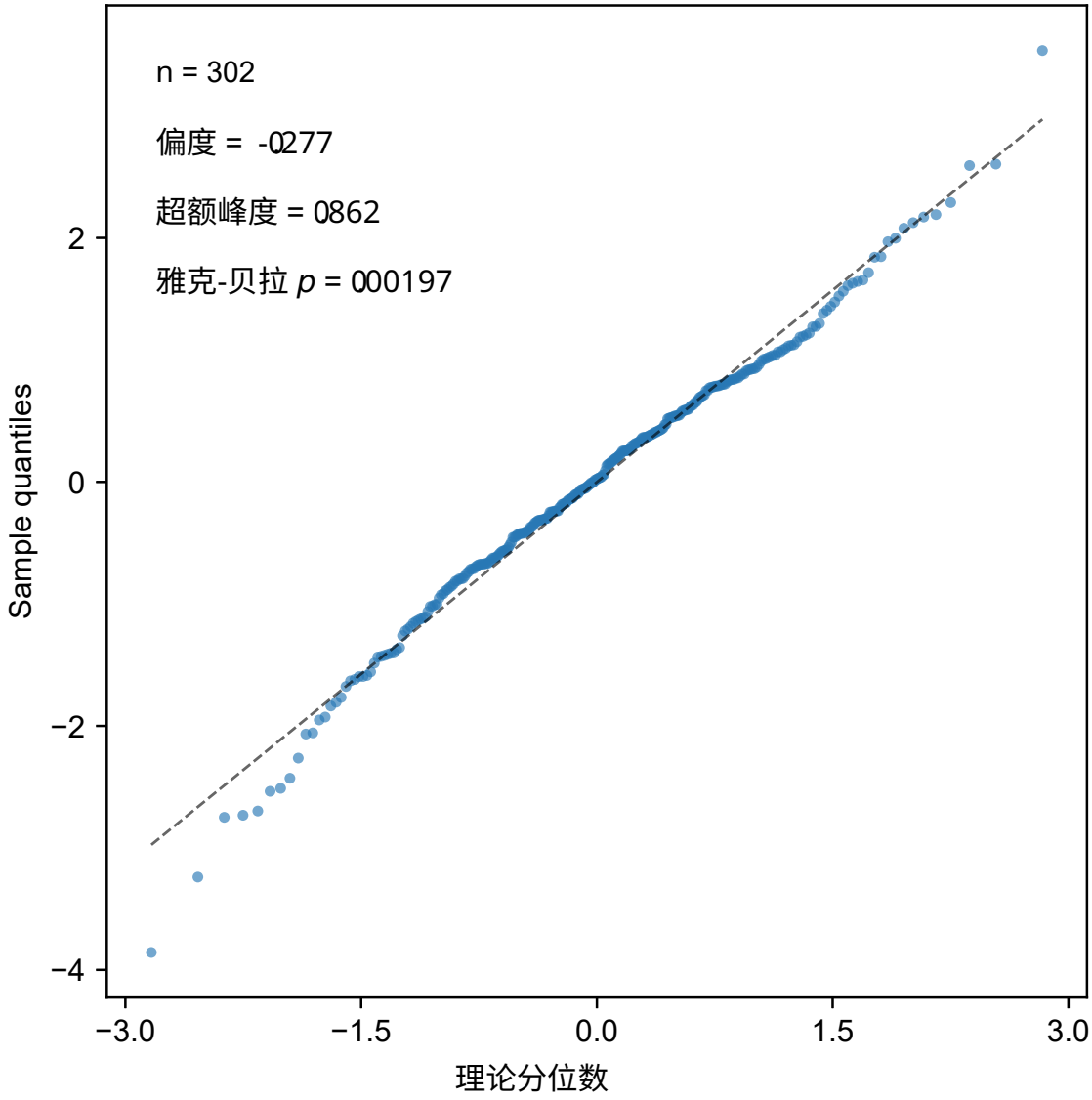


图 S4. r0测试集残差的 Q-Q图（$n = 302$）。

乘积形式无法吸收这一点。

另外进行两次重新拟合以确定其来源。首先，移除所有正活性变体，而不只是排名前 1% 的变体，并在评分子集（7 个分区，$R^2 = 0.729 \pm 0.040$）上重新拟合 r349，≤ 0结果显示重尾依然存在：七个分区中只有两个通过 Jarque–Bera检验（超额峰度为 0.39 和 -0.41），其余五个介于 0.99 至 9.86 之间。被标记的变体仍然全是双重错配，且如今仅限于负活性范围，其残差按符号分成两类：评分为 < -3 的变体被高估（均值为 $+2.8$ 至 $+4.0$），而评分为 $-3 <$ < 0 的变体则被低估（均值为 -3.6 至 -4.5）。这正是回归稀释的特征：乘积形式将双重错配的活性从两端向中心压缩，因此仅移除上尾无法消除重尾。

其次，进一步移除活性分布中最低的 1%，使剩下的仅为中心负活性变体和弱正活性变体，并再次拟合 r349（7 个分区，$R^2 = 0.719 \pm 0.010$，三种变体中可重复性最高的一种），结果仍有七个分区中的六个呈非高斯分布（超额峰度为 0.75 至 3.66，合并后的数值从 3.01 降至 1.60）。因此，去除两端极值后，重尾仍然存在。这表明问题出在模型，而非少数样本：强活性和强抑制变体的测量活动无法由逐位乘积重现，不过现有数据无法排除这类变体的实验误差较大。相比之下，r0 只需移除最高的 1% 就会变得类高斯：其正活性变体中包括模型能够处理的单错配，只有极端变体会被错误预测；而 r349 的所有正活性变体都是双错配，且其活性均被严重低估。

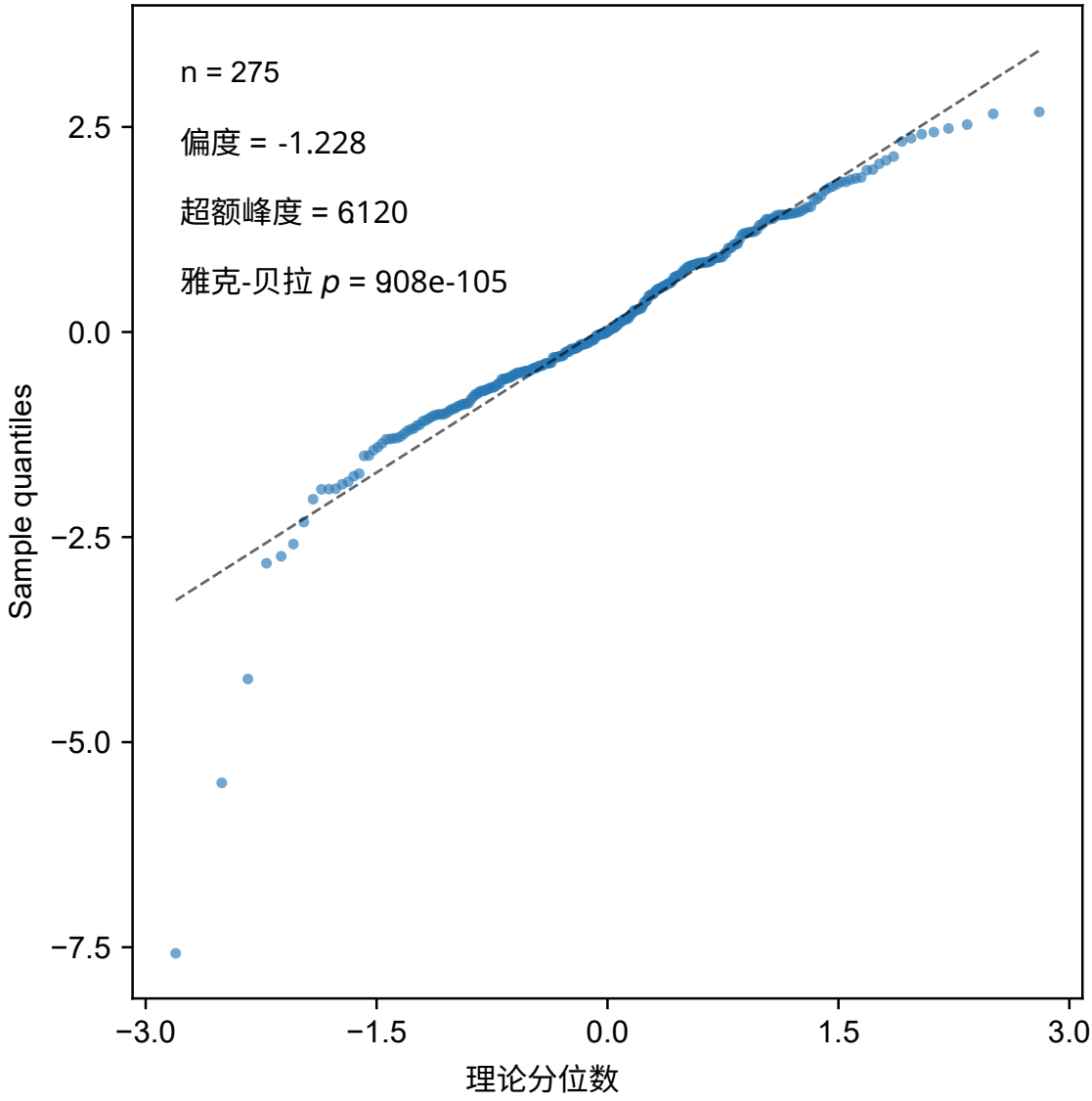


图 S5. r349测试集残差的 Q-Q图（$n = 275$）。

因此，我们通过单独拟合负活性主体，直接检验模型局限性解读。在评分为 ≤ 0 的子集上重新训练相同模型（r0 有 1597 个样本，r349 有 1696 个样本，优化器和训练轮数预算相同），尾部随之消失：在相应测试集上，r0 达到 $R^2 = 0.727$，偏度为 0.063，超额峰度为 0.023，JB $p = 0.923$，KS $p = 0.800$，夏皮罗 $p = 0.840$（$n = 248$）；r349 达到 $R^2 = 0.747$，偏度为 0.002，超额峰度为 0.387，JB $p = 0.432$，KS $p = 0.011$，夏皮罗 $p = 0.009$（$n = 309$）（图 S10、S11、S12 和 S13）。此时，两个文库均符合矩准则，但更严格的整体形状检验仍指出 r349 存在轻微的形状偏差。类高斯噪声因此能够描述负活性主体，其中包含 r0 的 94% 和 r349 的 99% 变体；而正活性双错配尾部则是当前自由能映射尚不完整的区域。

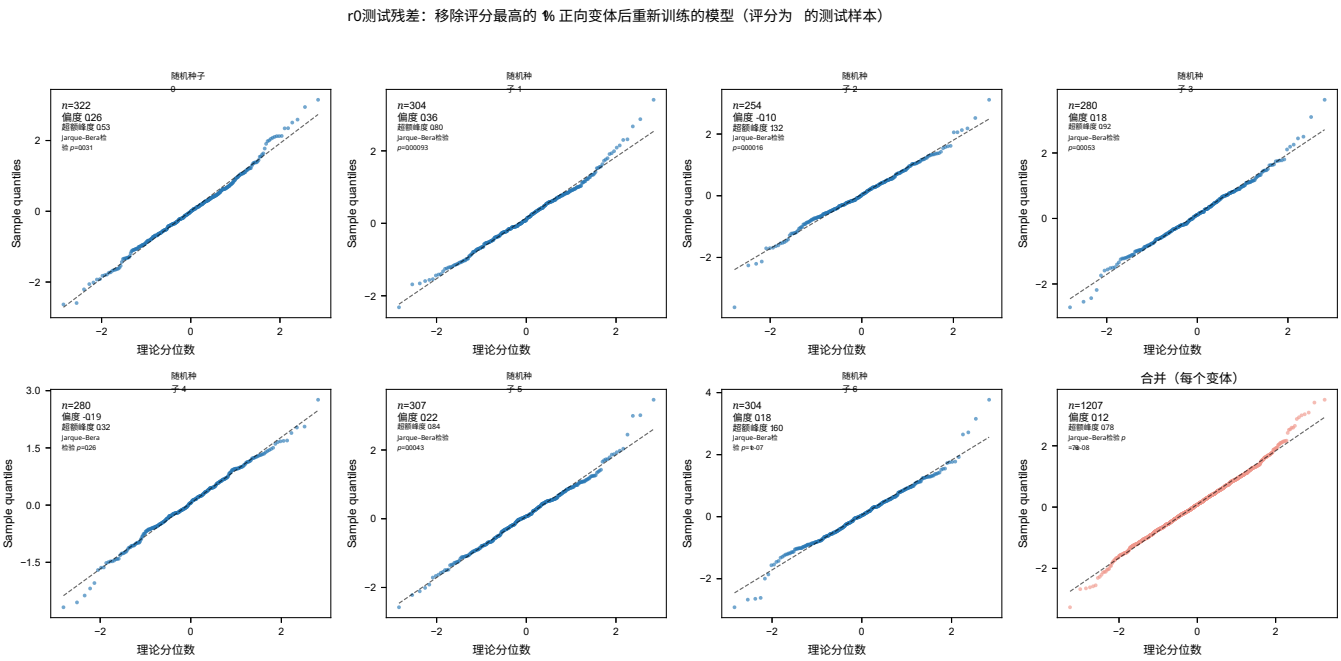
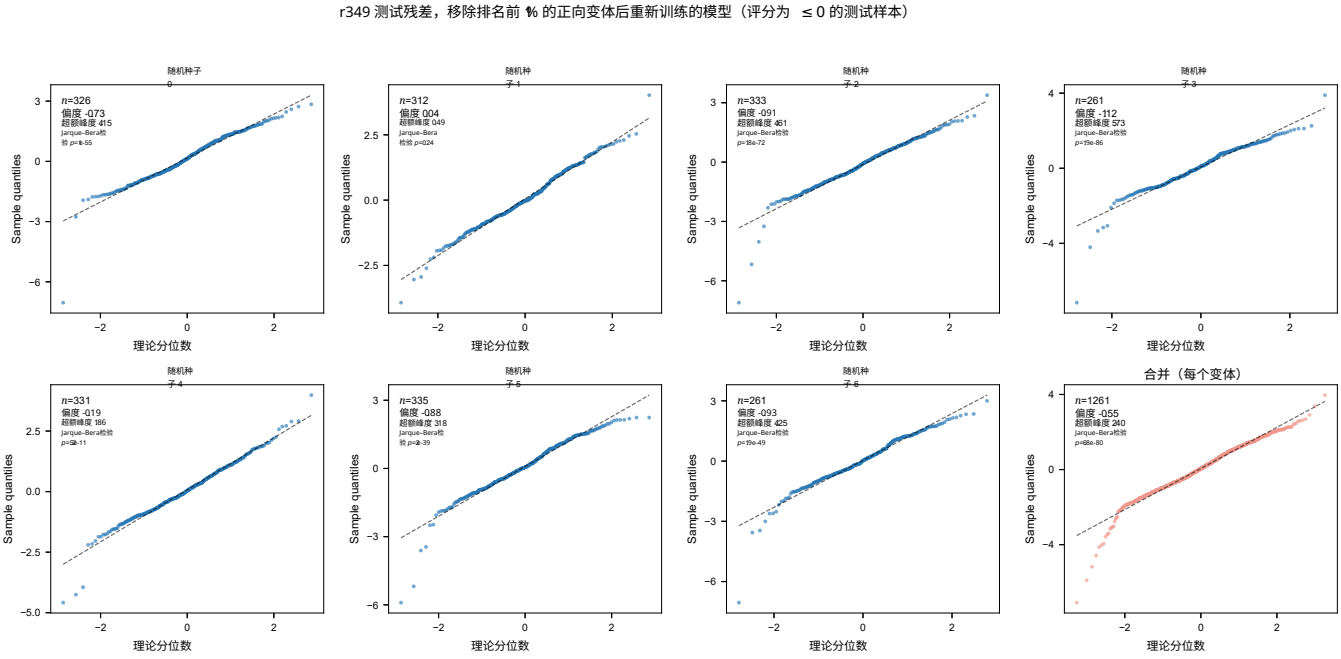


图 S6. 模型 *retrained* 移除同时为正向且活性排名前 1% 的 18个变体后，r0测试残差的 Q-Q 图（每次拟合使用 1687 个样本，而非 1705 个）。布局同图 S2，仅显示评分为 ≤ 0 的测试样本，即具有重尾的类别。移除这些变体后，每个分区的超额峰度均低于 1.7，而测试 R^2 上升（完整数据的均值为 0.678 ± 0.056，相比之下为 0.724 ± 0.028）。



图S7. 与图 S6 相同，但针对 r349（移除相同的 18 个变体，每次拟合使用 1701 个样本），同样仅限于评分为 ≤ 0的测试样本。此处，重新训练使测试 R^2 从 0.636 ± 0.053 升至 0.713 ± 0.030，但超额峰度平均仍为 3.5，且七个分区中有六个仍存在厚左尾，因此，对于这个文库而言，这种偏离并非仅由那几个变体造成。

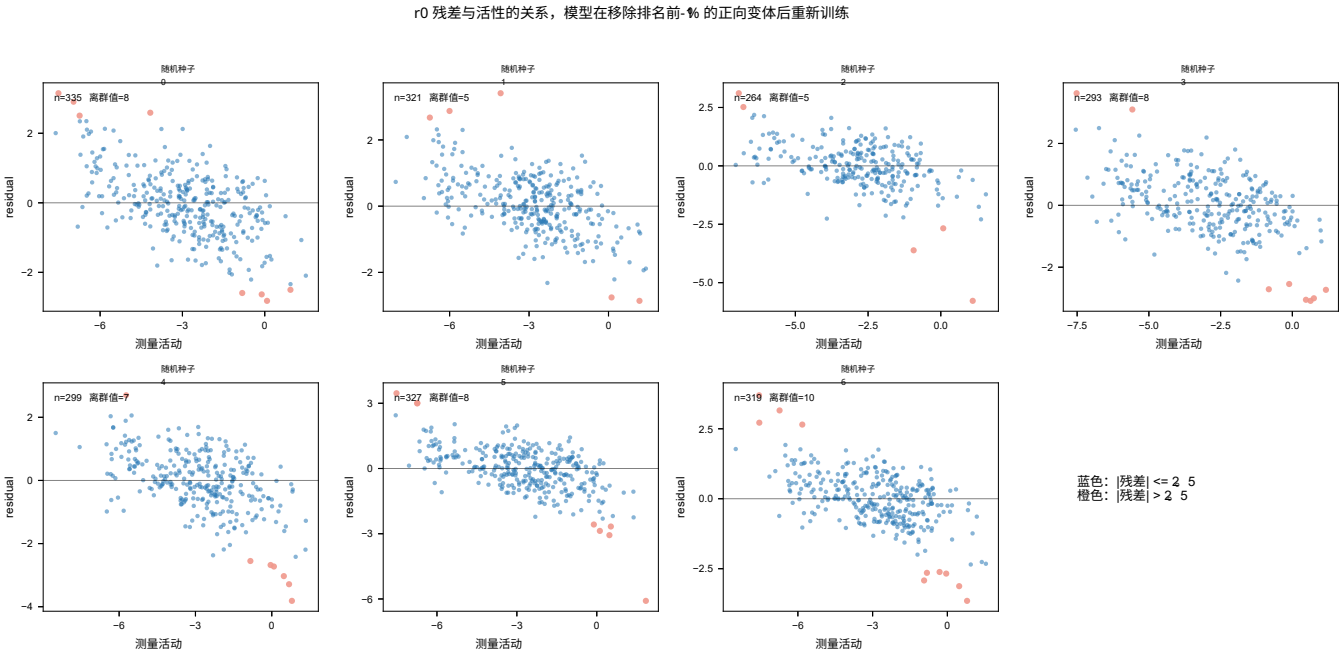


图 S8。重新拟合的模型中，每个分区一个面板，显示 r0 残差与测量活动的关系。突出显示了具有 $|r| > 2.5$ 的样本。对于非正活性，残差云以零为中心；对于正活性，残差云转为负值，而突出显示的样本集中在此处。

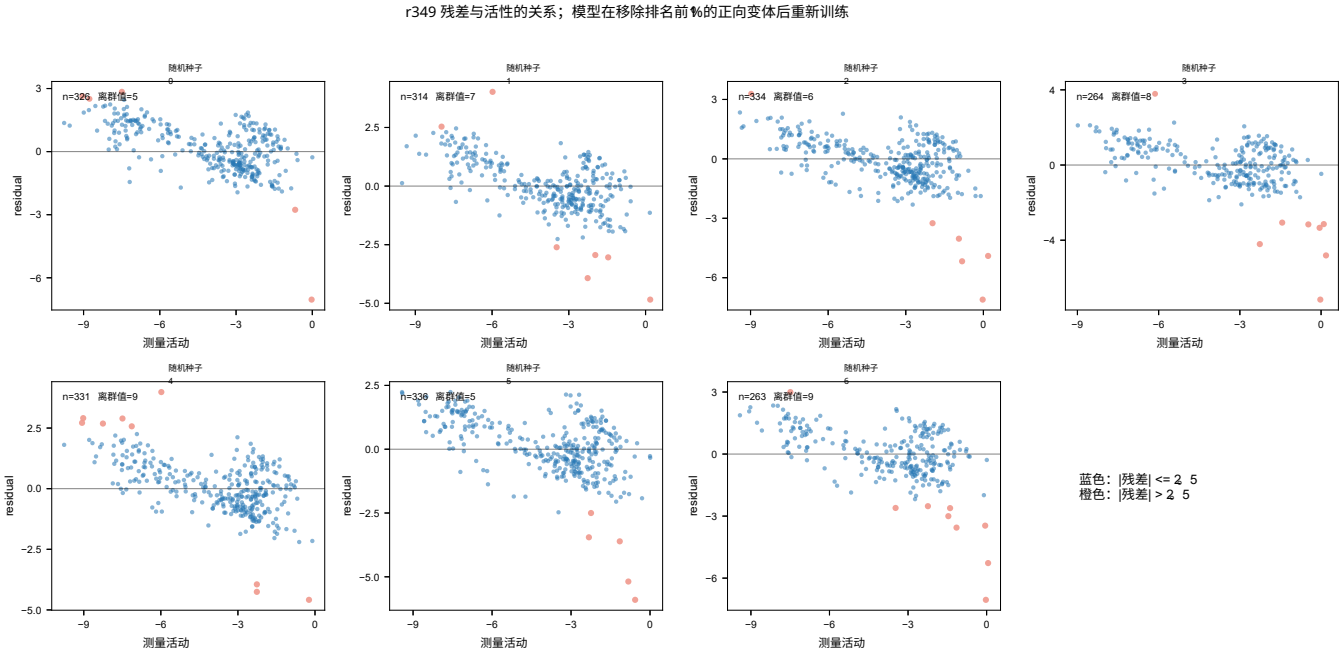


图 S9。与图 S8 相同，但针对 r349。尽管不再拟合排名前 1% 的正向变体，但每个分区中仍存在低估型双重错配变体，这也是 r349 残差保持重尾分布的原因。

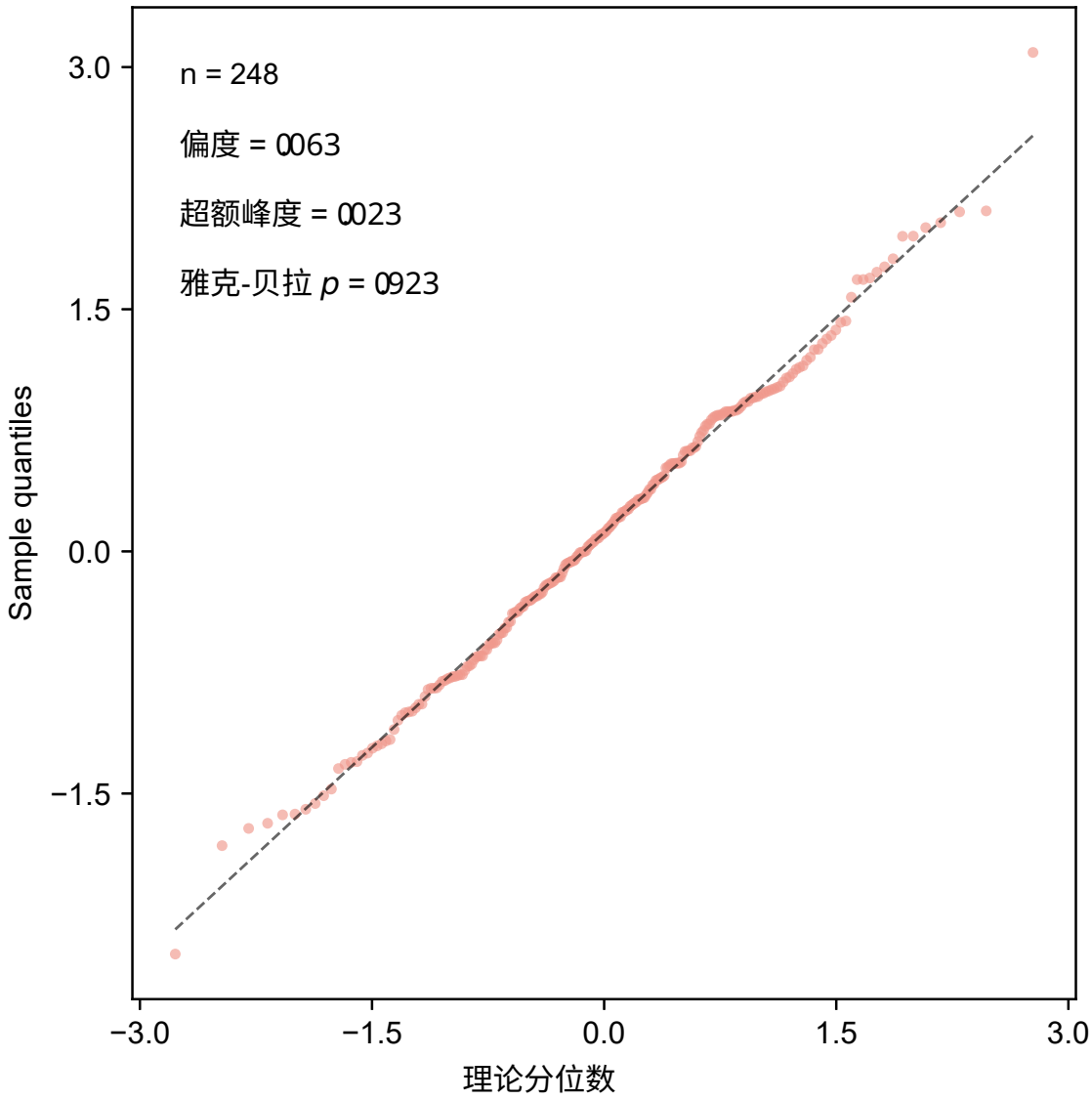
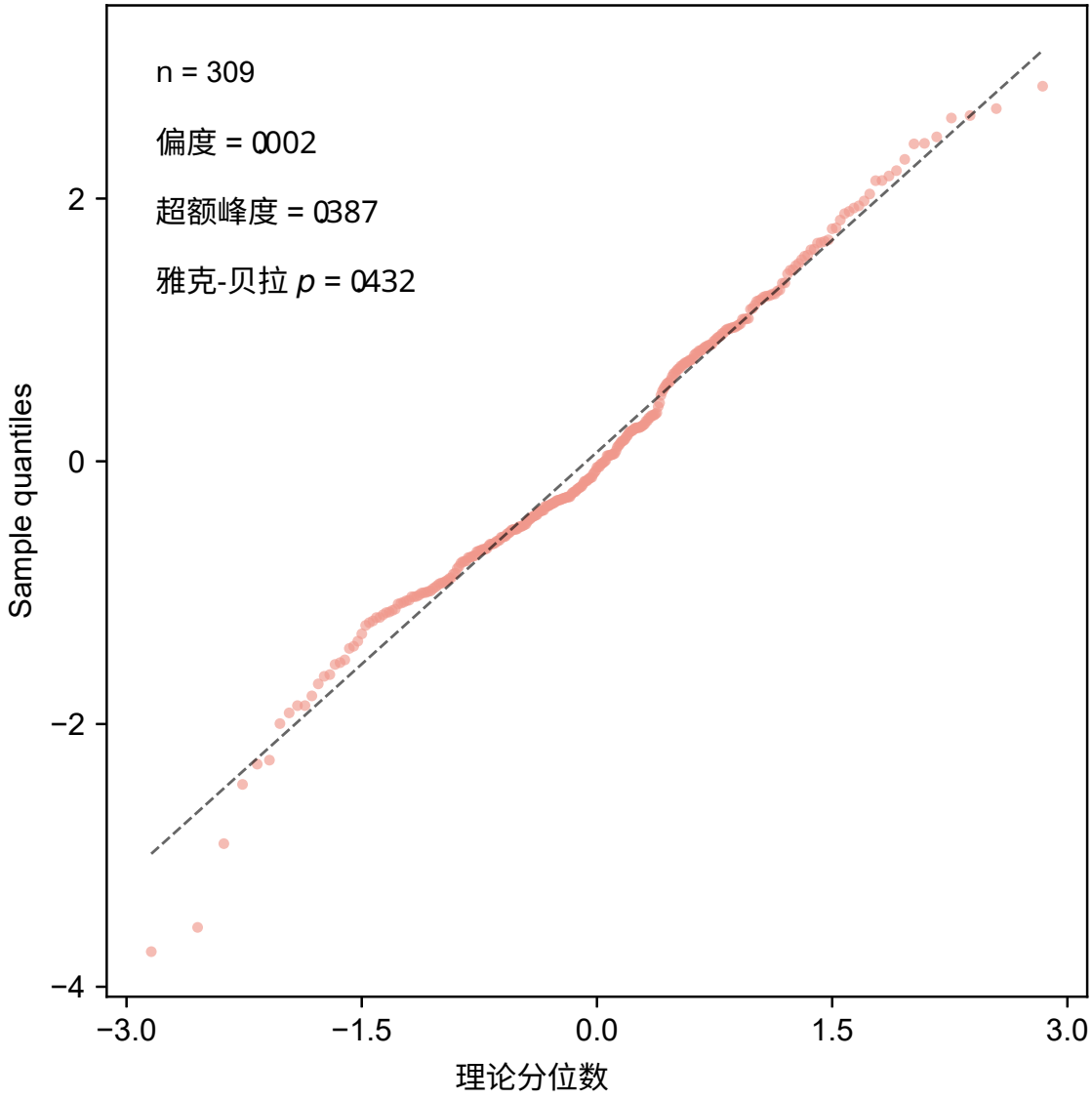


图 S10. 在负活性子集上重新训练的模型的 r0测试集残差 Q-Q图（评分 ≤ 0, $n = 248$）。参见图 S4。



图S11. 基于负活性子集重新训练的模型在 r349测试集上的残差 Q-Q图（评分 ≤ 0, $n = 309$）。对照图S5。

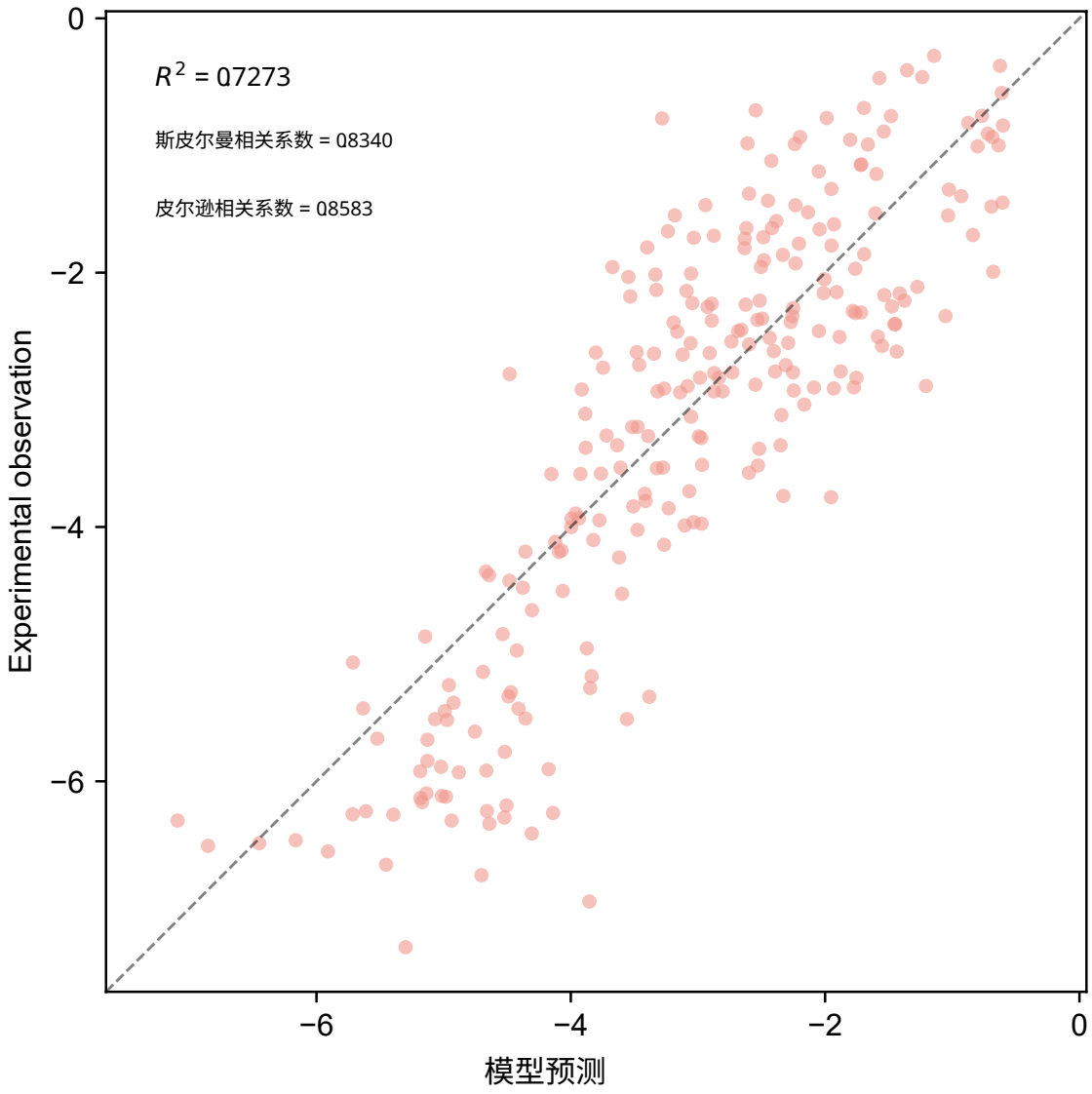
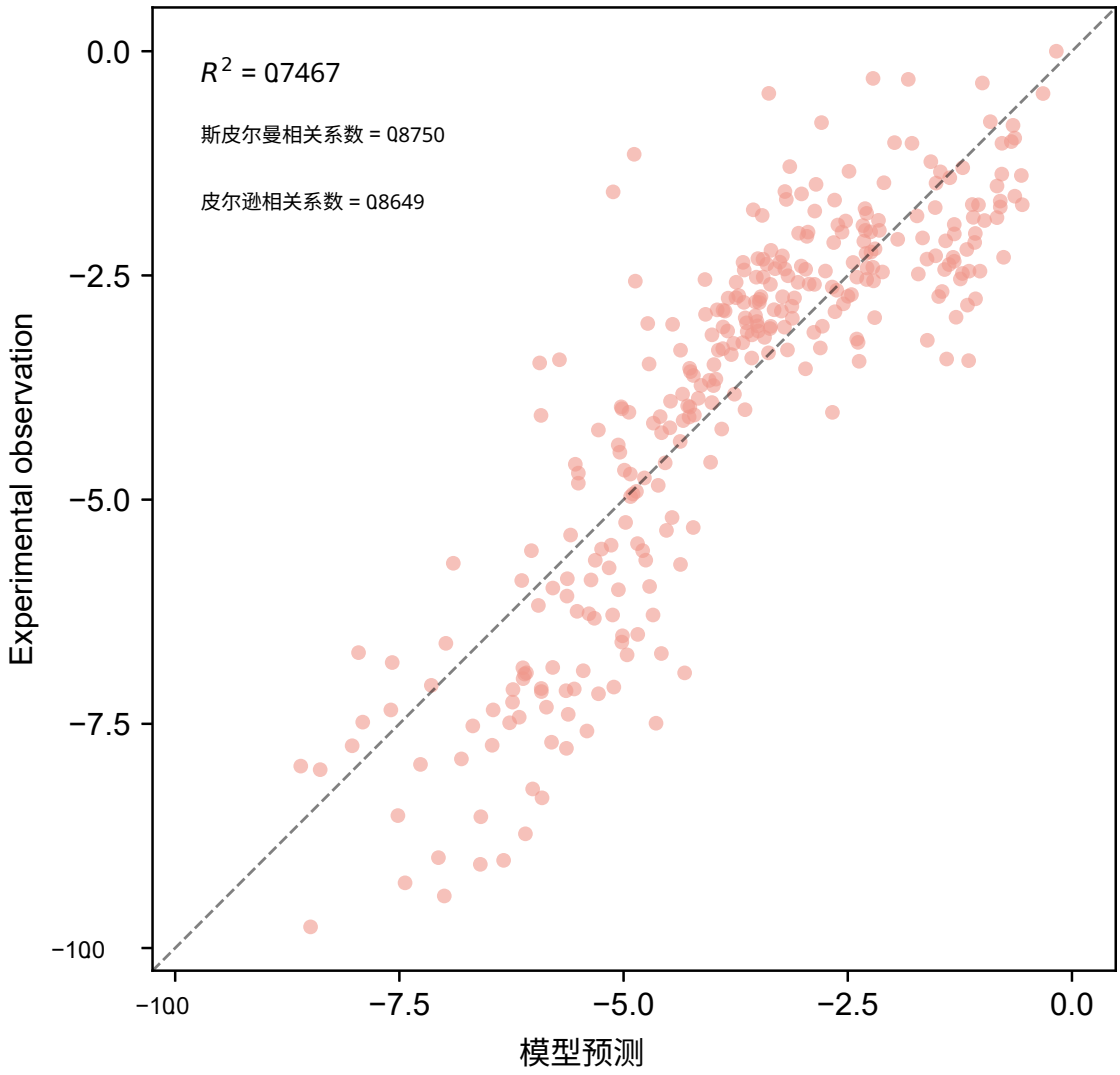


图 S12. r0模型在负活性子集上重新训练后的预测活性与测量活动的对比（评分 ≤ 0）。测试集，$n = 248$。



图S13. 负活性子集上重新训练的r349模型的预测活性与测量活性对比（评分 ≤ 0）。测试集， $n = 309$。