arXiv:2601.14710v1 [cs.LG] 21 Jan 2026

药物发现中的案例引导的序贯检测规划

Tianchi Chen[1], *、Jan Bima[2], 、Sean L. Wu[1]、Otto Ritter[2]、Bingjia Yang[3],和 Xiang Yu[3], †

[1]决策科学，默克公司，美国[2]决策科学，默克公司，捷克共和国[3]药代动力学、药效学、代谢与生物分析，默克公司，美国

摘要

在药物发现中对实验检测进行最优排序，是一个在极端不确定性和资源约束下的高风险规划问题。标准强化学习（RL）面临的主要障碍，是缺乏显式的环境模拟器或转移数据（s，a，s'）；规划只能完全依赖历史结果的静态数据库。我们提出了隐式贝叶斯马尔可夫决策过程（IBMDP），这是一种专为无模拟器环境设计的基于模型的强化学习框架。IBMDP通过利用相似的历史结果构建非参数信念分布，形成由案例引导的转移动态隐式模型。随着证据不断累积，该机制能够实现贝叶斯信念更新，并采用集成蒙特卡洛树搜索规划生成稳定的策略，在提高目标结果的信息增益与资源效率之间取得平衡。我们通过全面的实验验证了IBMDP。在一项真实世界的中枢神经系统（CNS）药物发现任务中，与成熟的启发式方法相比，IBMDP在保持决策置信度的同时，最多可减少92%的资源消耗。为严格评估决策质量，我们还在具有可计算最优策略的合成环境中对IBMDP进行了基准测试。与使用相同基于相似性的模型的确定性价值迭代替代方案相比，我们的框架与该最优策略实现了显著更高的一致性，证明了集成规划器的优越性。IBMDP为数据丰富但模拟器匮乏的领域中的序贯实验设计提供了一种切实可行的解决方案。

1 引言

为了发现新药，科学家需要依次决定开展多项检测，而时间、预算和材料往往有限 [Paul et al., 2010, Hughes et al., 2011]。这一过程通常始于从既往化合物的历史检测结果中获得的有限证据。对新的候选药物化合物开展检测，会产生一项观测，同时消耗资金和时间资源；每项检测都能探查化合物可开发性的一个不同方面（例如效力、ADME、安全性） [Segall, 2012]。例如，体外 检测可能成本低、速度快，但对后续决策提供的信息有限；而体内检测虽然更慢、成本更高，却能为 “ 继续 / 终止 ” 决策提供更有力的依据。在预算和进度都十分紧张的情况下，核心问题是：应当再开展一项检测，还是现在就停止？理想情况下，每项选定的检测都能降低后验不确定性，同时提高化合物符合预设可开发性标准的似然。这是一个在不确定性下的规划问题，而缺少转移元组（s, a, s'）使问题更加复杂——我们只能获取既往化合物的历史检测结果。在实践中，基于规则的操作手册和专家启发式方法往往过于厌恶风险或目光短浅，导致有限资源未能得到有效利用，投资组合结果也不尽理想。

为应对这些挑战，我们提出隐式贝叶斯马尔可夫决策过程（IBMDP），这是一种用于案例引导的序贯检测规划的强化学习（RL）框架，它利用历史化合物的检测结果，构建一个关于通过检测获得的信息增益的隐式概率模型。在每个步骤中，IBMDP 使用方差归一化距离核，对历史化合物记录构建类别分布，并采样与当前部分证据相符的合理检测结果，从而更新候选对象的已观测状态。这种隐式的非参数转移模型侧重于与候选对象最相关的情境，无需显式的机理模拟器。规划采用带双重渐进扩展的蒙特卡洛树搜索（MCTS-DPW） [Browne 等人，2012，Cou¨etoux 等人， 2011], ，并运行多个蒙特卡洛树搜索规划器组成的集成

*通讯作者：tianchi.chen@merck.com
†通讯作者：yuxiang822@gmail.com

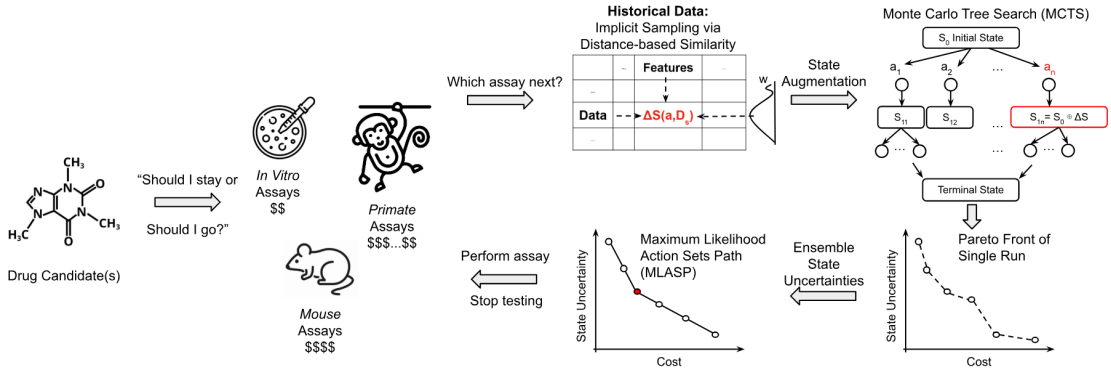

图 1：通过数据驱动、类似物引导的规划模拟器，在药物发现中进行序贯决策；该模拟器维护着对最相关历史化合物类似物的贝叶斯信念。

以降低随机采样和树搜索带来的方差；在多次运行中进行多数投票，得到在不同不确定性水平下均稳定的最大似然动作集路径（MLASP）。在搜索过程中模拟可能的动作序列时，IBMDP 会考虑不确定性向理想状态（例如，体内药物相似性高的状态）降低所带来的影响，并且仅当不确定性朝理想状态降低到足够幅度时，才推荐下一项检测。从部分可观测马尔可夫决策过程（POMDP）的角度来看，标准方法会维护关于隐藏状态的显式概率分布，并通过贝叶斯定理更新这些分布；而 IBMDP 则根据与当前观测状态的相似度对过往经验加权，并从中采样以作出决策（附录 A）。尽管 IBMDP 以放弃形式上的收敛保证为代价，换取在无模拟器环境中的实际适用性，但它通过集成蒙特卡洛树搜索规划提供了经验证稳健的策略（附录 A.6）。

贡献。(i) 基于证据自适应动态的强化学习规划：不同于使用固定转移函数的传统强化学习，IBMDP 的隐式动力学会随着观测的积累而演变——基于相似性的信念持续自适应，从而根据静态历史数据生成非平稳但有原则的状态转移。**(ii) 基于相似度加权**

贝叶斯信念机制：我们将历史结果转化为自适应生成模型，使转移概率能够根据不断积累的证据动态变化，从而无需显式动力学或（s，a，s'）轨迹即可进行规划。**(iii) 在非平稳动态下稳健的集成蒙特卡洛树搜索：**我们的集成方法采用多数投票（MLASP），即使转移模型不断演变，也能生成稳定的策略，并最优地平衡信息增益与资源效率。

2 预备知识

化合物、检测与历史数据令 $\mathcal{A} = \{a_1, \ldots, a_M\}$ 为可用测定的集合，$X = \{x_1, \ldots, x_N\}$ 为历史化合物的集合，其中每种化合物都有固定的分子表示。对于历史化合物 x_i 和检测 a_j，观测结果记为 $y_{i,\,j}$。完整的历史数据集表示为一个元组集合：

$$\mathcal{D} = \{(x_i, \mathbf{y}_i)\}_{i=1}^{N},$$

其中，$\mathbf{y}_i = (y_{i,\,1}, \ldots\ldots, y_{i,\,M})$ 是化合物 x_i 的所有检测结果构成的向量。我们将正在规划的候选药物化合物记为 $x_\star \equiv x_{N+1}$。此外，我们还可以使用针对各项检测的预测模型，例如定量构效关系（QSAR）模型。这些模型是函数 $f_j : x \mapsto \hat{y}_j = f_j(x)$，可在规划阶段 [Chen 等人，2024]针对候选对象 $x_\star$ 进行查询。为方便起见，附录 F 汇总了全文使用的所有符号，见全局符号参考。

目标属性 令 g 为我们关注的主要标量目标，例如决定化合物成败的明确体内 终点。该目标的历史值构成集合 $G = \{g_i\}_{i=1}^{N}$。在许多应用中，目标属性可能对应于某项可用测定。也就是说，对于某个特定的检测索引 $j \in \{1, \ldots, M\}$，所有化合物均满足 $g_i \equiv y_{i,\,j}$。至关重要的是，为防止数据泄漏，在规划过程中绝不会使用目标值集合 G 来

计算相似度或距离指标。我们定义 $I_g = \{i : \quad g_i \text{ 可用}\}$ 为目标值已测定的历史化合物索引集合。

状态与动作我们将候选对象 $x_\star$ 的实验规划问题表述为一个有限时域马尔可夫决策过程，其折扣因子为 $\gamma \in [0,1)$，最大时域为 T。在任意决策步骤 t，我们维护一个已执行检测的指标集 $M_t \subseteq \{1,\dots,M\}$，以及未测定检测项的集合 $U_t := \{1,\dots,M\} \setminus M_t$。该过程从空的已测检测集合 $M_0 = \emptyset$开始。

步骤 t 的**状态**概括了关于候选化合物的所有累积知识： $s_t = \big(x_\star, \quad \{y_{\star,j}\}_{j\in M_t}\big)$.步骤 t的**动作集合** $\mathcal{A}_t$包括选择一批最多包含 m 个当前未测定检测项进行检测，或决定结束实验。形式化表示为： $\mathcal{A}_t = \mathcal{P}_{\le m}(U_t) \cup \{\text{eox}\}$. 其中， $\mathcal{P}_{\le m}(U_t)$表示 U_t的所有子集构成的集合，且子集大小至多为 m；“eox”（实验结束）是终止动作。参数 $m \le M$ 是用户指定的通量上限，用于限制单个步骤中可并行执行的检测数量。执行动作 $A_t \subseteq U_t$ 会揭示结果 $\{y_{\star,j}\}_{j\in A_t}$，并更新下一步骤 $t+1$的已测集合和未测集合。

奖励函数 每个动作都会根据其消耗的资源（例如时间、材料和金钱开销）产生相应的成本。设 $c_j \in \mathbb{R}^q_{\ge 0}$ 为单项检测 a_j的成本向量。批次动作 A_t 的成本是其中各项检测成本之和，即 $c(s_t,\ A_t) = \sum_{a_j\in A_t} c_j$。设 $\boldsymbol{\rho} \in \mathbb{R}^q_{\ge 0}$ 为用户定义的权重向量，用于指定不同资源之间的权衡。标量步奖励 $R(s_t,\ A_t)$ 定义如下：

$$R(s_t, A_t) = \begin{cases} -\boldsymbol{\rho}^{\mathsf{T}} c(s_t, A_t), & \text{if } A_t \in \mathcal{A}_t \setminus \{\text{eox}\}, \\ 0, & \text{if } A_t = \text{eox}. \end{cases}$$

不确定性与目标似然泛函 为确保资源用于可行的候选药物，我们基于历史记录上的相似度权重 $w_i(s_t)$，定义两个关键的、依赖于状态的标量函数（将在后文正式定义）。首先，我们对权重重新归一化，仅考虑目标值 g 可用的历史化合物：

$$\tilde{w}_i(s_t) = \frac{w_i(s_t)}{\sum_{\ell\in I_g} w_\ell(s_t)} \qquad \text{for } i \in I_g.$$

请注意，当 $|I_g| \ll N$ 时，这种重新归一化可能会因有效样本量受限而导致方差低估。实验分析部分将讨论这一局限性。利用这些归一化后的权重，我们定义：

1. **状态不确定性 ($H(s_t)$)：** 目标属性 g 在相关历史数据上的加权方差，用于衡量候选对象潜在结果的不确定性。

$$H(s_t) = \sum_{i\in I_g} \tilde{w}_i(s_t)\,\big(g_i - \bar{g}(s_t)\big)^2, \quad \text{where} \quad \bar{g}(s_t) = \sum_{i\in I_g} \tilde{w}_i(s_t)\, g_i. \tag{1}$$

2. **目标似然度 ($L(s_t)$)：** 候选对象的目标属性落在预定义理想范围 $[g_{\min}, g_{\max}]$内的加权概率。

$$L(s_t) = \sum_{i\in I_g} \tilde{w}_i(s_t)\, \mathbf{1}[\, g_i \in [g_{\min}, g_{\max}]\,]. \tag{2}$$

Her其中， $\mathbf{1}[\cdot]$ 表示指示函数，当其参数为真时返回 $\mathbf{1}$，否则返回 $\mathbf{0}$。 e.

约束目标 最优策略 π^* 是指在满足终端不确定性和逐步可行性约束的前提下，使期望奖励总和最大化的策略。具体而言，我们旨在求解：

$$\begin{aligned} \pi^* \in \arg\max_{\pi}\ & \mathbb{E}_\pi\left[\sum_{t=0}^{T} \gamma^t\, R\big(s_t, \pi(s_t)\big)\right] \\ \text{subject to}\quad & \begin{cases} H(s_T) \le \epsilon, \quad \epsilon \in [0,1], \\ L(s_t) \ge \tau, \quad \forall t = 0,\dots,T-1, \end{cases} \end{aligned} \tag{3}$$

其中，$\epsilon > 0$ 是终止状态 s_T 下可容忍的最大不确定性，$\tau \in (0, 1)$ 是每个中间步骤可接受的最低目标似然度。可行性约束 $L(s_t) \geq \tau$ 确保规划过程沿着通向成功结果的轨迹进行，而终端约束 $H(s_T) \leq \epsilon$ 保证决策是在具备足够置信度的情况下作出的。

3 环境动态的隐式模型

关键挑战在于更新状态转移 $s_t \xrightarrow{A_t} s_{t+1}$——即执行一批检测后状态如何演变——而此时没有可用的显式模拟器，只能利用历史数据 $\mathcal{D}$ 推断动态。我们通过构建环境动态的隐式生成模型来应对这一挑战。该模型采用相似度度量 [Bellet 等人， 2013] 动态地重新加权历史化合物特征，从而形成对候选化合物 $x_\star$ 可能结果的信念。这样便无需对转移概率进行显式参数化，并通过从与当前状态 $x_\star$ 最相关的历史化合物类似物中采样，隐式地传播不确定性。

相似性权重计算 转移模型以相似度权重 $w_i(s_t)$ 为核心，该权重分配给每条历史化合物记录 $D_i = (x_i, \mathbf{y}_i) \in \mathcal{D}$。这些权重用于量化每个历史案例与当前状态 s_t 的相关性。权重采用方差归一化指数核计算：

$$w_i(s_t) = \exp\left(-\lambda_w \cdot d(s_t, D_i)\right), \tag{4}$$

其中，$d(s_t, D_i)$ 是一种距离度量。该距离基于候选对象 $x_\star$ 在步骤 t 时已知的所有特征集合计算，我们将其记为特征集 $\mathcal{K}_t$。该集合包括所有初始定量构效关系预测，以及 M_t 中所有已测定检测项的结果。该距离将候选对象的这些已知值与历史化合物 x_i 的相应值进行比较：

$$d(s_t, D_i) = \sum_{k \in \mathcal{K}_t} \lambda_k \cdot \frac{(\phi_k(s_t) - \phi_k(D_i))^2}{\sigma_k^2}, \tag{5}$$

这里，$\phi_k(\cdot)$ 是一个提取函数，用于从给定状态或历史记录中返回第 k 个特征的值。对于候选对象，$\phi_k(s_t)$ 要么是定量构效关系预测，要么是已测定检测项的结果 $\{y_{\star,j}\}_{j \in M_t}$。对于历史化合物，$\phi_k(D_i)$ 是相应的记录值。术语 σ_k^2 是特征 k 在历史数据集 $\mathcal{D}$ 中的经验方差，计算方式为 $\sigma_k^2 = \frac{1}{N}\sum_{i=1}^{N}(\phi_k(D_i) - \bar{\phi}_k)^2$ 其中 $\bar{\phi}_k = \frac{1}{N}\sum_{i=1}^{N}\phi_k(D_i)$。参数 λ_k 是特征专属权重，而 λ_w 是全局温度参数。方差归一化确保了不同尺度特征之间的无量纲比较。

基于相似性的状态转移 执行一个动作（一个检测批次）$A_t \subseteq U_t$ 后，从状态 s_t 转移到 s_{t+1} 的过程通过加权采样过程进行模拟。首先，从 $\mathcal{D}$ 中采样一个历史案例，其概率与其相似度权重成正比：

$$I \sim \text{Categorical}\left(\frac{w_1(s_t)}{Z}, \ldots, \frac{w_N(s_t)}{Z}\right), \quad \text{where } Z = \sum_{i=1}^{N} w_i(s_t).$$

设选中的历史案例为 $D_I = (x_I, \mathbf{y}_I)$。然后，通过从该采样案例中取出相应的值，“揭示”动作批次 A_t 中各项检测的结果：

$$\{y_{\star,j} := y_{I,j}\}_{j \in A_t}.$$

通过将这些新生成的结果添加到先前状态中，形成新状态 s_{t+1}。形式上，$M_{t+1} = M_t \cup A_t$，新状态为：

$$s_{t+1} = \left(x_\star, \{y_{\star,j}\}_{j \in M_{t+1}}\right).$$

这种生成过程确保新检测的模拟结果与合理且有历史记录的化合物特征一致，从而保留检测之间的相关性。

通过采样进行隐式转移建模 上述采样机制定义了一个隐式转移概率分布 $P(s_{t+1}|s_t,\ A_t)$。该分布是一个混合模型，其中每个分量对应于 $\mathcal{D}$ 中的一个历史案例。转移到特定下一个状态 s_{t+1} 的概率，是所有会产生该状态的历史案例的权重总和：

$$P(s_{t+1}|s_t, A_t) = \sum_{i=1}^{N} \frac{w_i(s_t)}{Z} \cdot \mathbf{1}[s_{t+1} = s_t \oplus \{(a_j, y_{i,j})\}_{j\in A_t}], \tag{6}$$

其中，$\oplus$ 表示状态更新操作：通过向 M_t 添加新的检测结果并更新观测值 $\{y_{\star,j}\}$来扩充当前状态，而 $\mathbf{1}[\cdot]$ 是指示函数。当历史数据集 $\mathcal{D}$ 足够代表底层系统时，这种基于采样的方法可以近似真实的转移动态。近似质量取决于 $\mathcal{D}$的覆盖范围、相似度度量的适当性以及数据集大小 $N = |\mathcal{D}|$。

贝叶斯权重更新 转移到新状态 s_{t+1}后，会重新评估相似度权重，以纳入新的证据。这一重新计算是贝叶斯信念更新的直接且有原则的实现。正如我们在附录 A中正式推导的那样，我们的框架等价于一个POMDP，其中隐藏状态是 $\mathcal{D}$中历史案例上的潜在索引。相似度权重 $w_i(s_t)$表示这些潜在"原型"上的后验信念，而在观测到新的检测结果后进行重新计算，则等价于应用贝叶斯定理。新的权重 $\{w_i(s_{t+1})\}_{i=1}^{N}$ 使用与之前相同的距离函数进行计算，但现在该函数应用于扩充后的状态 s_{t+1}：

$$w_i(s_{t+1}) = \exp\left(-\lambda_w \cdot d(s_{t+1}, D_i)\right). \tag{7}$$

这种更新机制会将模型的信念转向与候选对象 $x_\star$ 扩展后的证据集最为一致的历史案例，使规划器能够随着数据不断积累，完善预测和后续决策。

4 隐式贝叶斯马尔可夫决策过程 (IBMDP)

隐式贝叶斯马尔可夫决策过程 (IBMDP) 是一种规划框架，旨在求解方程 equation 3 中定义的约束优化问题。它将隐式的、基于案例的转移模型与强大的规划算法相结合，以寻找奖励最大化的检测序列。该框架的核心是蒙特卡洛树搜索 (MCTS) 规划器，它利用从历史数据中推导出的生成模型模拟潜在的实验路径，从而探索决策空间。为确保建议的稳健性，IBMDP 采用集成方法，汇总多次独立规划运行的结果。

整体工作流程如下：历史数据 $\mathcal{D}$ 为相似度模块提供信息，该模块为当前状态 s_t计算权重 $w_i(s_t)$。这些权重驱动 MCTS-DPW 规划器所使用的隐式转移模型。规划器生成一个策略，并在多次集成运行中重复这一过程。最后，将各策略汇总，构建最大似然动作集路径 (MLASP)，作为最终建议的实验计划。详细流程见算法 1。

IBMDP 从初始状态 s_0 开始，其中包含候选化合物 $x_\star$ 以及任何已有的定量构效关系预测，且已测定检测项集合为空（$M_0 = \emptyset$）。规划器采用双重渐进扩展蒙特卡洛树搜索（MCTS-DPW），特别适合处理这个问题，因为它能够应对规模庞大的组合动作空间——在此处，即未测定检测项的幂集，$\mathcal{P}_{\leq m}$（U_t）。

在蒙特卡洛树搜索算法的每个模拟步骤中，规划器都必须评估执行某个动作 A_t 的后果。它通过调用上一节介绍的隐式转移模型来完成此操作。根据当前的相似度权重 w_i（s_t），抽取一个历史案例 D_I，并从该案例中获取 A_t 中各检测项的结果。由此得到模拟的下一状态 s_{t+1}。随后，规划器重新计算该新状态的相似度权重，w_i（s_{t+1}），并评估状态不确定性 H（s_{t+1}）和目标似然度 L（s_{t+1}）。如果不确定性 H（s）低于阈值 ϵ，则该状态被视为终止状态；若在任何步骤中违反可行性约束 L（s） $< \tau$，规划器便会收到负奖励。即时步奖励 R（s_t，A_t）也会被记录下来。这个过程使蒙特卡洛树搜索能够构建一棵搜索树，准确反映奖励（资源效率）与面向期望状态的预期信息增益之间的权衡，并始终以历史数据为指导。

为减轻随机性的影响，我们多次运行规划过程，以形成一个集成。最终建议（MLASP）通过在每个阶段对集成策略所推荐的动作进行多数投票来构建。

算法 1 集成式 IBMDP 算法 要求：初始状态 $s_0 = (x_\star, M_0 = \emptyset)$，历史数据 $\mathcal{D}$，奖励函数 $R(s, A)$，泛函 $H(s)$ $L(s)$，阈值 ϵ, τ，时域 T，迭代次数 n_{itr}，集成规模 N_e。确保：最大似然动作集路径（MLASP）。 1: 初始化策略集 $\Pi \leftarrow \emptyset$。 2: 对 $j = 1$ 到 N_e 执行 ▷ 集成循环 3: 初始化蒙特卡洛树搜索树 $\mathcal{T}$，其根节点为 s_0。 4: 对 $i = 1$ 到 n_{itr} 执行 ▷ MCTS迭代 5: 选择：从 s_0 遍历 $\mathcal{T}$，使用树策略（例如 UCB1）选择一个叶节点 s_{leaf}。 6: 扩展：如果 s_{leaf} 不是终止状态（$H(s_{leaf}) > \epsilon$），则选择一个未尝试的动作 $A \in \mathcal{A}_t(s_{leaf})$，并创建一个新的子节点 s_{new}。 7: 模拟（rollout）：从 s_{new} 开始，使用奖励感知启发式策略模拟状态和动作的轨迹，直到达到终止状态或时域 T。 8: rollout 期间，对于转移 $(s, A) \to s'$，隐式模型生成下一个状态 s'：从 $\text{Cat}(\{w_k(s)/Z\}_{k=1}^N)$ 中采样 $I \sim$，其中 $Z = \sum_{k=1}^N w_k(s)$，并设置 $s' = s \oplus \{(a_k, y_{I,k})\}_{k \in A}$。 9: 总回报 Q 为累积奖励；如果轨迹中的任一状态满足 $L(s) < \tau$，则给予较大的负奖励（例如 -10^6）。 10: 反向传播：使用回报 Q，更新从 s_{new} 回溯至根节点路径上所有节点的访问次数和价值估计。 11: 结束循环 12: 从最终树 $\mathcal{T}$ 中提取最优策略 π_j^*，方法是在每个节点选择价值最高的动作。 13: $\Pi \leftarrow \Pi \cup \{\pi_j^*\}$。 14: 结束循环 15: 在每个决策步骤，通过多数投票汇总 Π中的策略，构建 MLASP。 16: 返回 MLASP

这确保了规划具有稳健性，而非单次模拟运行所造成的假象。

5 实验

我们通过两部分评估验证 IBMDP 的性能。首先，我们将其应用于中枢神经系统（CNS）药物发现中的一项真实世界序贯检测规划任务，以展示其实际效用和节省资源的潜力。为确保可复现性，我们在一个公开数据集上进行了相同实验，该数据集用于选择大鼠和犬的体内药代动力学检测，以确定人类的体内 清除率（附录 E）。其次，我们建立了一个具有已知最优策略的合成环境，以严格评估其决策过程的质量。

5.1 BrainPenetrationAssays：真实世界案例研究

问题设定与数据。我们在一项面向中枢神经系统（CNS）药物发现的序贯检测规划任务中评估 IBMDP，其目标是高效确定某种化合物的脑渗透潜力。这一性质在很大程度上取决于该化合物穿过血脑屏障（BBB）的能力。决策涉及在廉价、快速但信息量较少的体外 转运体检测（P-糖蛋白，P-糖蛋白；乳腺癌耐药蛋白，BCRP）与一种权威但耗时且昂贵的体内 检测之间进行选择；后者用于测量未结合脑/血浆分配系数（k_{puu}）。

我们的历史数据集 $\mathcal{D}$ 包含 $N = 220$ 个化合物，这些化合物均已完成所有相关检测（100 nM P-糖蛋白、1μM P-糖蛋白、100 nM BCRP）和目标属性 k_{puu} 的测量。所有化合物还具有相应的定量构效关系预测，可为检测结果（例如 P-糖蛋白和 BCRP 活性）以及平均滞留时间（MRT）等其他相关属性提供初始估计。运营成本定义为：每项体外 检测的成本为 400 美元、周转时间为 7 天；体内 k_{puu} 检测的成本为 4,000 美元、周转时间为 21 天。每个步骤的动作最多限于 3 项并行检测。如果某个化合物的 $k_{\text{puu}} > 0.5$，则认为其具有较高潜力。

实验设置。 规划目标是在降低状态不确定性 $H(s)$（关于目标 k_{puu}）、提高目标似然 $L(s)$（即 k_{puu} 处于理想范围内）以及最大化奖励（高效利用资源）之间取得平衡。当规划器达到足够置信度时，实验序列终止；足够置信度由联合标准定义，即对于所有中间步骤， $H(s_T) \leq \epsilon$ 和 $L(s_t) \geq \tau$。所得结果将与传统的基于规则的决策策略进行比较。

基于规则的基线。 在实践中，决策通常遵循基于定量构效关系预测的简单启发式方法。对于这项任务，基线启发式方法如下：

• 如果 QSAR （可能 $0.5 \leq k_{\text{puu}} \leq 1$）符合 $\text{QSAR}_{\text{1uM P-糖蛋白}} < 2$ 且 $\text{QSAR}_{\text{100nM BCRP}} < 2$，则认为该化合物**有望成功**。

• 如果 QSAR**显示该化合物不具前景** （可能 $k_{\text{puu}} \leq 0.5$），则将其判定为不具前景：$\text{QSAR}_{\text{1uM P-糖蛋白}} > 4$ 或 $\text{QSAR}_{\text{100nM BCRP}} > 4$。

我们从三个类别中选取了四个具有代表性的场景，评估 IBMDP 在应对这一启发式方法时的表现：**(i)基线确认** （QSAR 信号明确，场景 3），**(ii) 启发式方法挑战** （QSAR 信号相互矛盾或处于临界状态，场景 2 和 4），以及 **(iii)机会发现**（QSAR 表明某化合物不具前景，但该化合物实际上表现良好，场景 1）。

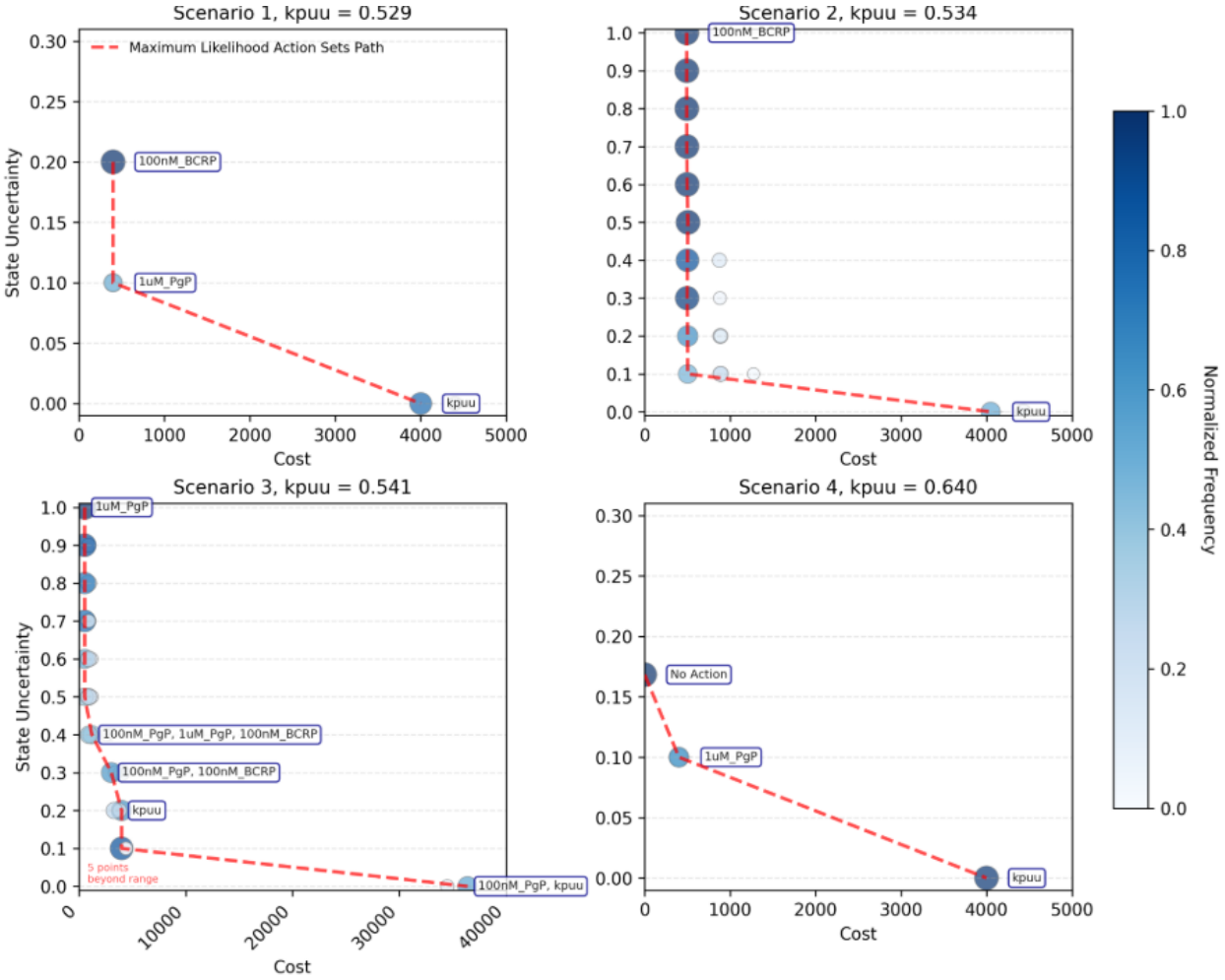

图2：IBMDP 对四种代表性化合物的货币优先结果。每个图均显示可实现的资源消耗与终端状态不确定性之间的帕累托前沿，并突出显示最大似然动作集路径（MLASP）。这展示了 IBMDP 如何提供权衡曲线，使决策者能够根据自身对风险和预算的承受程度选择计划。

结果。 如表1所示，IBMDP始终能够找出比传统方法更节省资源的实验方案；传统方法往往默认运行完整的检测项目组合，耗资5,200美元。表格各行与图2所示的场景一一对应。对于机会发现情景（第1行/场景1），启发式方法会错误地舍弃一种有价值的化合物，而IBMDP建议采用一项高效的800美元方案来揭示其真正潜力。对于下一个化合物（第2行/场景2），IBMDP找出一项仅需400美元的最简方案，以消除不确定性。在基线确认情景（第3行/场景3）中，IBMDP建议仅花费400至800美元来确认这一有前景的特征。最后，对于QSAR相互矛盾这一棘手情况（第4行/场景4），IBMDP仅需400至800美元便能有效消除不确定性。在这些具有代表性的案例中，IBMDP以最多减少92%的资源消耗，实现了相同或更高的决策置信度。

表1：传统启发式方法与 IBMDP 对代表性化合物进行检测的资源开销比较。各行按图2中的情景1-4排序。传统方法的开销为 $5200，反映了运行完整检测面板的费用（ k_{puu} 为 $4000，另加体外 检测的 $400）。IBMDP 始终避免了这项开销。

定量构效关系预测器		检测				开销（×$100）	
1uM PgP	100nM BCRP	kpuu	100nM PgP	1uM PgP	100nM BCRP	传统方法	IBMDP
5.0	9.6	0.53	15.9	12.9	8.2	52	8
0.9	8.5	0.53	2.2	1.1	14.2	52	4
1.7	1.3	0.54	1.1	0.8	1.3	52	4 - 8
21.4	0.7	0.64	17.4	19.7	0.8	52	4 - 8

5.2 合成数据模拟

基准测试设置。 为在受控环境中严格评估IBMDP的策略质量，我们使用一个可计算理论最优策略的合成数据集对其进行基准测试（完整细节见附录D）。我们使用真实的解析不确定性动态，通过价值迭代得出该最优策略（VI-Theo）。随后，我们将IBMDP与VI-Theo基线以及一个确定性变体进行比较；后者采用与IBMDP相同的基于相似性的估计，但使用价值迭代进行规划（VI-Sim）。

结果。 表 2 总结的结果展示了 IBMDP 的随机集成规划方法的有效性。在 100 次独立试验中，IBMDP 的主要推荐（前一推荐）有 47% 的情况与最优 VI-Theo 策略对齐。相比之下，确定性的 VI-Sim 方法仅实现了 36% 的对齐率。集成方法的优势还体现在：最优动作有 66% 的概率包含在 IBMDP 的前两项推荐中，从而为高价值策略空间提供稳健且有效的覆盖。

表 2：与理论基准的策略对齐

方法	匹配	匹配率（%）
IBMDP 前一推荐	47	47.0
IBMDP 前二名	66	66.0
基于 VI 的相似度	36	36.0

这种优异性能源于策略生成方式的根本差异。基于价值迭代的方法会收敛到单一的确定性策略，而 IBMDP 的蒙特卡洛树搜索智能体集成则更广泛地探索策略空间。这使其能够识别多个通常近乎等价的高价值动作；在测定选择中，这一点尤为有利，因为不同的特征组合可能带来相似的信息增益。结果证实，我们的基于集成的规划器能够有效应对策略空间本身固有的不确定性，因此比确定性替代方案提供更稳健、更可靠的建议。

6 结论

为实现案例引导规划，我们提出了 IBMDP，这是一种强化学习框架，可将历史案例转化为用于序贯检测选择的生成模型。通过根据相似度对历史案例加权，该算法无需显式的转移函数，即可借助蒙特卡洛树搜索进行稳健的多步规划。在一个现实世界的药物发现问题中的应用表明，该方法能够以更少、更廉价的检测揭示化合物的真实情况。这项工作建立了一种利用过往经验指导未来实验的强大方法，对于药物发现以外的众多领域同样具有广泛适用性；在这些领域中，历史数据丰富，但机理模型较为匮乏。

7 相关工作

马尔可夫决策过程与基于模型的强化学习。马尔可夫决策过程将序贯决策形式化 [Puterman, 2014] [Puterman, 1994, Alagoz 等人, 2010]；基于模型的强化学习学习动态以用于规划 [Sutton 和 Barto, 2018年, Kaiser 等人, 2019年, Moerland

等人，2023]。基于核的强化学习主要利用相似度来近似价值或平滑学习得到的转移[Ormoneit and Sen, 2002, Kveton and Theocharous, 2012, Xu et al., 2007]。IBMDP 利用相似度构建生成式非参数转移无(s, a, s') 元组——从历史记录中采样检测结果，而非学习关于后续状态的显式核函数。

贝叶斯强化学习与贝叶斯优化。贝叶斯强化学习维护模型参数或价值的后验分布，并对显式马尔可夫决策过程进行采样（例如，后验采样强化学习） [Ghavamzadeh 等，2015年，Osband 等，2013年，Agrawal 和 Jia， 2017]。贝叶斯优化旨在一次性改进目标函数 [Griffiths 和 Hern´andez-Lobato，2020年，G´omez-Bombarelli 等， 2018] Shahriari 等， 2015]。IBMDP 避免显式参数后验分布，通过对记录进行相似度加权重加权来实现贝叶斯基于案例的生成，从而支持通过奖励优化和可行性约束进行多步规划。

贝叶斯实验设计与隐式模型。经典的单步贝叶斯优化/贝叶斯实验设计方法具有短视性，并假设存在显式似然或模拟器[Chaloner 和 Verdinelli，1995年，Rainforth 等， 2024]；隐式贝叶斯实验设计通过信息论代理或策略学习处理难以处理的似然 [Kleinegesse 和 Gutmann，2020年，Kleinegesse 等，2021年，Ivanova 等， 2021] [Blau 等， 2022]。IBMDP 将隐式模型嵌入强化学习规划器（MCTS-DPW）中，兼顾奖励、时间和可行性，而不只关注信息增益。

受约束马尔可夫决策过程和部分可观测马尔可夫决策过程。 CMDP 通常对累积成本施加约束 [Achiam 等人， 2017]；我们的约束针对状态 属性（终端不确定性、逐步似然），并在规划过程中加以执行。这一设定与部分可观测马尔可夫决策过程类似 [Kaelbling 等人， 1998]；我们基于记录构建的相似度加权后验充当隐式信念。多步强化学习/部分可观测马尔可夫决策过程求解器需要模拟器或 (s, a, s') 元组，而 IBMDP 则直接基于历史检测概况构建相似度加权的隐式生成模型 ，在不学习显式动力学的情况下保留检测间依赖关系。因此，若不进行大量调整，便无法进行严格可比的直接基准比较：(i) 将效用重新定义为针对检测 而非输入，(ii) 增加资源感知型 批次处理，并制定与我们对 $H(s)$ 和 $L(s)$ 的约束相一致的原则性停止 规则，以及 (iii) 提供与无模拟器设定相一致的后验预测分布 （附录 A）。

强化学习中的集成方法。集成方法能够提升稳健性并提供不确定性估计 [Dietterich, 2000, Zhou, 2012, Wier- ing and Van Hasselt, 2008, Osband et al., 2016,Lakshminarayanan et al., 2017]。IBMDP务实地采用集成方法来稳定随机规划。

应用背景。既有生物医学领域的强化学习研究主要聚焦于试验或分子生成/合成 [Bennett and Hauser, 2013, Eghbali-Zarch et al., 2019, Abbas et al., 2007, Fard et al., 2018, Wang et al., 2021, Bengio et al., 2021, You et al., 2018, Zhou et al., 2019, Segler et al., 2018, Coronato et al., 2020,Escandell-Montero et al., 2014, Schneider et al., 2020]。早期发现阶段的测定选择仍未得到充分探索。IBMDP提供了一个实用的规划器，可将历史测定记录转化为连贯的生成式转移模型，并纳入操作约束，从而应对发现研究中典型的“无（s， a， s'）”情形。

关于与相关方法进行公平比较。 虽然上述方法看似相关，但直接进行基准比较在根本上并不公平——每种方法都基于不同的数学假设和问题表述。基于模型的强化学习需要 $(s$， a， $s')$ 数据或模拟器；贝叶斯强化学习从参数后验分布中采样；贝叶斯优化执行单步优化；部分可观测马尔可夫决策过程需要显式的转移模型。IBMDP 独特地针对仅存在静态历史结果的情形，因此这些比较可谓“苹果比橘子”。详细分析请参见附录 C。

8 局限性

历史数据覆盖度。 方法的有效性取决于 $\mathcal{D}$ 的质量和代表性；数据缺口或偏差可能导致次优选择。与可进行探索的无模型强化学习不同，基于相似性的采样无法发现 $\mathcal{D}$ 中不存在的策略——不过，在数据稀缺的情况下，稳定的物理化学规律在一定程度上可以降低这种风险。

相似度度量假设。 对（归一化）欧几里得距离采用指数核，意味着假设这些距离能够反映测定行为。非线性或阈值生物学可能不符合这一假设；要捕捉构效关系，可能需要采用领域定制度量。

可扩展性。 最坏情况下的总复杂度为 $O(N_e \cdot n_{\text{itr}} \cdot \min(b^H,\ n_{\text{itr}}) \cdot |\mathcal{D}| \cdot d)$，其中 b 为有效分支因子，H 为最大树深度。每次迭代的开销主要来自相似度权重计算（$O(|\mathcal{D}| \cdot d)$）。大型数据集 $|\mathcal{D}|$ 或高特征维度 d 可能会给内存和计算带来压力，在企业级应用中可能需要分布式基础设施或数据子采样。

超参数敏感性。性能取决于对 $\{\lambda_w, \lambda_k, N_e, c, \epsilon, \tau\}$ 的调优；要在不同项目中保持稳健性，可能需要专家先验或相当充足的验证预算。

参考文献

Ismail Abbas、Joan Rovira 和 Josep Casanovas。临床试验优化：用于规划临床试验招募的蒙特卡罗模拟马尔可夫模型。载于《当代临床试验》，第 28 卷，第 220–231 页。爱思唯尔，2007年。Joshua Achiam、David Held、Aviv Tamar 和 Pieter Abbeel。约束策略优化。载于国际机器学习会议，第 22–31 页。PMLR，2017年。Shipra Agrawal 和 Randy Jia。强化学习后验采样：最坏情况下的遗憾界。神经信息处理系统进展，第 30 卷，2017年。Oguzhan Alagoz、Heather Hsu、Andrew J Schaefer 和 Mark S Roberts。马尔可夫决策过程：在不确定性下进行序贯决策的工具。《医学决策》，30(4):474–483，2010年。Aur´elien Bellet、Amaury Habrard 和 Marc Sebban。特征向量和结构化数据度量学习综述。*arXiv* 预印本 *arXiv:1306.6709*，2013年。Emmanuel Bengio、Moksh Jain、Maksym Korablyov、Doina Precup 和 Yoshua Bengio。基于流网络的生成模型：用于非迭代式多样化候选生成。神经信息处理系统进展，第 34 卷，第 27381–27394 页，2021年。Casey C Bennett 和 Kris Hauser。用于模拟临床决策的人工智能框架：一种马尔可夫决策过程方法。《医学人工智能》，57(1):9–19，2013年。Tom Blau、Edwin V Bonilla、Iadine Chades 和 Amir Dezfouli。利用深度强化学习优化序贯实验设计。载于国际机器学习会议，第 2107–2128 页。PMLR，2022年。

Cameron B Browne、Edward Powley、Daniel Whitehouse、Simon M Lucas、Peter I Cowling、Philipp Rohlfshagen、Stephen Tavener、Diego Perez、Spyridon Samothrakis 和 Simon Colton。蒙特卡洛树搜索方法综述。《*IEEE*计算智能与游戏人工智能汇刊》，4(1):1–43，2012。

Kathryn Chaloner 和 Isabella Verdinelli。贝叶斯实验设计：综述。《统计科学》，第273–304页，1995。

Jacky Chen、Yunsie Chung、Jonathan Tynan、Chen Cheng、Song Yang 和 Alan Cheng。小分子药物发现模型的性能洞见：数据规模扩展、多任务学习与泛化。2024。

Antonio Coronato、Muddasar Naeem、Giuseppe De Pietro 和 Giovanni Paragliola。面向智能医疗应用的强化学习：综述。《医学人工智能》，109:101964，2020。Adrien Cou¨etoux、Jean-Baptiste Hoock、Nataliya Sokolovska、Olivier Teytaud 和 Nicolas Bonnard。连续上置信树。载于学习与智能优化：第五届国际会议 *LION 5*，意大利罗马，*2011*年*1*月*17—21*日。论文选集 *5*，第433–445页。Springer，2011。Thomas G Dietterich。机器学习中的集成方法。载于多分类器系统国际研讨会，第1–15页。Springer，2000。Maryam Eghbali-Zarch、Reza Tavakkoli-Moghaddam、Fatemeh Esfahanian、Amir Azaron 和 Mohammad Mehdi Sepehri。用于模拟2型糖尿病药物治疗中药物不良反应的马尔可夫决策过程。《机械工程师学会会刊*H*辑：医学工程杂志》，233(8):793–811，2019。Pablo Escandell-Montero、Milena Chermisi、Jose M Martinez-Martinez、Juan Gomez-Sanchis、Carlo Barbieri、Emilio Soria-Olivas、Flavio Mari、Joan Vila-Franc´es、Andrea Stopper、Emanuele Gatti 等。通过强化学习优化血液透析患者的贫血治疗。《医学人工智能》，62(1):47–60，2014。Mahdi M Fard、Sandor Szalma、Shashikant Vattikuti 和 Gyan Bhanot。用于临床试验最优决策的贝叶斯马尔可夫决策过程框架。《*IEEE*生物医学与健康信息学杂志》，22(6):2061–2068，2018。

Mohammad Ghavamzadeh、Shie Mannor、Joelle Pineau 和 Aviv Tamar。贝叶斯强化学习：综述。《机器学习基础与趋势》，8(5-6):359–483，2015。Rafael G´omez-Bombarelli、Jennifer N Wei、David Duvenaud、Jos´e Miguel Hern´andez-Lobato、Benjam´ın S´anchez-Lengeling、Dennis Sheberla、Jorge Aguilera-Iparraguirre、Timothy D Hirzel、Ryan P Adams 和 Al´an Aspuru-Guzik。利用分子的数据驱动连续表示实现自动化学设计。*ACS Central Science*，4(2):268–276，2018。Ryan-Rhys Griffiths 和 Jos´e Miguel Hern´andez-Lobato。利用变分自编码器进行自动化学设计的约束贝叶斯优化。《化学科学》，11(2):577–586，2020。John P Hughes、Simon Rees、Sonya B Kalindjian 和 Roger K Philpott。早期药物发现的原则。《英国药理学杂志》，162(6):1239–1249，2011。Desislava R Ivanova、Adam Foster、Simon Kleinegesse、Michael U Gutmann 和 Tom Rainforth。隐式深度自适应设计：无需似然的基于策略的实验设计。载于神经信息处理系统进展，第34卷，第25785–25798页，2021。Leslie Pack Kaelbling、Michael L Littman 和 Anthony R Cassandra。部分可观测随机领域中的规划与行动。《人工智能》，101(1-2):99–134，1998。Lukasz Kaiser、Mohammad Babaeizadeh、Piotr Milos、Blazej Osinski、Roy H Campbell、Konrad Czechowski、Dumitru Erhan、Chelsea Finn、Piotr Kozakowski、Sergey Levine 等。面向 Atari 的基于模型的强化学习。国际学习表征会议，2019。Steven Kleinegesse 和 Michael U Gutmann。通过互信息神经估计，为隐式模型开展贝叶斯实验设计。载于国际机器学习会议，第5316–5326页。PMLR，2020。Steven Kleinegesse、Christopher Drovandi 和 Michael U Gutmann。利用互信息下界，为隐式模型开展基于梯度的贝叶斯实验设计。*arXiv*预印本 *arXiv:2105.04379*，2021。Branislav Kveton 和 Georgios Theocharous。基于代表性状态的核强化学习。*AAAI*人工智能会议，26(1):977–983，2012。Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。利用深度集成进行简单且可扩展的预测不确定性估计。神经信息处理系统进展，30，2017。Thomas M Moerland、Joost Broekens、Aske Plaat 和 Catholijn M Jonker。基于模型的强化学习：综述。《机器学习基础与趋势》，16(1):1–118，2023。´Dirk Ormoneit 和 Saunak Sen。基于核的强化学习。《机器学习》，49:161–178，2002。Ian Osband、Daniel Russo 和 Benjamin Van Roy。通过后验采样实现更有效的强化学习。载于神经信息处理系统进展，第26卷，2013。Ian Osband、Charles Blundell、Alexander Pritzel 和 Benjamin Van Roy。通过自助法DQN实现深度探索。神经信息处理系统进展，29，2016。Steven M Paul、Daryl S Mytelka、Christopher T Dunwiddie、Christian C Persinger、Bernard H Munos、Stacy R Lindborg 和 Aaron L Schacht。如何提高研发生产力：制药行业面临的重大挑战。*Nature Reviews Drug Discovery*，9(3):203–214，2010。Martin L Puterman。马尔可夫决策过程：离散随机动态规划。John Wiley & Sons，1994。Martin L Puterman。马尔可夫决策过程：离散随机动态规划。John Wiley & Sons，2014。Tom Rainforth、Adam Foster、Desi R Ivanova 和 Freddie Bickford Smith。现代贝叶斯实验设计。《统计科学》，39(1):100–114，2024。

Petra Schneider、W Patrick Walters、Alleyn T Plowright、Norman Sieroka、Jennifer Listgarten、Robert A Goodnow Jr、Jasmin Fisher、Johanna M Jansen、Jos´e S Duca、Thomas S Rush 等。《人工智能时代重新思考药物设计》。*Nature Reviews Drug Discovery*，19(5):353–364，2020年。Matthew D Segall。多参数优化：识别具有均衡属性的高质量化合物。*Current Pharmaceutical Design*，18(9):1292–1310，2012。Marwin HS Segler、Mike Preuss 和 Mark P Waller。利用深度神经网络和符号人工智能规划化学合成。《自然》，555(7698):604–610，2018年。Bobak Shahriari、Kevin Swersky、Ziyu Wang、Ryan P Adams 和 Nando De Freitas。让人类退出闭环：贝叶斯优化综述。*Proceedings of theIEEE*，104(1):148–175，2015年。Richard S Sutton 和 Andrew G Barto。《强化学习：导论》。MIT出版社，2018年。Jike Wang、Chang-Yu Hsieh、Mingyang Wang、Xiaorui Wang、Zhenxing Wu、Dejun Jiang、Benben Liao、Xujun Zhang、Bo Yang、Qiaojun He 等。基于条件变换器、知识蒸馏和强化学习的多约束分子生成。*NatureMachine Intelligence*，3(10):914–922，2021年。Marco A Wiering 和 Hado Van Hasselt。强化学习中的集成算法。*IEEE*系统、人和控制论汇刊*B*辑（控制论），38(4):930–936，2008年。Xin Xu、Dewen Hu 和 Xicheng Lu。用于强化学习的基于核的最小二乘策略迭代。*IEEE*神经网络汇刊，18(4):973–992，2007年。Bingjia Yang、Yunsie Chung、Archer Y Yang、Bo Yuan、Tianchi Chen 和 Xiang Yu。QComp：一种基于定量构效关系的药物发现插补框架。*Journal of Chemical Information and Modeling*，65(15):7862–7873，2025年。Jiaxuan You、Bowen Liu、Zhitao Ying、Vijay Pande 和 Jure Leskovec。用于目标导向分子图生成的图卷积策略网络。神经信息处理系统进展，31，2018年。Zhenpeng Zhou、Steven Kearnes、Li Li、Richard N Zare 和 Patrick Riley。通过深度强化学习优化分子。*Scientific Reports*，9(1):10752，2019年。Zhi-Hua Zhou。《集成方法：基础与算法》。CRC出版社，2012。

附录

A 理论框架：IB-MDP作为POMDP

本附录为 IB-MDP 框架提供正式的概念基础。我们证明，我们基于相似度加权的案例方法并非临时启发式方法，而是在针对信息收集问题定制的部分可观测马尔可夫决策过程（POMDP）中，对贝叶斯信念更新的一种计算上易于处理的实现。

A.1 POMDP预备知识

POMDP 的形式化定义为元组 $(\mathcal{S}, \mathcal{A}, \Omega\ P, O, R, \gamma)$，其中 $\mathcal{S}$是隐藏状态集合， $\mathcal{A}$ 是动作集合，Ω 是观测集合。由于智能体无法观测真实状态 $s \in \mathcal{S}$，因此会维护一个信念状态 $b_t(s)$，即定义在 $\mathcal{S}$上的概率分布。采取动作 A_t 并收到观测 ω_t后，信念会通过贝叶斯滤波器进行更新：

$$b_{t+1}(s') \ \propto\ O(\omega_t \mid s', A_t) \sum_{s \in \mathcal{S}} P(s' \mid s, A_t)\, b_t(s). \tag{8}$$

信息获取情形。 序贯测定规划任务属于一种信息获取问题。候选化合物 $x_\star$ 的内在属性是固定的；执行检测会揭示有关这些属性的信息，但不会改变这些属性。这对应于静态潜在状态，此时转移概率为恒等函数：$P(s' \mid s, A_t) = \mathbf{1}[s' = s]$。在这一常见的特殊情形下，方程 equation 8 中的信念更新可简化为乘法形式的贝叶斯定理：

$$b_{t+1}(s) \ \propto\ O(\omega_t \mid s, A_t)\, b_t(s). \tag{9}$$

A.2 IB-MDP潜在索引模型及其与POMDP的对应关系

为了将我们的框架映射到POMDP，我们引入一个离散潜变量 $Z \in \{1, \ldots, N\}$，其中每个取值 i都对应一个历史记录 $D_i \in \mathcal{D}$。我们将 Z视为隐藏状态，代表候选化合物 $x_\star$ 在已知历史案例中的“真实原型”。核心思想是，通过维护关于Z的信念，我们实际上是在维护关于 $x_\star$完整且未观测到的特征概况的信念。具体对应关系详见表3。

A.3 相似度更新与贝叶斯滤波的等价性

建立映射后， 我们现在证明，IB-MDP中的相似度权重更新机制是 t
一种直接实现 根据方程 9 进行贝叶斯信念更新。

设步骤 t 之前潜在索引的先验信念为权重 $w_i(s_t) \equiv P(Z = i \mid s_t)$。执行测定批次 A_t 后，得到观测 $\omega_t \equiv \{y_{\star,j}\}_{j \in A_t}$。将 IB-MDP 中的对应项代入方程 9，可得 IB-MDP 的权重信念更新规则：

$$w_i(s_{t+1}) = \frac{p(\omega_t \mid Z = i, A_t)\ w_i(s_t)}{\sum_{\ell=1}^{N} p(\omega_t \mid Z = \ell, A_t)\ w_\ell(s_t)}. \tag{10}$$

这证实了 IB-MDP 中权重的演化遵循一种有原则的贝叶斯递推。

将似然与相似性核联系起来。 最后一步是说明，我们对相似度权重的具体实现对应于一个有效的概率似然模型。若用以历史值 y_i为中心的高斯核来建模候选对象观测到检测结果 y_a 的似然，则：$_{,a}$

$$p(y_a \mid Z = i, a) \ \propto\ \exp\!\left(-\frac{\lambda_a}{2}\,\frac{(y_a - y_{i,a})^2}{\sigma_a^2}\right),$$

并假设在给定原型 Z 的条件下，批次中的各项检测相互独立（这一建模假设使推断易于处理；即使已知化合物性质，生化测定实验仍可能存在相关性，但我们的实证结果表明，通过集成平均机制，即使违反这一假设，模型仍具有稳健性），则观测 ω_t 的联合似然是各个单独似然的乘积。从对所有已观测检测 $\{y_a\}_{a\in M_t}$ 的均匀先验出发递归应用贝叶斯更新，可得关于 Z 的后验分布，其形式与我们的相似度权重完全一致：

$$w_i(s_t) \;\propto\; \prod_{a\in M_t} p(y_a \mid Z=i,a) = \exp\!\left(-\frac{1}{2}\sum_{a\in M_t}\lambda_a\,\frac{(y_a-y_{i,a})^2}{\sigma_a^2}\right) \equiv \exp\!\big(-\lambda_w\, d(s_t,D_i)\big),$$

在这里，我们可以确定全局温度参数 $\lambda_w=\beta/2$，其中 β 是温度化后验分布的逆温度。当 $\beta=1$（标准后验分布）时，我们有 $\lambda_w=1/2$。因此，我们的相似度函数并非任意的启发式方法，而是对应于潜在历史原型上的温度化贝叶斯后验分布。

A.4 后验预测转移模型

信念状态（权重向量 $w(s_t)$）用于规划。在蒙特卡洛树搜索规划器中用于模拟未来轨迹的转移模型，是通过对潜变量 Z 的不确定性进行边缘化得到的。转移到下一状态 s_{t+1} 的概率是以当前信念为条件的结果后验预测分布：

$$P(s_{t+1}\mid s_t,A_t) = \sum_{i=1}^{N} P(s_{t+1}\mid Z=i,s_t,A_t)P(Z=i\mid s_t) \tag{11}$$

$$= \sum_{i=1}^{N} w_i(s_t)\;\delta_{\,s_t\oplus\{(a_j,y_{i,j})\}_{j\in A_t}}(s_{t+1}), \tag{12}$$

其中，$\delta_x(y)$ 表示狄拉克 δ 测度；当 $y=x$ 时其值为 1，否则为 0。这证实了我们的采样机制——根据权重 $w_i(s_t)$ 抽取一个历史案例 D_i，并使用其结果——是从后验预测分布中采样的一种合理方法，使规划器能够探索与迄今收集的所有证据相符的合理未来情景。

A.5 含义与总结

在 POMDP 框架下阐释 IB-MDP 为其提供了坚实的概念基础，并带来了若干关键洞见，详见表3。

• **动态性的合理性**：在一次 MCTS模拟中观察到的相似度权重变化并非任意的非平稳性，而是代表了智能体不断变化的信念状态。随着信息的收集（状态s_t 得到扩充），用于后续预测的模型会自然且正确地发生变化，反映出更加精细的信念。

• **MCTS 的适用性**：MCTS 非常适合此任务，因为它是一种基于模拟的规划器，旨在处理复杂的状态空间。它能够有效探索动作对未来信念状态及其相关奖励的影响，而无需对信念空间本身进行显式表示。

• **有原则的近似**：IB-MDP 为求解形式化 POMDP 提供了一种实用的数据驱动的近似方法。其有效性依赖于两个关键假设：历史数据 $\mathcal{D}$ 的质量与覆盖范围，以及所选核（例如高斯核）对于观测似然模型的适用性。尽管高斯核具有计算上的可处理性，并且符合关于生化测定实验中测量噪声的常见假设，但我们承认，其他核（例如拉普拉斯核、Student-t 核）可能更适合捕捉重尾分布或离群值。我们的方法对核选择的稳健性仍是未来实证验证中的一个重要研究方向。

• **计算效率**：通过使用案例库 $\mathcal{D}$ 上的权重隐式表示信念状态，IB-MDP 避免了在可能极其庞大的隐藏状态空间上维护并更新显式概率分布所需的复杂计算。

总之，将 IB-MDP 解读为一种近似 POMDP 框架，明确表明重新计算相似度权重是贝叶斯信念更新的直接实现。这为我们的方法以及在仅有历史数据可用时使用蒙特卡洛树搜索进行有原则的不确定性下规划提供了依据。

表 3：POMDP 与 IB-MDP 之间概念映射的摘要。

POMDP 组件	**IB-MDP 概念对应项**	**备注**
隐藏状态（$s \in \mathcal{S}$）	历史案例的潜在索引 $Z = i$ $D_i \in \mathcal{D}$	候选对象的“真实”但未知的特征。
信念状态（$b_t(s)$）	相似度权重 $w_i(s_t) \equiv P(Z = i \mid s_t)$	对可能的原型进行概率分布，原型。
动作（$A_t \in \mathcal{A}_t$）	要执行的检测批次，$A_t \subseteq U_t$	直接等价。
观测（ω_t）	检测结果集合 $\{y_{\star,j}\}_{j \in A_t}$	收集到的新证据。
观测模型（$O(\omega_t \mid s', A_t)$）	似然 $p(\omega_t \mid Z = i, A_t)$	通过相似度核实现-nel.
信念更新	权重的重新计算 $w_i(s_{t+1})$	一种直接且有理论依据的贝叶斯更新-日期。

A.6 收敛保证与理论考量

IBMDP 的收敛特性与传统贝叶斯强化学习有着根本区别，因为它独特地依赖历史数据，而非与环境交互。本小节将探讨其能够和不能够提供哪些收敛保证。

A.6.1 传统贝叶斯强化学习保证

诸如强化学习后验采样（PSRL）的方法，为有限马尔可夫决策过程提供形式化的遗憾界 $\tilde{O}(\sqrt{SAT})$，其中 S 表示状态，A 表示动作，T 表示时域。这些方法保证以较高概率实现 PAC 风格收敛，达到 ϵ-最优策略；其依据的原则是，随着数据不断累积，后验分布会逐渐集中于真实马尔可夫决策过程。

A.6.2 IBMDP 收敛特性

由于采用隐式模型构建，IBMDP 的收敛行为更加微妙：

可实现的保证。 标准的 MCTS 配合 UCB1，在假设马尔可夫决策过程模型固定的情况下，随着迭代次数趋于无穷，可渐近收敛到最优策略。具体而言，对于 IBMDP：

- MCTS-DPW 收敛性：IBMDP 中采用的双重渐进拓宽变体，在大型组合动作空间中仍能保持收敛性
- 线性情形一致性：对于具有线性关系和独立特征的合成数据，我们证明（见第 D 节），基于相似度的方差估计量依概率收敛于真实条件方差，当$N \to \infty$
- 经验稳健性：采用多数投票的集成方法可提供稳定的建议，其最优策略一致率达到 47%，而确定性方法为 36%

根本局限。 与传统的贝叶斯强化学习不同，IBMDP 无法提供：

- 形式化遗憾界：隐式模型引入的近似误差受历史数据覆盖度限制，而非收敛至真实动力学
- PAC 保证：由于依赖数据代表性，无法以高概率确保达到 ϵ 最优性
- 真实动力学恢复：基于相似性的模型只能近似真实转移函数，而无法学习该函数$P(s' \mid s, a)$

近似质量取决于三个关键因素：（i）历史数据$\mathcal{D}$的覆盖范围和代表性，（ii）相似度度量是否适用于该领域，以及（iii）历史数据集 $|\mathcal{D}|$的规模。虽然在无法获取真实动力学的情况下，无法确立形式化收敛速率，但实证验证表明，在满足这些因素时，该方法具有稳健的性能。

收敛性权衡。 IBMDP以形式化收敛保证换取实际适用性。在数据覆盖充分、相似度指标适当且满足正则性条件的假设下，随着MCTS迭代次数 $n_{\text{itr}} \to \infty$ 和集成规模 $N_e \to \infty$:

$$||\pi_{\text{IBMDP}} - \pi^*_{\text{empirical}}||_\infty \xrightarrow{P} 0 \tag{13}$$

其中， $\pi^*_{\text{empirical}}$ 是由 $\mathcal{D}$所诱导的经验马尔可夫决策过程的最优策略。然而，该经验最优策略与真实最优策略之间的差距取决于数据质量和覆盖范围——这是在没有模拟器时运行所面临的根本局限。

这种权衡并非 弱点，而是一种必要的适应：IBMDP 提供了一种有原则的解决方案，而传统 方法因具 有更强的保证，却会因缺乏与环境交互的能力而完全无法运行 s.

B 实现与算法细节

B.1 超参数选择方法

IB-MDP框架的性能取决于一组关键超参数，这些参数决定了相似性模型、蒙特卡洛树搜索规划器以及问题的目标约束。我们实验中使用的取值综合参考了既有文献指南、针对特定数据集的实证测试以及领域相关考量，以在决策质量与计算可行性之间取得平衡。

相似性模型参数。这些参数构成了隐式生成式转移模型的核心。

• **相似性带宽（λ_w）：** 此参数控制相似度函数的"平滑度"。根据 A节中的理论推导， $\lambda_w = \beta/2$ 其中 β 为逆温度。对于标准后验分布（$\beta = 1$），我们有 $\lambda_w = 0.5$。不过在实践中，我们测试了范围为 $[0.5, 2.0]$ 的取值，发现$\lambda_w = 1.0$ 在我们的数据集（$|\mathcal{D}| = 220$， $d = 6$）上表现更好。这对应于 $\beta = 2.0$的温度化后验分布，其取值足够大，能够确保局部性（赋予真正相似的化合物更高权重），同时又足够小，使模型能够从足够多的历史样本中获取支持，从而做出稳健的预测。

• **特征权重（$\lambda_k = 1.0$ 针对所有 k）：** 这些权重可用于在距离计算中提高或降低信息量不同的特征的重要性。由于我们缺乏关于定量构效关系预测和测定结果相对可靠性的详细先验信息，因此将所有权重设为相等，以避免引入主观偏差。这种做法将所有已知特征视为对确定相似度同等重要。

集成与规划器参数。这些参数控制蒙特卡洛树搜索算法及其最终策略的稳健性。

• **集成规模（$N_e = 50$）：** 为减轻转移模型和MCTS规划器的随机性所带来的方差，我们采用由多个独立运行组成的集成。我们测试了从20到100的不同规模，发现 $N_e = 50$能够提供稳定的策略建议（即一致的MLASP），且不会带来过高的计算成本。图3展示了集成的多数投票如何得出稳健的动作选择。

• **MCTS迭代次数（$n_{\text{迭代}} = 20,\ 000$）：** 该参数决定每次MCTS运行的搜索预算。我们的分析表明，对于本问题的复杂程度，策略建议在约20,000次迭代后趋于稳定，继续增加迭代次数的收益则会逐渐降低。

• **探索常数（$c = 5.0$）：** 遵循蒙特卡洛树搜索的标准做法，该值用于平衡对新动作的探索与对搜索树中已知高价值动作的利用。在我们的实验中，5.0 的取值实现了有效探索。

问题约束参数。这些参数定义了规划问题本身的终止条件和可行性约束。

• **终端不确定性阈值 ($\epsilon = 0.10$):** 当 $H(s_T)$ 低于 0.10 时，我们便停止。在我们的运行中，初始不确定性介于 0.2 和 0.6 之间，因此该阈值可确保在判定策略具有足够置信度之前，不确定性至少降低两至六倍。

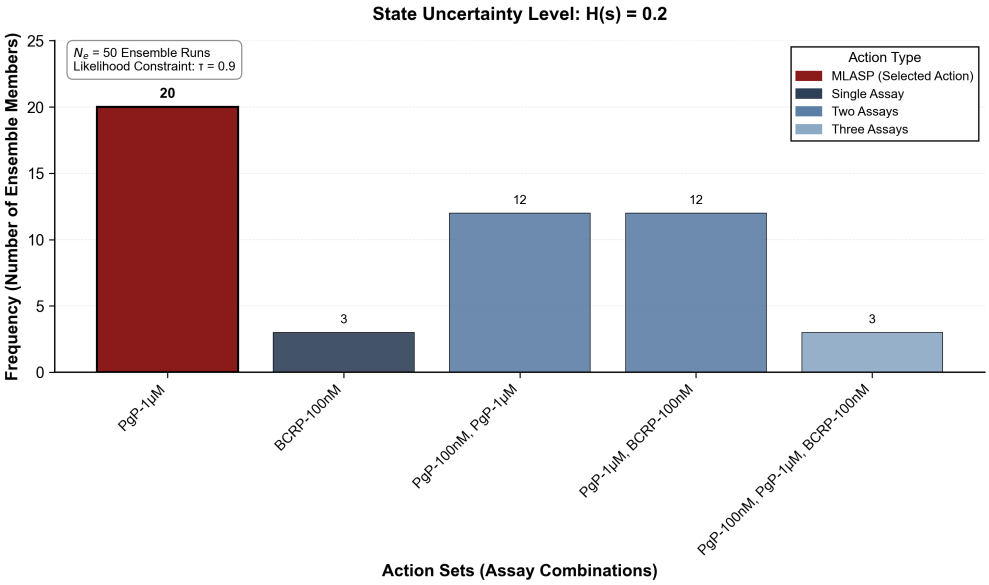

图3：在包含 $N_e = 50$ 次运行的集成中所提出动作的示例直方图。对于具有不确定性 $\mathcal{H}(s) = 0.2$且满足似然约束 $\tau = 0.9$的给定状态，频率最高的动作将被选为 MLASP。此示例展示了集成方法如何通过多数投票生成稳健且稳定的建议。

- **目标似然阈值 ($\tau \in \{0.6, 0.9\}$):** 对目标似然 $L(s_t)$ 设置阈值，可确保规划器只追求成功可能性足够高的轨迹。我们测试了两个取值，以探索成本与置信度之间的权衡。较低的取值（$\tau = 0.6$）允许制定更多探索性、可能成本更低的计划；较高的取值（$\tau = 0.9$）则要求采用更保守、更有把握但可能成本更高的策略。图 4 明确展示了较高的 τ 如何导致采用不同且成本更高的 MLASP，以满足更严格的置信度要求。

B.2 计算复杂度分析

IB-MDP算法的计算复杂度是其实际应用中的一个关键因素。最坏情况下的总复杂度为:

$$O(N_e \cdot n_{\text{itr}} \cdot \min(b^H, n_{\text{itr}}) \cdot |\mathcal{D}| \cdot d),$$

其中， N_e 为集成规模， n_{itr} 为 MCTS迭代次数， b 为有效分支因子（每个节点探索的动作平均数量）， H 为最大树深度（受时域 T 限制）， $|\mathcal{D}|$ 为历史案例数量， d 为特征空间的维度。$\min(b^H, n_{\text{itr}})$ 表示可扩展的最大节点数，其上限由树结构或迭代预算决定。在实践中，采用渐进式扩展和 UCT选择时，实际扩展的数量通常远小于这一最坏情况上界。

单次蒙特卡洛树搜索模拟步骤中的主要耗时因素是计算相似度权重，这需要计算当前状态与 $\mathcal{D}$中的每个历史案例之间的距离。该操作的复杂度为 $O(|\mathcal{D}| \cdot d)$，并在搜索树的每次节点扩展时执行。

与替代方法的比较。尽管这种复杂性不容忽视，但与其他用于严谨地进行不确定性下规划的方法相比，仍有优势。对于这种尺度的问题，精确的 POMDP 求解器在计算上难以处理，因为其复杂度随信念空间的大小呈指数级增长。传统的价值迭代需要对状态空间进行离散化；随着连续值检测数量增加，这种做法将变得不可行。IB-MDP 的基于采样的方法能够有效地探索这一高维空间，无需显式枚举。

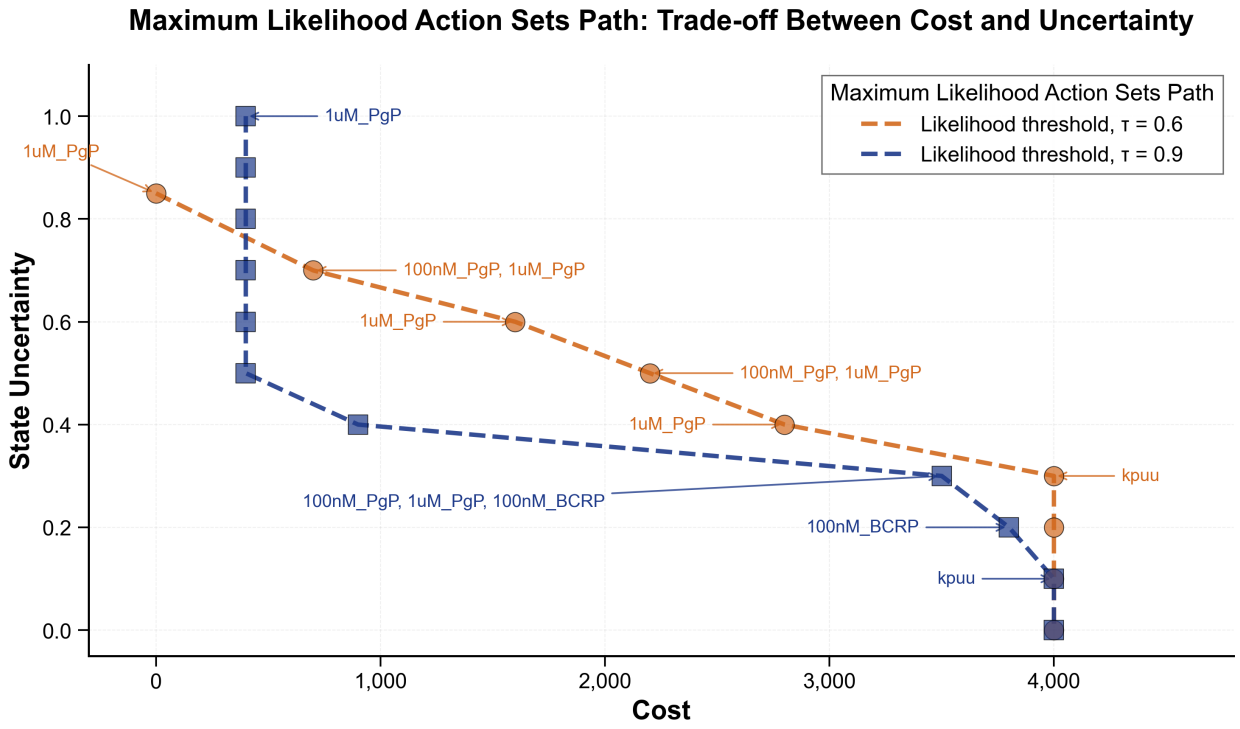

图 4：在两种不同目标似然阈值下，同一化合物的 MLASP 路径比较： τ =0。6（蓝色）和 $\tau = 0.9$ （红色）。更严格的约束（$\tau = 0.9$）会促使规划器推荐成本更高的检测序列，以达到更高的置信度，体现了该参数对成本与决策置信度之间直接权衡的调控作用。

实际性能与可扩展性。在我们的实验设置（$N_e = 50$、 $n_{\text{itr}} = 20,\ 000$、 $|\mathcal{D}| = 220$、 $d = 6$）中，在配备 16GB 内存的 Apple M1 Pro 芯片上，为单个化合物生成策略总共约需一小时。该算法的复杂度与历史数据集的大小（$|\mathcal{D}|$）、特征维度（d）以及集成运行次数（N_e）呈线性关系。这种可预测的扩展趋势表明，对于现实药物发现项目中通常遇到的更大数据集和更高维度问题，该方法在计算上仍然可行，尤其是在具备并行计算资源的情况下。

C 框架　　差异化及直接比较的不公平性　　on

C.1 关键差异特征

表 4 总结了 IBMDP 与传统强化学习框架之间的根本区别。这些差异源于 IBMDP 针对无模拟器、数据丰富环境中的序贯实验规划所进行的独特设计——现有方法若不进行根本性重构，便无法应对这一类问题。

表 4：IBMDP 与传统强化学习框架的根本区别

方面	传统框架	IBMDP
数据需求	$(s,\ a,\ s')$ 元组或模拟器	仅静态历史结果
转移模型	显式学习的 $P(s'\|s,\ a)$	通过相似性采样隐式实现
信念表示	显式概率分布	相似度权重 $w_i(s_t)$
规划方法	单一策略或参数化	采用多数投票的集成蒙特卡洛树搜索
约束	累积型：$\sum_t c_t \leq C$	基于状态：$H(s_T) \leq \epsilon, L(s_t) \geq \tau$
动作空间	参数优化	组合式测定选择
动作效果	改变底层状态	揭示固定属性
相关性处理	需要显式建模	通过采样实现经验保留
目标	单一奖励最大化	多目标优化

C.2 根本性创新

IBMDP 所处的问题设定与传统强化学习截然不同。传统框架假设可以访问环境模拟器或转移数据，而 IBMDP 仅依赖包含历史实验结果的静态数据库。这种限制在药物发现中很常见：机理模型不可用，且实验不可逆，因此必须采用一种根本不同的方法。

The co该机制通过相似度加权采样构建隐式生成模型：

$$w_i(s_t) = \exp\left(-\lambda_w \cdot d(s_t, D_i)\right) \tag{14}$$

$$P(s_{t+1}|s_t, A_t) = \sum_{i=1}^{N} \frac{w_i(s_t)}{Z} \cdot \mathbf{1}[s_{t+1} = s_t \oplus \{(a_j, y_{i,j})\}_{j \in A_t}] \tag{15}$$

该机制通过从与当前实验状态最相似的历史案例中进行采样，生成合理的转移，从而保留真实化合物中观察到的检测间自然相关性——鉴于当前的科学认知水平，这些相关性难以甚至无法通过显式建模来刻画。

C.3 与现有框架类别的比较

C.3.1 与马尔可夫决策过程和基于模型的强化学习的区别

基于模型的强化学习从根本上依赖于从（s，a，s'）元组中学习转移动态，通常通过用于逼近 $P(s'|s,\ a)$ 的参数模型来实现。即使是采用相似度指标的基于核的强化学习方法，也主要将其用于价值函数逼近，而非生成转移。

IBMDP 的不同之处在于，它并非将相似度用作平滑机制，而是将其作为完整生成过程的基础。在无法获取任何转移数据的情况下，它会从历史化合物中抽取完整的检测结果概况，并根据这些化合物与当前状态的相关性进行加权。这种非参数方法绕过了显式动力学建模的需求，同时自然保留了历史数据中存在的跨检测依赖关系。

C.3.2 与贝叶斯方法的区别

贝叶斯强化学习和贝叶斯优化会维护显式的后验分布：前者针对模型参数（以 PSRL 为例），后者针对目标函数。这些方法要么从参数后验分布中抽取完整的马尔可夫决策过程，要么进行贪心的单步优化。

IBMDP 实现了我们所称的贝叶斯基于案例的生成：相似度权重充当历史化合物原型的隐式后验，并随着证据累积而通过重新加权进行更新。与贝叶斯优化只关注单个步骤不同，IBMDP 支持多时域规划，同时考虑实验成本、时间约束以及实现预期结果的概率——这是一种与传统贝叶斯方法有本质区别的多目标优化。

C.3.3 与实验设计的区别

经典贝叶斯实验设计假设可以使用似然函数或模拟器，并以即时信息增益为优化目标。即使是针对难以处理的似然的隐式贝叶斯实验设计方法，也依赖于假设存在某种形式生成模型的信息论替代量。

IBMDP 将其隐式模型直接嵌入强化学习规划器（MCTS-DPW），优化的是完整的实验序列，而非单个实验。框架的状态不确定性泛函 $H(s_t)$ 和目标似然泛函 $L(s_t)$ 提供了可解释的领域特定度量，可直接关联实验目标，不同于抽象的信息论量。

C.3.4 与 POMDP 的区别

标准 POMDP 表述会显式维护隐藏状态上的信念分布，因此需要指定转移模型 $P(s'|s,\ a)$ 和观测模型 $O(o|s,\ a)$。信念更新遵循贝叶斯滤波器，因此需要这些显式模型。

IBMDP 的相似度加权后验充当隐式信念表示，从而无需维护高维信念状态。该框架的约束——终端不确定性 $H(s_T) \leq \epsilon$ 以及每步骤

可行性 $L(s_t) \geq \tau$——针对目标状态属性而非累积型量，直接编码实现决策置信度和轨迹可行性的实验要求。

C.4 为什么不适用与标准强化学习直接基准比较

IBMDP 与传统框架之间存在根本性不兼容，因此无法进行直接基准比较——这些方法需要特定的输入数据格式，而在 IBMDP 所针对的静态历史数据库场景中，这些格式并不存在。

C.4.1 不兼容的前提条件

传统强化学习方法通常都要求具备环境模拟器以按需生成转移，或拥有一组 (s, a, s') 元组来学习动态。IBMDP 恰恰应用于这些前提条件缺失的场景：只有静态历史化合物特征存在，且没有查询反事实结果的机制。创建模拟器需要对生化相互作用具备当前科学尚不具备的机理性理解，而通过穷举实验收集转移数据则违背了高效规划的根本目的。

值得注意的是，保守Q学习或决策变换器等离线（批次）强化学习方法同样无法应用：这些方法需要轨迹数据 (s_t, a_t, r_t, s_{t+1})，用于记录某种行为策略下的序贯决策。历史测定数据库仅包含静态终点特征 $\{(x_i, \mathbf{y}_i)\}$——测定执行的顺序、中间决策状态以及任何评估信号均未被记录。

C.4.2 根本结构差异 s

动作空间存在本质差异。传统方法在连续或离散参数空间中进行优化，其中动作会影响状态转移。IBMDP则从实验检测的组合集合中进行选择——$\mathcal{P}_{\leq m}(U_t) \cup \{\text{eox}\}$——这些动作揭示的是不变分子性质的信息。这种控制与信息收集之间的区别，要求采用完全不同的规划范式。

此外，两者的约束结构并不兼容。传统的受约束马尔可夫决策过程限制轨迹上的累积成本，而IBMDP则施加瞬时可行性约束和终端不确定性界限，以直接体现实验要求。

C.4.3 所需变换

将传统方法应用于此情境，需要：

1. 将动作空间彻底重新定义，从参数优化转为带有批次约束的组合式测定选择；2. 实施与不确定性和可行性泛函相匹配的奖励感知停止规则，而非简单的累积目标；3. 在无法访问模拟器或转移数据的情况下，创建后验预测分布；4. 从单目标优化转为带有基于状态的约束的多目标优化

如此大幅度的修改将从根本上改变这些方法的性质，实质上创造出全新的算法，而非现有算法的变体。由此产生的比较将是 IBMDP 与这些新方法之间的比较，而不是与原有框架的比较。

C.5 理论基础

尽管运行于这一独特设定中，IBMDP 仍有严谨的理论基础。A节证明，该框架在数学上等价于一个 POMDP，其中隐藏状态表示历史案例上的一个潜在索引。相似度权重构成有效的贝叶斯后验分布，权重更新则实现了精确的贝叶斯信念更新。这种等价性：

$$w_i(s_{t+1}) = \frac{p(\omega_t | Z = i, A_t) \cdot w_i(s_t)}{\sum_{\ell=1}^{N} p(\omega_t | Z = \ell, A_t) \cdot w_\ell(s_t)} \tag{16}$$

为我们的实验中观察到的实证成功提供了有力的理论依据：IBMDP 在保持决策质量的同时，最多可将资源消耗降低 92%。

C.6 启示

IBMDP 解决了一类现有强化学习框架并非为之设计的问题：在没有模拟器的情况下进行序贯实验规划。无法进行直接基准比较，反映了 IBMDP 的设置（静态结果向量）与标准强化学习的要求（轨迹数据或模拟器）之间存在数据格式差异；这一差异促使我们在附录D中构建合成基准测试，在其中可以计算真实最优策略。IBMDP 利用历史数据进行相似度加权采样和集成规划，为药物发现及其他传统强化学习假设不成立的类似实验科学领域中的案例引导式序贯决策，提供了首个实用解决方案。

D 合成数据基准测试

概述与动机。本附录通过合成数据开展严格的基准测试研究，将 IBMDP 与理论最优基线及确定性基线进行比较。合成环境具有独特优势：我们能够精确计算真正的最优策略，从而以严谨的方式验证基于相似度的规划方法。在这一受控设置下，我们可以针对真实情况，分离并评估 IBMDP 核心创新的有效性——相似度加权信念机制和集成规划。

目标。 我们提供了一个受控基准测试，用于比较三种序贯检测选择规划器：**(i)** 一种具有精确不确定性动态的理论价值迭代基线（**VI-Theo**）；**(ii)** 一种具有基于相似度的不确定性的确定性价值迭代方法（**VI-Sim**）；以及 **(iii)** 一种随机IBMDP规划器，它在由MCTS-DPW树组成的集成中使用相似度加权后验预测分布转移（**IBMDP**）。具有已知结构的合成数据生成器能够精确计算VI-Theo策略，并为两种基于相似度的规划器提供有原则的测试平台。

与正文的符号保持一致。 我们沿用正文主体部分的符号（状态/动作元组、相似度权重、 $H(s)$、 $L(s)$等；参见公式 equation 1–equation 3 和表 8）。在本附录中，当重点是集合结构而非单个测量值时，为方便表示，我们可能将状态写作 $s = (x_\star,\ \mathcal{M})$，其中 $x_\star$是固定的候选化合物， $\mathcal{M} \subseteq \{1, \dots, M\}$表示已测定检测项的集合。这等同于正文中的符号 $s_t = (x_\star,\ \{y_{\star,j}\}_{j \in M_t})$，其中$\mathcal{M} = M_t$ 为已测定检测项及其数值的索引。当候选化合物固定且上下文明确时，我们也可以将状态简写为 $\mathcal{M}$。

D.1 合成数据：通用模型与实例化

目的。 我们构建了一个合成数据环境，其中真实条件方差可以通过解析方法计算，为评估基于相似度的估计量提供真实情况。具有独立特征的线性结构代表了一个简化但信息丰富的测试案例，在该案例中，理论最优性是可处理的。

D.1.1 一般数据生成过程（通用表述）。

固定整数 N（历史案例数量）和 M（检测项目/特征数量）。对于每个历史案例 $i \in \{1, \dots, N\}$，我们抽取一个特征向量

$$\mathbf{y}_i = (y_{i,1}, \dots, y_{i,M})^\top \in \mathbb{R}^M.$$

对于每个 检测 $a \in \{1, \dots, M\}$ 指定分布参数（μ_a， σ_a， a_a， b_a），并独立抽取 y

$$y_{i,a} \ \sim\ \mathcal{TN}(\mu_a, \sigma_a;\ a_a, b_a),$$

在 $[a_a, b_a]$ 上取具有位置参数 μ_a 和尺度 σ_a的（一元）截断正态分布。[1] 令 $\beta = (\beta_1, \ldots, \beta_M)^\top \in \mathbb{R}^M$，并独立抽取测量噪声

$$\epsilon_i \sim \mathcal{TN}(\mu_\epsilon, \sigma_\epsilon; a_\epsilon, b_\epsilon).$$

标量目标为

$$g_i = \sum_{a=1}^{M} \beta_a y_{i,a} + \epsilon_i = \beta^\top \mathbf{y}_i + \epsilon_i.$$

历史数据集为 $\mathcal{D} = \{(x_i, \mathbf{y}_i)\}$；目标值 $G = \{g_i\}$ 单独存储（且绝不用于任何距离/权重计算）。

独立性假设下的方差闭式解。 令 $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_M^2)$ 表示每项检测的方差参数（视为生成的 $y_{i,a}$ 的经验方差）。那么

$$\mathrm{Var}(g) = \beta^\top \Sigma \beta + \sigma_\epsilon^2.$$

为推导条件方差，我们将检测划分为已测量的 $\mathcal{M}$ 和未测量的 $\mathcal{U}$，并记 $\beta = (\beta_\mathcal{M}, \beta_\mathcal{U}), \Sigma = \mathrm{diag}(\Sigma_\mathcal{M}, \Sigma_\mathcal{U})$。在独立性假设下，推导如下：

$$\begin{aligned}
\mathrm{Var}(g \mid \mathbf{y}_\mathcal{M}) &= \mathrm{Var}\big(\beta_\mathcal{M}^\top \mathbf{y}_\mathcal{M} + \beta_\mathcal{U}^\top \mathbf{y}_\mathcal{U} + \epsilon \mid \mathbf{y}_\mathcal{M}\big) && (17)\\
&= \mathrm{Var}\big(\beta_\mathcal{U}^\top \mathbf{y}_\mathcal{U} + \epsilon \mid \mathbf{y}_\mathcal{M}\big) \quad \text{(since } \beta_\mathcal{M}^\top \mathbf{y}_\mathcal{M} \text{ is fixed given } \mathbf{y}_\mathcal{M}) && (18)\\
&= \mathrm{Var}\big(\beta_\mathcal{U}^\top \mathbf{y}_\mathcal{U}\big) + \mathrm{Var}(\epsilon) \quad \text{(by independence of } \mathbf{y}_\mathcal{U}, \epsilon \text{ from } \mathbf{y}_\mathcal{M}) && (19)\\
&= \beta_\mathcal{U}^\top \mathrm{Var}(\mathbf{y}_\mathcal{U}) \beta_\mathcal{U} + \sigma_\epsilon^2 && (20)\\
&= \beta_\mathcal{U}^\top \Sigma_\mathcal{U} \beta_\mathcal{U} + \sigma_\epsilon^2. && (21)
\end{aligned}$$

这一恒等式是下文 VI-Theo 推导的核心。

D.1.2 基准测试中使用的实例化。

我们设定 $M = 6$ 和 $N = 200$。对于 $a = 1, \ldots, 6$,

$$\mu_a = 50 \cdot \frac{a}{6}, \qquad \sigma_a = 0.3\,\mu_a, \qquad (a_a, b_a) = (0,\, 2\mu_a),$$

$$\beta = (0.3,\, 0.25,\, 0.2,\, 0.15,\, 0.07,\, 0.03), \qquad \epsilon \sim \mathcal{TN}(0, 5;\, -10, 10).$$

所有特征抽样在 a 和 i之间相互独立；噪声与特征相互独立。

采样方案（用于复现）。 对于每次试验：(1) 固定 $M, N, \{\mu_a, \sigma_a, a_a, b_a\}$ 和 β；(2) 逐分量抽取 $\{\mathbf{y}_i\}_{i=1}^N$；(3) 抽取 $\{\epsilon_i\}_{i=1}^N$；(4) 设定 $g_i = \beta^\top \mathbf{y}_i + \epsilon_i$；(5) 存储 $\mathcal{D} = \{(x_i, \mathbf{y}_i)\}$ 和 $G = \{g_i\}$，其可用集合为 $I_g = \{i : g_i \text{ 用于评估}\}$。

D.2 理论基准（VI-Theo）：完整推导

概述。 VI-Theo 表示在完全掌握数据生成过程的情况下的理论最优策略。它使用上文推导的精确条件方差公式，计算每一步的最优不确定性降低。此基准仅适用于已知真实模型参数的合成环境。

状态、动作和动态。 状态是已测定检测项的集合 $\mathcal{M} \subseteq \{1, \ldots, 6\}$（与我们的记号约定一致）。动作空间是未测定的检测项的幂集：

$$A(\mathcal{M}) = \mathcal{P}\big(\{1, \ldots, 6\} \setminus \mathcal{M}\big).$$

转移是确定性的：执行 $a \in A(\mathcal{M})$ 会得到 $\mathcal{M} \leftarrow \mathcal{M} \cup a$。

[1] 在我们的实验中，我们将 (μ_a, σ_a) 视为生成样本的经验均值/尺度；截断会对理论矩造成轻微扰动。

精确条件方差及其约简。写作 $g = \beta_{\mathcal{M}}^{\top}\mathbf{y}_{\mathcal{M}} + \beta_{\mathcal{U}}^{\top}\mathbf{y}_{\mathcal{U}} + \epsilon$。在给定已观测到的测量值 $\mathbf{y}_{\mathcal{M}}$ 的条件下，

$$\mathrm{Var}(g \mid \mathbf{y}_{\mathcal{M}}) = \mathrm{Var}\big(\beta_{\mathcal{U}}^{\top}\mathbf{y}_{\mathcal{U}} + \epsilon\big) = \beta_{\mathcal{U}}^{\top}\Sigma_{\mathcal{U}}\beta_{\mathcal{U}} + \sigma_{\epsilon}^{2},$$

由于$\mathbf{y}_{\mathcal{U}}$与$\mathbf{y}_{\mathcal{M}}$和 ϵ相互独立。因此，通过测量批次 $a \subseteq \mathcal{U}$所实现的精确的不确定性降低计算如下：

$$\begin{aligned}
\Delta\sigma_a^2 &= \mathrm{Var}(g \mid \mathbf{y}_{\mathcal{M}}) - \mathrm{Var}(g \mid \mathbf{y}_{\mathcal{M}\cup a}) && (22)\\
&= \big(\beta_{\mathcal{U}}^{\top}\Sigma_{\mathcal{U}}\beta_{\mathcal{U}} + \sigma_{\epsilon}^{2}\big) - \Big(\beta_{\mathcal{U}\setminus a}^{\top}\Sigma_{\mathcal{U}\setminus a}\beta_{\mathcal{U}\setminus a} + \sigma_{\epsilon}^{2}\Big) && (23)\\
&= \beta_{\mathcal{U}}^{\top}\Sigma_{\mathcal{U}}\beta_{\mathcal{U}} - \beta_{\mathcal{U}\setminus a}^{\top}\Sigma_{\mathcal{U}\setminus a}\beta_{\mathcal{U}\setminus a} && (24)\\
&= \sum_{k\in a}\beta_k^2\,\sigma_k^2. && (25)
\end{aligned}$$

最后一个等式成立，是因为 Σ 是对角矩阵，且每项被测量检测 k 的贡献恰好为 $\beta_k^2\sigma_k^2$。

成本与奖励。设每项检测成本为

$$c_1 = 1.0,\quad c_2 = 1.2,\quad c_3 = 1.5,\quad c_4 = 1.8,\quad c_5 = 2.0,\quad c_6 = 2.2,$$

以及（可选的）终端目标测量成本 $c_{\text{target}} = 10.0$。批次成本为$c_a = \sum_{k\in a} c_k$，即时奖励为单位成本对应的不确定性降低：

$$R(\mathcal{M}, a) = \frac{\Delta\sigma_a^2}{c_a}.$$

贝尔曼递推。设折扣因子为 $\gamma = 0.95$,

$$V(\mathcal{M}) \;=\; \max_{a\in A(\mathcal{M})}\Big\{\, R(\mathcal{M}, a) \;+\; \gamma\, V(\mathcal{M}\cup a)\,\Big\},$$

初始化为 $V\,(\{1,\ldots,6\})\; = 0$（没有剩余不确定性，也没有剩余动作）。我们迭代至误差容限 10^{-6} 或达到 1000 次迭代，以获得最优策略 $\pi_{\text{Theo}}\,(\mathcal{M})$。

关于最优性的说明。 由于转移是确定性的，且奖励与折扣因子相加，该递推在合成不确定性模型下给出了精确的最优策略。该策略作为我们比较基于相似性的规划器时所采用的真实基线。

D.3 基于确定性相似度的 VI（VI-Sim）：完整推导

概述。 VI-Sim 表示一种确定性规划器，它采用与 IBMDP 相同的基于相似性的方差估计，但使用价值迭代而不是随机树搜索。该基线通过使用相同的隐式模型、但采用确定性优化，分离出了 IBMDP 集成规划方法的贡献。

相似度权重和距离。 在状态 $s = (x_\star,\ \mathcal{M})$ 下（保持我们一致的记号；此处我们使用 s 表示一般状态，而不是用 s_t 表示特定的时间步骤），定义：

$$w_i(s) = \frac{\exp\{-\lambda_w\, d(s, D_i)\}}{\sum_{j=1}^{N}\exp\{-\lambda_w\, d(s, D_j)\}}, \tag{26}$$

$$d(s, D_i) = \sum_{k\in\mathcal{K}(s)}\lambda_k\,\frac{\big(\phi_k(s) - \phi_k(D_i)\big)^2}{\sigma_k^2}. \tag{27}$$

其中， $\mathcal{K}(s)$为已知特征（初始定量构效关系及任何已测定检测项）； $\phi_k(\cdot)$提取特征 k； λ_k 为特征权重（默认值为 $= 1$）， $\lambda_w > 0$ 为带宽， σ_k^2 为 $\mathcal{D}$上的经验方差。目标值 $\{g_i\}$ 绝不会用于 $d(\cdot,\ \cdot)$。

加权 g-均值　　和方差（在 I_g上重新归一化）。定义重新归一化权重和加权均值　　：

$$\tilde{w}_i(s) = \frac{w_i(s)}{\sum_{\ell \in I_g} w_\ell(s)} \quad \text{for } i \in I_g, \tag{28}$$

$$\bar{g}(s) = \sum_{i \in I_g} \tilde{w}_i(s)\, g_i. \tag{29}$$

状态 s处估计的条件方差为：

$$\widehat{\sigma}^2_{\text{cond}}(\mathcal{M}) = \sum_{i \in I_g} \tilde{w}_i(s) \left(g_i - \bar{g}(s)\right)^2. \tag{30}$$

执行批次 a 后，我们将观测状态更新为 $s' = (x_\star,\ \mathcal{M} \cup a)$，根据扩展后的特征集重新计算权重，并得到 $\widehat{\sigma}^2_{\text{cond}}(\mathcal{M} \cup a)$。

估计的方差缩减量和奖励。使用相似度加权方差估计，我们定义方差缩减量和奖励：

$$\Delta\widehat{\sigma}^2_a = \widehat{\sigma}^2_{\text{cond}}(\mathcal{M}) - \widehat{\sigma}^2_{\text{cond}}(\mathcal{M} \cup a), \tag{31}$$

$$R(\mathcal{M}, a) = \frac{\Delta\widehat{\sigma}^2_a}{c_a}. \tag{32}$$

然后，我们将此 R 应用于与 D.2节相同的确定性变分推断递归，以获得策略 π_{Sim}。

D.4 IBMDP：隐式后验预测模型与规划细节

概述。IBMDP 在基于相似度的方法上引入了两项关键创新：（1）一种随机后验预测模型，按照相似度对历史案例加权后进行采样；（2）集成蒙特卡洛树搜索规划，用于探索多条策略轨迹。这种组合使规划即使面对隐式模型固有的不确定性，也能保持稳健。

潜在索引视角与似然。引入一个离散潜在索引 $Z \in \{1, \ldots, N\}$，用于索引历史案例 D_i。给定 $Z = i$ 并选择检测 a时，高斯差异模型可得：

$$p(y_a \mid Z = i, a) \propto \exp\!\left(-\frac{\lambda_a (y_a - y_{i,a})^2}{2\sigma_a^2}\right), \tag{33}$$

其中每项检测的权重为 $\lambda_a > 0$。对于批次 A_t，并假设在给定 Z 的条件下各项检测相互独立（关于这一假设的讨论见 A节）：

$$p(\mathbf{y}_{A_t} \mid Z = i, A_t) \propto \prod_{a \in A_t} \exp\!\left(-\frac{\lambda_a (y_a - y_{i,a})^2}{2\sigma_a^2}\right). \tag{34}$$

作为（经温度调节的）后验分布的权重与增量递归。令 O_t 表示截至时间t已观测检测的集合。定义方差归一化距离

$$d(s_t, D_i) = \sum_{(a, y_a) \in O_t} \frac{\lambda_a}{\sigma_a^2} (y_a - y_{i,a})^2.$$

在 Z 上采用均匀先验，温度为 λ_w时，相似度权重等于温度化后验分布

$$w_i(s_t) = \frac{\exp\{-\lambda_w\, d(s_t, D_i)\}}{\sum_{j=1}^{N} \exp\{-\lambda_w\, d(s_t, D_j)\}}.$$

如果随后测量检测 a_t 并观察到 y_{a_t}，距离将以加法方式更新：

$$d(s_{t+1}, D_i) = d(s_t, D_i) + \frac{\lambda_{a_t}}{\sigma_{a_t}^2} (y_{a_t} - y_{i,a_t})^2. \tag{35}$$

由此得到乘法权重更新：

$$w_i(s_{t+1}) \propto w_i(s_t) \cdot \exp\left\{ -\lambda_w \frac{\lambda_{a_t}}{\sigma_{a_t}^2} (y_{a_t} - y_{i,a_t})^2 \right\} \tag{36}$$

$$\propto w_i(s_t) \cdot \left[p(y_{a_t} \mid Z = i, a_t) \right]^{2\lambda_w}, \quad \text{where } \lambda_w = \beta/2, \tag{37}$$

随后进行归一化。因此，该重新加权过程是一种（温度化的）贝叶斯信念更新。

后验预测分布（隐式转移）。对 Z 进行边缘化后，可得到下一步信息状态的后验预测分布：

$$P(s_{t+1} \mid s_t, A_t) = \sum_{i=1}^{N} w_i(s_t) \cdot \delta\Big(s_{t+1} = s_t \oplus \{(a, y_{i,a})\}_{a \in A_t} \Big), \tag{38}$$

在实际操作中，通过采样一个历史案例 $i \sim \mathsf{Cat}(\{w_k(s_t)\})$，并将批次结果 $\{y_{i,a}\}_{a \in A_t}$ 复制到候选对象中来实现——从而保留采样所得历史特征中的跨检测相关性。

使用 MCTS-DPW 和集成方法进行规划。在每次 MCTS rollout 中，我们根据上述后验预测分布生成随机的下一状态，累计步骤成本 $R(s_t, A_t)$，并在违反可行性约束时施加惩罚（例如，任意步骤中出现 $L(s) < \tau$），直至到达终止状态（例如，$H(s) \le \epsilon$）或时域 T。为减少随机采样和树搜索带来的方差，我们运行 N_e 棵相互独立的树，并通过多数投票汇总建议；在每个决策步骤中，同时报告前1动作和前*2*动作集合，从而形成 MLASP。

D.5 示例：相似度权重的演变

目的。 这个简化示例展示了随着证据不断积累，相似度权重如何演变，从而帮助理解 IBMDP 隐式动力学的自适应特性。

设置。 三个历史记录各包含一个特征，其取值为 $\{0, 1, 2\}$；令 $\sigma^2 = 1$、 $\lambda_w = \lambda_1 = 1$。初始候选状态为 $s^{(0)}$，其取值为 **1**。**1**。

步骤 0（初始）。 原始权重：

$$w_1 = e^{-(1.1-0)^2} = e^{-1.21} = 0.297, \tag{39}$$

$$w_2 = e^{-(1.1-1)^2} = e^{-0.01} = 0.990, \tag{40}$$

$$w_3 = e^{-(1.1-2)^2} = e^{-0.81} = 0.445. \tag{41}$$

归一化后 $Z = 0.297 + 0.990 + 0.445 = 1.732$，得到：

$$\tilde{w} = (0.171,\, 0.571,\, 0.257). \tag{42}$$

步骤 1（动作将状态移动到 1.6 之后）。原始权重：

$$w_1 = e^{-(1.6-0)^2} = e^{-2.56} = 0.077, \tag{43}$$

$$w_2 = e^{-(1.6-1)^2} = e^{-0.36} = 0.697, \tag{44}$$

$$w_3 = e^{-(1.6-2)^2} = e^{-0.16} = 0.852. \tag{45}$$

使用 $Z' = 1.626$ 进行归一化得到：

$$\tilde{w} = (0.047,\, 0.429,\, 0.524). \tag{46}$$

随着证据向右移动，后验分布逐渐向位置 2 的历史记录偏移。

D.6 理论分析：VI-Sim 的一致性

定理（基于相似性的估计的一致性）。在特征相互独立的合成线性模型下，VI-Sim 使用的基于相似度的方差估计量 $\widehat{\sigma}^2_{\text{cond}}(\mathcal{M})$ 依概率收敛到 VI-Theo 使用的精确条件方差 $\sigma^2_{\text{cond}}(\mathcal{M})$，即 $\widehat{\sigma}^2_{\text{cond}}(\mathcal{M}) \xrightarrow{P} \sigma^2_{\text{cond}}(\mathcal{M})$ 当 $N \to \infty$ 时。

证明。

假设。(A1) 特征 $\{y_{i,a}\}$ 在 a 之间相互独立，并且在 i 之间独立同分布，且各自具有有限方差 σ_a^2；(A2) 噪声 ϵ_i 与特征独立，且具有有限方差 σ_ϵ^2；(A3) 权重函数 $w_i(s)$ 仅取决于每条记录的已测量检测 $\mathcal{M}$，以及候选对象在这些检测中的观测值；(A4) 对 I_g 上的权重 $\tilde{w}_i(s)$ 重新归一化后构成一个概率向量；(A5) I_g 随 N 增大，使得 $|I_g| \to \infty$。

步骤 1：设置与符号。固定状态 $s = (x_\star,\ \mathcal{M})$，其已测集合为 $\mathcal{M}$，未测集合为 $\mathcal{U} = \{1, \ldots, M\} \setminus \mathcal{M}$。定义记录 i 的目标：

$$g_i = \beta_{\mathcal{M}}^\top \mathbf{y}_{i,\mathcal{M}} + \beta_{\mathcal{U}}^\top \mathbf{y}_{i,\mathcal{U}} + \epsilon_i.$$

（在模型下的）精确条件方差为：

$$\sigma^2_{\text{cond}}(\mathcal{M}) = \beta_{\mathcal{U}}^\top \Sigma_{\mathcal{U}} \beta_{\mathcal{U}} + \sigma_\epsilon^2.$$

价值迭代-Sim 在 s 处使用的估计量为：

$$\widehat{\sigma}^2_{\text{cond}}(\mathcal{M}) = \sum_{i \in I_g} \tilde{w}_i(s) \left(g_i - \sum_{j \in I_g} \tilde{w}_j(s)\, g_j \right)^2.$$

步骤 2：利用独立性进行分解。我们将每个目标分解为 $g_i = \beta_{\mathcal{M}}^\top \mathbf{y}_{i,\ \mathcal{M}} + \underbrace{\quad z_i \quad}$，其中$\underbrace{\text{已测项}}$ $\underbrace{\text{未测项}}$

$$z_i := \beta_{\mathcal{U}}^\top \mathbf{y}_{i,\mathcal{U}} + \epsilon_i.$$

根据关于独立性的假设 (A1)–(A2)，z_i 与$\mathbf{y}_{i,\ \mathcal{M}}$ 独立，因此也与$\mathbf{y}_{i,\ \mathcal{M}}$的任何可测函数独立，包括 $w_i(s)$ 和 $\tilde{w}_i(s)$。由此可得：

$$\mathbb{E}(z_i \mid \mathbf{y}_{i,\mathcal{M}}) = \mathbb{E}(z_i) = \mathbb{E}(\beta_{\mathcal{U}}^\top \mathbf{y}_{i,\mathcal{U}}) + \mathbb{E}(\epsilon_i) = 0, \tag{47}$$

$$\mathrm{Var}(z_i \mid \mathbf{y}_{i,\mathcal{M}}) = \mathrm{Var}(z_i) = \mathrm{Var}(\beta_{\mathcal{U}}^\top \mathbf{y}_{i,\mathcal{U}}) + \mathrm{Var}(\epsilon_i) \tag{48}$$

$$= \beta_{\mathcal{U}}^\top \Sigma_{\mathcal{U}} \beta_{\mathcal{U}} + \sigma_\epsilon^2. \tag{49}$$

步骤 3：加权方差估计量的分析。定义加权均值：$\bar{g}_w := \sum_{i \in I_g} \tilde{w}_i(s) g_i$, $\bar{y}_{\mathcal{M},\ w} := \sum_{i \in I_g} \tilde{w}_i(s) \beta_{\mathcal{M}}^\top \mathbf{y}_{i,\ \mathcal{M}}$，以及 $\bar{z}_w := \sum_{i \in I_g} \tilde{w}_i(s) z_i$。则方差估计量变为：

$$\widehat{\sigma}^2_{\text{cond}}(\mathcal{M}) = \sum_{i \in I_g} \tilde{w}_i(s) \left(g_i - \bar{g}_w \right)^2 \tag{50}$$

$$= \sum_{i \in I_g} \tilde{w}_i(s) \left((\beta_{\mathcal{M}}^\top \mathbf{y}_{i,\mathcal{M}} + z_i) - (\bar{y}_{\mathcal{M},w} + \bar{z}_w) \right)^2 \tag{51}$$

$$= \sum_{i \in I_g} \tilde{w}_i(s) \left(\beta_{\mathcal{M}}^\top \mathbf{y}_{i,\mathcal{M}} - \bar{y}_{\mathcal{M},w} + z_i - \bar{z}_w \right)^2. \tag{52}$$

在给定整个测量面板 $\{\mathbf{y}_{i,\ \mathcal{M}}\}_{i\in I_g}$（其决定了 $\{\tilde{w}_i(s)\}_{i\in I_g}$）的条件下取期望，并利用$\mathbb{E}(z_i|\mathbf{y}_{i,\ \mathcal{M}}) = 0$在 i之间的独立性：

$$\mathbb{E}\big[\widehat{\sigma}^2_{\text{cond}}(\mathcal{M}) \,\big|\, \{\mathbf{y}_{i,\mathcal{M}}\}\big] = \sum_{i\in I_g} \tilde{w}_i(s) \left(\beta_{\mathcal{M}}^\top \mathbf{y}_{i,\mathcal{M}} - \bar{y}_{\mathcal{M},w}\right)^2 + \sum_{i\in I_g} \tilde{w}_i(s)\, \mathbb{E}\big[(z_i - \bar{z}_w)^2 \,\big|\, \{\mathbf{y}_{i,\mathcal{M}}\}\big]\,.$$

步骤 4：简化第二项。第二个求和项可以简化，因为 $\{z_i\}$ 独立同分布、均值为零，且与权重独立：

$$\begin{aligned}
\mathbb{E}\big[(z_i - \bar{z}_w)^2 \,\big|\, \{\mathbf{y}_{i,\mathcal{M}}\}\big] &= \mathbb{E}[z_i^2] + \mathbb{E}[\bar{z}_w^2] - 2\mathbb{E}[z_i \bar{z}_w] \quad \text{(expanding the square)} && (53)\\
&= \mathrm{Var}(z_i) + \mathrm{Var}(\bar{z}_w) - 2\,\mathrm{Cov}(z_i, \bar{z}_w) && (54)\\
&= \mathrm{Var}(z_i) + \mathrm{Var}(z_i) \sum_{j\in I_g} \tilde{w}_j(s)^2 - 2\tilde{w}_i(s)\mathrm{Var}(z_i) && (55)\\
&= \mathrm{Var}(z_i)\,(1 + \sum_{j\in I_g} \tilde{w}_j(s)^2 - 2\tilde{w}_i(s)), && (56)
\end{aligned}$$

其中，我们利用了 $\mathrm{Var}(\bar{z}_w) = \mathrm{Var}(z_i)\sum_{j\in I_g} \tilde{w}_j(s)^2$（由独立性可得），以及 $\mathrm{Cov}(z_i,\ \bar{z}_w) = \tilde{w}_i(s)\,\mathrm{Var}(z_i)$。因此

$$\begin{aligned}
\sum_{i\in I_g} \tilde{w}_i(s)\, \mathbb{E}\big[(z_i - \bar{z}_w)^2 \,\big|\, \{\mathbf{y}_{i,\mathcal{M}}\}\big] &= \mathrm{Var}(z_i) \sum_{i\in I_g} \tilde{w}_i(s)\,(1 + \sum_{j\in I_g} \tilde{w}_j(s)^2 - 2\tilde{w}_i(s)) && (57)\\
&= \mathrm{Var}(z_i) \Big(\sum_{i\in I_g} \tilde{w}_i(s) + \sum_{i\in I_g} \tilde{w}_i(s) \sum_{j\in I_g} \tilde{w}_j(s)^2 - 2\sum_{i\in I_g} \tilde{w}_i(s)^2\Big) && (58)\\
&= \mathrm{Var}(z_i) \Big(1 + \sum_{j\in I_g} \tilde{w}_j(s)^2 - 2\sum_{i\in I_g} \tilde{w}_i(s)^2\Big) && (59)\\
&= \mathrm{Var}(z_i) \Big(1 - \sum_{i\in I_g} \tilde{w}_i(s)^2\Big). && (60)
\end{aligned}$$

注：在第三行中，我们利用了 $\sum_{i\in I_g} \tilde{w}_i(s) = 1$ 和 $\sum_{i\in I_g} \tilde{w}_i(s) \sum_{j\in I_g} \tilde{w}_j(s)^2 = \sum_{j\in I_g} \tilde{w}_j(s)^2$，因为权重之和为 1。因此

$$\mathbb{E}\big[\widehat{\sigma}^2_{\text{cond}}(\mathcal{M}) \,\big|\, \{\mathbf{y}_{i,\mathcal{M}}\}\big] = \underbrace{\sum_{i\in I_g} \tilde{w}_i(s) \left(\beta_{\mathcal{M}}^\top \mathbf{y}_{i,\mathcal{M}} - \bar{y}_{\mathcal{M},w}\right)^2}_{\text{weighted variance of measured part}} + \mathrm{Var}(z_i) \Big(1 - \sum_{i\in I_g} \tilde{w}_i(s)^2\Big).$$

渐近分析与结论。由 (A

5 $\sum_i \tilde{w}_i\ s\ = 1 \qquad 0 \le \tilde{w}_i\ s\ \le$) 以及权重

1 有界（因为 ( ) 和 ( )）$^{s\ 2} \to 0$，可知当 $|I_g| \to \infty$ 且权重非退化时，$\sum_{i\in I_g} \tilde{w}_i(\ )$ 依概率成立。此外，根据针对随机但关于实测部分可测的权重且具有有限二阶矩的三角阵的（加权）大数定律，第一项依概率收敛于给定实测面板时实测贡献的真实条件方差。然而，精确的 VI-Theo 条件方差并不取决于实测面板（检测之间相互独立），因此

$$\sum_{i\in I_g} \tilde{w}_i(s) \left(\beta_{\mathcal{M}}^\top \mathbf{y}_{i,\mathcal{M}} - \bar{y}_{\mathcal{M},w}\right)^2 \xrightarrow{p} 0.$$

合并后，得到

$$\widehat{\sigma}^2_{\text{cond}}(\mathcal{M}) \xrightarrow{P} \mathrm{Var}(z_i)\,(1 - 0) = \beta_{\mathcal{U}}^\top \Sigma_{\mathcal{U}} \beta_{\mathcal{U}} + \sigma_\epsilon^2 = \sigma^2_{\text{cond}}(\mathcal{M}).$$

因此，在合成线性／独立模型下，VI-Sim 的方差估计量与精确的 VI-Theo 条件方差相一致。

[2]具体而言，如果 $\max_i \tilde{w}_i(s) \xrightarrow{P} 0$，那么 $\sum_i \tilde{w}_i(s)^2 \le \max_i \tilde{w}_i(s) \sum_i \tilde{w}_i(s) \to 0$。此条件要求选择合适的相似性核带宽，使得当 $N \to \infty$ 时，没有任何单个历史案例会主导权重。

启示。 由于在独立性假设下， $\sigma^2_{\text{cond}}(\mathcal{M})$ 关于$\mathbf{y}_{\mathcal{M}}$ 为常数，因此，对已测定检测项采用任何合理的相似度加权，都会得到相同的极限条件方差。在更一般的（相关或非线性）情形下，只要核和带宽满足标准的非参数条件，该估计量的目标值就是 $\text{Var}(g \mid \mathbf{y}_{\mathcal{M}} = \text{候选对象})$；IBMDP 的随机集成还可进一步减轻有限样本偏差/方差。

D.7 实验方案与指标

概述。 我们在 100 次独立试验中，对三种规划方法进行系统比较，重点关注它们与理论最优策略的一致性。

对于 100 次独立试验中的每一次：

(i) 按照 D.1 节生成一个新的合成数据集。

(ii) 计算初始状态下 VI-Theo 的最优首个动作。

(iii) 计算 VI-Sim 推荐的首个动作。

(iv) 使用由多个 MCTS-DPW 规划器组成的集成运行 IBMDP；记录 (a) 前1 动作（集成中出现频率最高的动作），以及 (b) 前2 动作集合（出现频率最高的两个动作）。

我们报告每次试验的三项一致性指标：

- **T1 匹配：** 指示变量，表示 IBMDP 的前1 是否与 VI-Theo 的动作相同。
- **T2 匹配：** 指示变量，表示 VI-Theo 的动作是否出现在 IBMDP 的前2 动作集合中。
- **相似度匹配：** 指示变量，表示 VI-Sim 是否与 VI-Theo 相同。

D.8 实验结果与分析

摘要。表5展示了主要结果。在100次试验中，IBMDP 的前1与 VI-Theo 最优解相匹配的情况有47次；IBMDP 的前2涵盖最优解的情况有66次；VI-Sim 与最优解相匹配的情况有36次。这些结果表明，与确定性相似度规划器相比，随机集成规划器能够找回更多近似等效的高价值动作，从而验证了 IBMDP 集成方法的价值。

表5：100次试验中与理论基准的策略对齐情况。

方法	匹配项	匹配率（%）
IBMDP 前一推荐	47	47.0
IBMDP 前2	66	66.0
VI-Sim	36	36.0

独立试验间的统计一致性。为验证我们方法的统计可重复性，我们在下文展示完整的逐次试验对齐结果。这项详细分析表明，IBMDP 的策略建议在各种问题实例中始终与理论最优值一致，为该方法的稳健性和可靠性提供了经验证据（特征索引对应检测 $a \in \{1, \ldots, 6\}$）。

表6：VI-Theo 与 IBMDP 和 VI-Sim 的逐次试验比较。

Iter	VI-Theo	IBMDP Top-1 特征	T1 匹配	IBMDP Top-2 特征	T2 匹配	VI-Sim	相似度匹配
1	5	{3, 4}	0	{3, 5}	1	3	0
2	5	{3, 4}	0	{3, 5, 6}	1	3	0
3	3	{3, 4}	1	{3, 5}	1	5	0
4	5	{3, 4}	0	{3, 5}	1	3	0
5	3	{3, 4}	1	{3, 5}	1	3	1

续见下页

表6: （续）

Iter	VI-Theo	IBMDP 前1 特征	T1 匹配	IBMDP 前2 特征	T2 匹配	VI-Sim	相似度匹配
6	4	{3, 4}	1	{3, 5}	0	3	0
7	3	{3, 4}	1	{3, 5}	1	3	1
8	5	{3, 4}	0	{3, 5, 6}	1	3	0
9	3	{3, 4}	1	{3, 5}	1	3	1
10	4	{3, 4}	1	{3, 5, 6}	0	6	0
11	4	{3, 4}	1	{3, 5}	0	4	1
12	4	{3, 4}	1	{3, 5}	0	4	1
13	4	{3, 4}	1	{3, 5}	0	6	0
14	4	{3, 4}	1	{3, 5}	0	6	0
15	5	{3, 4}	0	{3, 5}	1	3	0
16	5	{3, 4}	0	{3, 5}	1	3	0
17	4	{3, 4}	1	{3, 5}	0	4	1
18	4	{3, 4}	1	{3, 5, 6}	0	3	0
19	5	{3, 4}	0	{3, 5}	1	3	0
20	5	{3, 4}	0	{3, 5}	1	3	0
21	5	{3, 4}	0	{3, 5}	1	3	0
22	4	{3, 4}	1	{3, 5}	0	4	1
23	5	{3, 4}	0	{3, 5}	1	3	0
24	4	{3, 4}	1	{3, 5}	0	6	0
25	5	{3, 4}	0	{3, 6}	0	5	1
26	5	{3, 4}	0	{3, 5}	1	3	0
27	5	{3, 4}	0	{3, 5}	1	3	0
28	4	{3, 4}	1	{3, 5}	0	3	0
29	4	{3, 4}	1	{3, 5}	0	3	0
30	5	{3, 4}	0	{3, 5}	1	3	0
31	5	{3, 4}	0	{3, 5, 6}	1	3	0
32	4	{3, 4}	1	{3, 5}	0	6	0
33	5	{3, 4}	0	{3, 5}	1	5	1
34	4	{3, 4}	1	{3, 5}	0	4	1
35	4	{3, 4}	1	{3, 5}	0	4	1
36	5	{3, 4}	0	{3, 5}	1	3	0
37	5	{3, 4}	0	{3, 5}	1	3	0
38	5	{3, 4}	0	{3, 5}	1	3	0
39	5	{3, 4}	0	{3, 5}	1	3	0
40	3	{3, 4}	1	{3, 5}	1	3	1
41	4	{3, 4}	1	{3, 5}	0	4	1
42	4	{3, 4, 5}	1	{3, 4, 5}	1	6	0
43	3	{3, 4}	1	{3, 5}	1	3	1
44	5	{3, 4}	0	{3, 5, 6}	1	3	0
45	5	{3, 4}	0	{3, 5}	1	3	0
46	5	{3, 4}	0	{3, 5}	1	3	0
47	3	{3, 4}	1	{3, 5}	1	3	1
48	5	{3, 4}	0	{3, 5}	1	5	1
49	4	{3, 4}	1	{3, 5}	0	3	0
50	5	{3, 4}	0	{3, 5}	1	3	0
51	4	{3, 4}	1	{3, 5, 6}	0	4	1
52	5	{3, 4}	0	{3, 5}	1	3	0
53	5	{3, 4}	0	{3, 5}	1	5	1
54	5	{3, 4}	0	{3, 5}	1	3	0
55	4	{3, 4}	1	{3, 5}	0	3	0
56	5	{3, 4}	0	{3, 5}	1	3	0
57	4	{3, 4}	1	{3, 4, 5}	1	3	0
58	3	{3, 4}	1	{3, 5}	1	3	1
59	4	{3, 4}	1	{3, 5}	0	6	0
60	3	{3, 4}	1	{3, 4, 5}	1	3	1
61	3	{3, 4}	1	{3, 5, 6}	1	3	1
62	6	{3, 4}	0	{3, 5}	0	3	0

续下页

表6：（续）

Iter	VI-Theo	IBMDP 前1 特征	T1 匹配	IBMDP 前2 特征	T2 匹配	VI-Sim	相似度匹配
63	5	{3, 4}	0	{3, 5}	1	5	1
64	4	{3, 4}	1	{3, 5}	0	3	0
65	5	{3, 4}	0	{3, 5}	1	3	0
66	5	{3, 4}	0	{3, 5}	1	3	0
67	5	{3, 4}	0	{3, 5}	1	3	0
68	4	{3, 4}	1	{3, 5}	0	6	0
69	6	{3, 4}	0	{3, 5}	0	6	1
70	4	{3, 4}	1	{3, 5}	0	4	1
71	5	{3, 4}	0	{3, 5}	1	3	0
72	5	{3, 4}	0	{3, 5}	1	3	0
73	4	{3, 4}	1	{3, 5}	0	6	0
74	5	{3, 4}	0	{3, 5}	1	3	0
75	5	{3, 4}	0	{3, 5}	1	5	1
76	5	{3, 4}	0	{3, 5}	1	5	1
77	5	{3, 4}	0	{3, 5}	1	3	0
78	4	{3, 4}	1	{3, 5}	0	4	1
79	5	{3, 4}	0	{3, 5}	1	3	0
80	5	{3, 4}	0	{3, 5}	1	5	1
81	5	{3, 4}	0	{3, 5}	1	3	0
82	5	{3, 4}	0	{3, 5}	1	3	0
83	4	{3, 4}	1	{3, 5}	0	4	1
84	5	{3, 4}	0	{3, 5}	1	3	0
85	5	{3, 4}	0	{3, 5}	1	3	0
86	5	{3, 4}	0	{3, 5}	1	3	0
87	4	{3, 4}	1	{3, 5, 6}	0	3	0
88	5	{3, 4}	0	{3, 5}	1	3	0
89	5	{3, 4}	0	{3, 5}	1	3	0
90	5	{3, 4}	0	{3, 5}	1	3	0
91	3	{3, 4}	1	{3, 5}	1	3	1
92	4	{3, 4}	1	{3, 5}	0	4	1
93	3	{3, 4}	1	{3, 5}	1	3	1
94	3	{3, 4}	1	{3, 5}	1	3	1
95	4	{3, 4}	1	{3, 5, 6}	0	3	0
96	5	{3, 4}	0	{3, 5}	1	5	1
97	5	{3, 4}	0	{3, 5}	1	3	0
98	4	{3, 4}	1	{3, 5}	0	4	1
99	3	{3, 4}	1	{3, 5}	1	3	1
100	4	{3, 4}	1	{3, 5}	0	6	0

解读。 VI-Theo 和 VI-Sim 为每个状态返回一个确定性的动作。IBMDP 通过随机滚动模拟探索后验预测策略空间，并通过集成揭示多个价值相近的高价值选项。其 Top-2 覆盖率更高（66%，而 VI-Sim 的匹配率为 36% ），这反映出其策略空间探索能力更强，也更能抵御有限样本效应。

E 使用公开数据集进行基准测试

E.1 高成本差异化清除优化

我们复用一个公开的药代动力学数据集（大鼠、犬和人类的清除率以及定量构效关系预测因子），以在检测成本差异较大的情况下对 IBMDP 进行压力测试。该数据集由 [Yang 等人 2025] 描述，并可供下载。规划器每个决策步骤最多可提出两项检测；成本高昂的人体清除率检测与大鼠和犬的检测同等对待（即可以安排在任意批次）。操作目标是在尽可能降低支出的同时，使最终的人体清除率超过 1.0 mL/min/kg。物种特异性成本列于表7。

表7：高成本差异实验的检测成本结构

检测	成本（美元）	相对成本
大鼠清除率	400	1.0×
犬清除率	800	2.0×
人体清除率	4,000	10.0×

不同于传统门控式推进（例如“先大鼠、再犬、最后人类”），每个情景都从未测量状态 $s_0 = \{\mathrm{CL}^{\mathrm{pred}}{}_{\text{大鼠}}\ ,\ \mathrm{CL}^{\mathrm{pred}}{}_{\text{犬}}\ ,\mathrm{CL}^{\mathrm{pred}}{}_{\text{人类}}\}$ 开始，因此求解器可以选择任何符合条件的批次。IBMDP 集成（30 次运行，$c = 5.0$，每次运行 5,000 次迭代，$\tau \in \{0.6, 0.9\}$）通过对推荐检测批次进行多数投票，生成最大似然动作集路径（MLASP）。投票统计显示出三种状态：（i）高不确定性状态倾向于在确定进行人类检测之前先开展大鼠/犬检测；（ii）低不确定性状态直接跳转至人体清除率；（iii）中等状态则根据该信念阈值 τ 改变行为。图5展示了由此产生的帕累托前沿，并突出显示 MLASP 如何在总支出与终端不确定性之间进行权衡。

E.2 解读帕累托前沿

图5汇总了在每步两项测定约束下，容差为 $\tau \in \{0.0, 0.1, \ldots, 1.0\}$ 时的规划结果。对于每个容差，我们执行一个包含 30 次运行的集成，并记录每次运行提出的首个测定批次。因此，每个标记都代表一条规则：“如果需要 $H(s_T) \le \tau$，则从批次 A_0 开始”；横轴表示相应的检测支出（大鼠 + 、犬 + 、人类 τ），纵轴表示目标不确定性 τ。蓝色轨迹连接帕累托有效点，揭示了收紧 τ 时出现的支出与不确定性之间的权衡。星形标记表示最大似然动作集路径（MLASP）——对于图示容差，在集成成员中出现频率最高的批次。从较宽松到较严格的 H 逐步推进可以看出，宽松的容差倾向于选择成本较低的大鼠/犬检测，而诸如 $s_T(\le 0.10)$ $10\times$ 这样的严格要求最终会迫使人们采用人体清除率检测，尽管它在表7中的成本 。执行首个批次后，IBMDP策略会更新信念状态并重新计算下一步动作，因此该图展示的是初始决策，而完整策略仍会根据情况自适应调整。

F 全局符号参考

本附录为全文使用的所有数学符号提供了全面的参考。下表按类别整理了各符号，方便查阅。

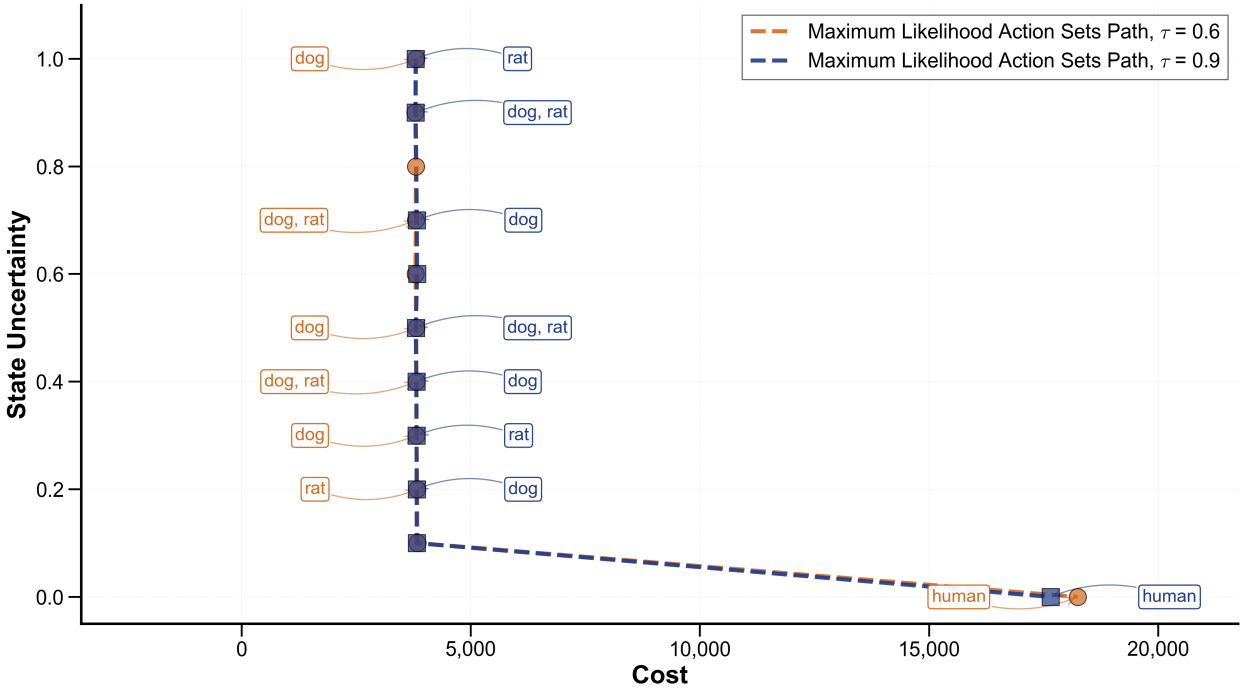

图5：ADME 清除率优化结果，比较 IBMDP 在公开的中枢神经系统清除率基准中某个代表性化合物上的表现，分别采用两种信念阈值（$\tau = 0.6$和 $\tau = 0.9$）。该图展示了 IBMDP 集成在 30 次运行中实现的帕累托最优权衡：总检测支出（横轴）与终端状态不确定性 $H(s_T)$（纵轴）。$\tau = 0.6$（较宽松）和 $\tau = 0.9$（较严格）对应的两条不同曲线表明，更严格的信念阈值需要更高的检测费用，才能实现更低的不确定性。值得注意的是，两种 tau 配置的帕累托前沿高度一致，这证实 IBMDP 在不同置信度要求下能够产生一致且稳健的规划策略。图中标出了每种阈值对应的最大似然动作集路径（MLASP），展示了集成共识如何在人体清除率检测的高昂费用（4,000 美元）与将决策不确定性降至指定阈值以下的需求之间进行权衡。

表 8：全局符号参考，汇总了全文中使用的符号 .

符号	含义
集合与索引	
N, M	历史化合物数量和可用检测总数。
$X = \{x_i\}_{i=1}^{N}$	由具有固定表征的 N 历史化合物组成的集合。
$\mathcal{A} = \{a_1, \ldots, a_M\}$	可用检测集合 M 。
$\mathcal{D} = \{(x_i, \mathbf{y}_i)\}_{i=\ 1}^{N}$	化合物及其测定结果向量的历史数据集。
i, j, k, t	历史案例、检测、特征和决策步骤的索引。
候选化合物与状态	
$x_\star$	制定计划所针对的候选化合物。
$s_t = (x_\star,\ \{y_{\star,j}\}_{j \in M_t})$	步骤 t 时的状态，包括候选对象及截至目前测得的所有结果。
$M_t \subseteq \mathcal{A}$	截至步骤 t，已为 $x_\star$ 测定的检测项集合。
$U_t = \mathcal{A} \setminus M_t$	步骤 t 时 $x_\star$ 尚未测定的检测项集合。
动作、成本与策略	
$\mathcal{A}_t = \mathcal{P}_{\le m}(U_t) \cup \{\text{eox}\}$	s_t处的动作集合：最多包含 m 个未测定的检测项的批次，以及停止动作。
m	每个步骤中可并行执行的检测项数量上限。
$A_t \in \mathcal{A}_t$	在步骤 t 选择的动作（一个检测项批次）。
$c(s_t,\ A_t) \in \mathbb{R}^q_{\ge 0}$	执行动作 A_t 所需的 q 资源成本向量。
$\boldsymbol{\rho} \in \mathbb{R}^q_{\ge 0}$	用于权衡不同成本类型的用户定义的权重。
$R(s_t,\ A_t)$	标量步成本：$\boldsymbol{\rho}^{\mathsf{T}} c(s_t,\ A_t)$。$R(s_t,\ \text{eox}) = 0$。
$\pi, \pi^\star$	将状态映射到动作的策略，以及最优策略。
相似性模型与目标泛函	
g	所关注的主要标量目标属性（例如，体内 终点）。
$G = \{g_i\}$、I_g	历史目标值的集合及其可用的指标集。
$d\ (s_t,\ D_i)$	当前状态与历史案例 i 之间的方差归一化距离。
$w_i\ (s_t)$	给定当前状态 s_t 时，历史案例 i 的相似度权重。
$\tilde{w}_i(s_t)$	相似度权重 $w_i(s_t)$ 在集合 I_g 上重新归一化。
$H(s_t)$	状态不确定性：基于 $\tilde{w}_i(s_t)$ 的目标 g 的加权方差。
$L(s_t)$	目标似然：g 处于理想范围内的加权概率。
$\mathbf{1}[\cdot]$	指示函数：当方括号内的条件成立时返回 1，否则返回 0
	否则。
超参数与约束	
λ_u, λ_k	超参数：全局相似性带宽和逐特征权重。
ϵ, τ	约束目标的阈值：最大终端不确定性和最小目标
	似然 .
γ, T	马尔可夫决策过程的折扣因子和最大时域。
$N_e,\ n_{\text{itr}}$	规划参数：集成规模和每次运行的MCTS迭代。
算法组件	
MCTS-DPW	双重渐进拓宽蒙特卡洛树搜索。
MLASP	最大似然动作集路径：集成多数投票得出的最终规划
	投票。
eox	实验动作结束（停止动作）。