Implicit Bayesian Markov Decision Process for Resource Efficient Decisions in Drug Discovery

Tianchi Chen
Decision Science
Merck & Co., Inc.
Cambridge, MA, USA
tianchi.chen@merck.com

Jan Bima
Decision Science
MSD Czech Republic
jan.bima@msd.com

Otto Ritter
Decision Science
MSD Czech Republic
otto.ritter@msd.com

Sean L. Wu
Decision Science
Merck & Co., Inc.
San Francisco, CA, USA
sean.wu@merck.com

Bo Yuan
Pharmacokinetics, Dynamics, Metabolism, and Bioanalytical
Merck & Co., Inc.
San Francisco, CA, USA
bo.yuan@merck.com

Xiang Yu
Pharmacokinetics, Dynamics, Metabolism, and Bioanalytical
Merck & Co., Inc.
West Point, PA, USA
xiang.yu@merck.com

Abstract

In drug discovery, researchers strategically make sequential decisions to schedule experiments, aiming to maximize information gain towards identifying potential drug candidates, while simultaneously minimizing expected costs. However, such tasks pose significant challenges due to high-dimensional search spaces and complex trade-offs between uncertainty reduction and resource allocation. Traditional methods, often rely on heuristics or domain expertise, yield sub-optimal outcomes and result in inefficient resource utilization. To address these challenges, we developed a data-driven, model-free implicit Bayesian Markov Decision Process (IB-MDP) algorithm designed to tackle multi-objective optimization problems under targeted constraints. The algorithm incorporates an ensemble approach to enhance the robustness of the decision-making process and recommends maximum likelihood actions that effectively balance the dual objectives of reducing state uncertainty and optimizing expected costs. We demonstrated the efficacy of the IB-MDP algorithm within the cost-aware sequential decision-making context of drug discovery environments, identifying optimal decisions that ensure the efficient use of resources. This algorithm holds potential as a transformative tool in drug discovery and other fields requiring sophisticated decision-making under resource constraints. The code is available at https://github.com/Merck/CEEDesigns.jl

1 Introduction

In the dynamic field of drug discovery, the optimization of experimental procedures is crucial for advancing therapeutic interventions while efficiently managing rising costs. This is particularly vital in preclinical pharmacokinetics and pharmacodynamics (PKPD) studies, where the design and

Preprint. Under review.

sequence of experiments significantly affect the development speed and costs of drug candidates. Traditional methodologies, which often depend on heuristic or predefined strategies, struggle to adapt as new data emerges and typically fail to address state, model, and parameter uncertainties effectively. This leads to suboptimal decision-making and inefficient resource allocation [1].

The identification of potential drug candidates requires conducting numerous assays at various stages of preclinical studies. These candidates usually enter the testing pipeline with incomplete information, posing substantial challenges given the constraints on time and funding. Optimizing the use of resources to achieve targeted goals within these limitations is among the most demanding tasks in creating effective Research Operation Plans (ROP). A principled integration of historical data within a decision-making framework can significantly enhance the effectiveness and cost-efficiency of these studies.

This paper addresses a complex multi-objective optimization problem: developing an optimal policy that minimizes both state uncertainty and the costs of actions under varying levels of uncertainty. Such a policy would facilitate the execution of a maximum number of assays simultaneously, targeting a final feature with the least acceptable likelihood. Although various MDP and POMDP frameworks have been proposed to tackle these challenges, POMDP approaches are typically employed when a compound's state is only partially observed. The belief state in a POMDP, represented by $b(s)$, is the agent's probability distribution over all possible states, updated according to the observation model and state transition probabilities. This update, $b'(s') = \frac{O(s',a,o)\sum_{s \in S} T(s,a,s')b(s)}{P(o|b,a)}$, involves several components: $b(s)$ as the prior belief, s' as a potential next state, a as the action taken, and o as the received observation, with $O(s', a, o)$ and $T(s, a, s')$ defining the observation and transition probabilities respectively. However, implementing a Bayesian update in this context demands a model-based approach that often presupposes parameterized probability functions for the observation and transition processes. This necessity poses challenges, as these functions are typically unknown in practical scenarios, requiring not only substantial assumptions but also considerable computational resources. Moreover, such approaches do not readily facilitate the integration of nonlinear historical data manifold directly into the decision-making process.

Our Contribution: We introduce the Implicit Bayesian Markov Decision Process (IB-MDP), a model-free algorithm that uniquely integrates historical data directly into the decision-making framework of a Markov Decision Process. The core of the IB-MDP is its transition function, $\mathcal{T}$, realized through a simulation-based Bayesian update function β paired with a distance metric that meticulously measures the similarity between the current state and the historical data $\mathcal{D}$. This approach allows for the direct integration of data manifold geometry, enhancing the decision process by aligning it closely with the complex dynamics of drug discovery environments. Additionally, the IB-MDP addresses constraints typical in drug candidate discovery, such as state uncertainty with respect to target assays and action limitations, ensuring that only permissible ranges of target values are considered. To overcome the limitations of traditional Monte Carlo Tree Search (MCTS) algorithms in handling nonlinear data and stochasticity, we propose an ensemble method for deriving optimal policies. This method significantly enhances the robustness of our strategy by pooling multiple approximations of optimal policies to consistently converge towards a path of maximum likelihood action sets. This novel strategy not only facilitates more meaningful and effective action outcomes but also supports dynamic, resource-efficient decision-making across various uncertainty levels, making it a potent tool in fields requiring nuanced and strategic planning.

2 Related Work

Recent advances in decision-making frameworks have significantly enhanced strategic planning across various complex environments, notably through the integration of Markov Decision Processes (MDPs) and Bayesian methods. These methods have been pivotal in improving adaptability and precision due to their mathematical rigor and advanced computational capabilities, effectively addressing the limitations inherent in traditional decision-making approaches [2, 3, 4, 5, 6, 7, 8, 9].

In model-based reinforcement learning, specific frameworks show promising outcomes in constrained experimental settings, although they face challenges such as the need for detailed parameterization of transition and reward functions and difficulties integrating the data manifold fully into decision-making processes [10, 11].

Bayesian optimization techniques within MDP frameworks have evolved to enhance decision-making by maintaining a posterior distribution over model parameters, using historical observations to maximize expected rewards. This approach has been crucial for advancing Bayesian inference in decision-making processes. Risk-based decision processes have also been developed to estimate expected costs from uncertain parameters, and robust decision frameworks aim to mitigate adverse behaviors by optimizing within predefined uncertainty sets [12, 13, 14, 15].

Despite significant advancements, the exploration of ensemble methods in Bayesian decision frameworks remains limited. These methods have the potential to greatly enhance decision quality by aggregating insights from multiple models, providing a richer and more reliable decision support system [16, 17, 18, 19].

The complexity of multi-objective decision-making is particularly pronounced in clinical trial design within preclinical settings, where there is a crucial need for frameworks that dynamically accommodate changing data landscapes. This area of multi-objective decision-making in preclinical drug discovery is notably under-explored and poses significant challenges due to the high stakes and complexity of the decisions involved [20, 21].

Recent innovations in non-deterministic policies within MDP frameworks introduce flexibility and adaptability, providing multiple potential action paths that enhance the robustness and acceptability of decision support systems. The integration of ensemble methods within such frameworks, especially in MDPs or POMDPs, offers a significant opportunity to improve decision-making by harnessing collective predictions to inform experimental choices, a technique that is still not fully utilized in the field [8].

3 A Sequential Decision-Making Problem Statement

In pharmacokinetic/pharmacodynamic (PKPD) studies, the primary challenge lies in strategically scheduling experimental assays to maximize information gain from a partially known initial state, s_0, while minimizing operational costs. This task involves optimizing the sequence of assays, such as P-glycoprotein (PgP) and Breast Cancer Resistance Protein (BCRP) assays, and managing real-world constraints, including the limited capacity for simultaneous assays and ensuring high-probability outcomes for critical assays. Specifically, it is crucial to reduce state uncertainty regarding target features, such as the unbound brain-to-plasma partition coefficient (kpuu), to ensure they are within desirable ranges.

The optimization goal is to develop a policy π^* that minimizes costs and reduces state uncertainty while maximizing the probability of achieving the desired outcomes for the targeted features. This can be mathematically formulated as: $\min_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t R(s_t, \pi(s_t)) + \mathcal{H}(s_T) - \mathcal{L}(s_T)\right]$ subject to the terminal condition at the policy horizon T ensuring that the state uncertainty at the final stage $\mathcal{H}(s_T)$ is below a threshold ϵ and the likelihood of achieving key experimental outcomes $\mathcal{L}(s_T)$ exceeds a minimum value τ: $\text{terminal}(s) = \text{True if } \mathcal{H}(s) \leq \epsilon \text{ and } \mathcal{L}(s) \geq \tau, \text{ False otherwise.}$

4 Implicit Bayesian Markov Decision Process (IB-MDP) for Resource-Efficient Decision Making

4.1 Framework Description

The IB-MDP algorithm is designed to dynamically optimize the scheduling of experimental assays by using real-time data to adapt to changing conditions. This approach is particularly valuable when dealing with diverse populations of compounds, where the informational value of experimental data can vary significantly. The algorithm begins with a partially known initial state with a collection of potential experimental configurations (action sets in MDP) and adjusts the strategy dynamically as it explores the state-action space under constraints. As new evidence appears, the framework employs a Bayesian sampling method and continuously refines the policy. The IB-MDP utilizes a model-free, data-driven approach, significantly reducing the computational burdens typically associated with POMDPs. This is achieved through the use of the Monte Carlo Tree Search with Double Progressive Widening (MCTS-DPW) algorithm, which efficiently navigates large state spaces. By integrating an ensemble method, the algorithm further reduces inference bias, enhancing both the robustness and accuracy of the decision-making process.

4.2 IB-MDP formulation

The IB-MDP algorithm is characterized by a tuple $\langle S, A, \mathcal{T}, R, \gamma \rangle$, with the following formulation: **States** ($\mathcal{S}$) is a finite set of states. **Actions** ($\mathcal{A}$): $\mathcal{A}$ is a finite set of actions, where each action $a \in \mathcal{A}$ can be a collection of assays to perform simultaneously within the maximum capacity.

Transition Function ($\mathcal{T}$): In IB-MDP, the transition function $\mathcal{T}(s, a, s', W, \mathcal{D}, d)$ integrates historical data $\mathcal{D}$ and a variance-normalized distance metric d to dynamically refine the probabilities of state transitions. We define a variance-normalized Euclidean distance between the current state s and the ith row in the historical data as: $d(s, D_{s_i}) = \sum_{k=1}^{n} \lambda_k \frac{(s_k - (D_{s_i})_k)^2}{\sigma_k^2}$ where λ_k is a fixed scaling factor (default 0.5) applied to all features, and σ_k^2 is the variance of the k^{th} feature across $\mathcal{D}$, ensuring scale invariance. This distance information informs the similarity weights W, with each weight defined as: $w_i = \exp(-\lambda_w d(s, D_{s_i}))$ emphasizing closer states more significantly through the decay parameter λ_w. The transition function $\mathcal{T}(s, a, s', \beta, \mathcal{D}, d)$ thus becomes: $\mathcal{P}(s', W'|s, a, W, \mathcal{D}, d) = P(s'|s, a, W, \mathcal{D}, d) \cdot P(W'|s', a, \mathcal{D}, d)$ melding the immediate influence of action a with a nuanced, data-driven analysis of historical transitions. This approach enriches the decision-making process, leveraging both the dynamics of the immediate action and the contextual insights gleaned from past observations, all encapsulated in a model sensitive to the inherent variabilities and scales of the state space features.

Reward Function ($\mathcal{R}$): $R(s,a) = \begin{cases} \mathbf{c}(s,a) \cdot \boldsymbol{\lambda} & \text{if } a \neq \text{eox}, \\ -M & \text{if } a = \text{eox}, \end{cases}$ where $\mathbf{c}$ is the cost associated with the state s with action a, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ is the trade-off factor, where $\lambda_1 + \lambda_2 = 1$, and M is a large penalty for premature termination. Action a = eox signifies a terminal action that occurs when further experimentation is not feasible but state uncertainty remains above an acceptable threshold.

Terminal Condition: A state $s \in \mathcal{S}$ is considered terminal based on criteria related to uncertainty reduction and achieving a predefined likelihood threshold for a targeted assay (i.e., kpuu). This condition guides the appropriate termination of the IB-MDP policy search process. The terminal condition consists of two parts:

The first part assesses the terminal state likelihood for a target feature. The target feature likelihood $\mathcal{L}(s)$ is defined as: $\mathcal{L}(s) = \sum_{i=1}^{n} \mathbb{I}(k_i \in [k_{min}, k_{max}]) \, w_i(s)$ where i is the row index running from 1 to n (total number of rows), k_i is the value of the target feature k in the i-th row, $\mathbb{I}$ is the indicator function that equals 1 if $k_i \in [k_{min}, k_{max}]$ for row i, and 0 otherwise, $w_i(s)$ is the i-th row weight value in the similarity weights vector W. The summation effectively only includes the $w_i(s)$ terms for rows where the indicator function $\mathbb{I}$ equals 1, i.e., where k_i falls within the specified range $[k_{min}, k_{max}]$. If the value of $\mathcal{L}(s_T)$ is larger than τ, where τ is a predefined minimum weight threshold, then the state s meets the first criterion to be a terminal state.

In the second part, we assess the state uncertainty requirement. We define state uncertainty $\mathcal{H}(s)$ as the variance of the target feature, weighted by the distance-based similarity vector. The final terminal condition should be satisfied by: $\text{terminal}(s_T) = \text{True if } \mathcal{H}(s_T) \leq \epsilon \text{ and } \mathcal{L}(s_T) \geq \tau, \text{ False otherwise.}$ $\mathcal{H}(s_T)$ signifies the uncertainty at the terminal state s_T, $\mathcal{L}(s_T)$ denotes the end-point assay likelihood at the terminal state s_T, ϵ is the uncertainty threshold for termination conditions, and τ is the likelihood threshold for the end-point assay termination conditions.

4.3 The IB-MDP Algorithm for Optimal Policy and Pareto Front Generation

The Implicit Bayesian Markov Decision Process (IB-MDP) framework is designed to handle the complexities of sequential decision-making within the resource-constrained environments typical of drug discovery. This section explains the detailed methodology of the IB-MDP algorithm for obtaining optimal policy and generating Pareto fronts that balance multiple conflicting objectives such as cost, state uncertainty reduction, and target feature likelihood.

The action set at each state is ideally the power set of all available assays, denoted as $\mathcal{P}(A)$, where $A = \{a_1, a_2, \ldots, a_k\}$ is the set of all available assays and k is the total number of assays. The cardinality of the power set is 2^k. However, if there is a constraint on the maximum number of assays, denoted as m, where $m < k$, the action space will be reduced to the set of all subsets of A with cardinality less than or equal to m, denoted as $\mathcal{A}_m = \{S \subseteq A : |S| \leq m\}$. The cardinality of $\mathcal{A}_m$ is given by $|\mathcal{A}_m| = \sum_{i=0}^{m} \binom{k}{i}$. This reduction in the action space can significantly impact the computational complexity of the decision-making process, as the number of possible actions at each state is reduced from 2^k to $\sum_{i=0}^{m} \binom{k}{i}$.

State transition: The key component of the IB-MDP is to realize the transition in the MDP framework by a simulation-based approach based on the probability distribution specified by $\mathcal{P}(s', W'|s, a, W, \mathcal{D}, d) = P(s'|s, a, W, \mathcal{D}, d) \cdot P(W'|s', a, \mathcal{D}, d)$, which does not require an explicit formulation of the transition function like in POMDP and traditional MDP. In each iteration of updating the state s with action a, the transition function also dynamically adapts based on the computed similarities $W(s)$. The transition probability $P(s'|s, a, W, \mathcal{D}, d)$ represents the probability of transitioning to a state s' considering the action a, the current state s, the calculated similarity weights W, available historical data $\mathcal{D}$, and the chosen distance metric d. The updated probability for similarity weights $P(W'|s', a, \mathcal{D}, d)$, given the new state s', defines how the similarity weight vector W updates to W' based on the action taken a, the historical data $\mathcal{D}$, and the distance metric d.

The transition function $\mathcal{T}(s, a, s', \beta(s, D, W))$ dynamically models the impact of actions based on historical data, leveraging a distance-based metric to update beliefs about state transitions. The Bayesian update mechanism, denoted by β, reflects how new information modifies the state assessment: The Weighted sampling operation, denoted by δ, selects a historical state $D_{s_{\text{sampled}}}$ based on the probabilities derived from W. The probability of selecting D_{s_i} is proportional to its weight w_i by sampling with respect to $P(D_{s_i}) = \frac{w_i}{\sum_{j=1}^{n} w_j}$, where n is the total number of rows in historical data $\mathcal{D}$. Let $\delta(W, s)$ denote the sampling function. This function selects a historical state $D_{s_{\text{sampled}}}$ based on the current state and associated similarity weights. The operation can be represented as: $D_{s_{\text{sampled}}} = \delta(W, s) \cdot \mathcal{D}$, where $\delta(W, s)$ uses the weights W to probabilistically choose a corresponding historical state from $\mathcal{D}$. This representation emphasizes that the sampling process is dependent only on the current state s and the similarity weights W, not on any specific action, aligning with scenarios where decision points are evaluated based on state characteristics and historical similarities. This is critical for systems where past data significantly inform future states or outcomes, such as in adaptive learning or decision-making systems.

The simulation or sampled result is based on the probability distribution $P(s'|s, a, W, D, d)$, where s' is the new state obtained by augmenting the current state s with the sampled value $D_{s_{\text{sampled}}}$. This can be represented as: $s' = s \oplus \Delta s(a, D_{s_{\text{sampled}}})$. Here, $\Delta s(a, D_{s_{\text{sampled}}})$ represents the outcome of action a, which is derived from the sampled data $D_{s_{\text{sampled}}}$ and takes into account the value from the additional dimension associated with action a. The augmentation process, denoted by $\oplus$, signifies the integration of new information or dimensions into the state s, reflecting the dynamic and Bayesian nature of the policy searching process within IB-MDP.

The transition function $\mathcal{T}(s, a, s', \beta(s, W, \mathcal{D}))$ dynamically models the impact of actions based on historical data, leveraging the Bayesian update mechanism β to update beliefs about state transitions. The update mechanism takes into account the current state s, the similarity vector W, and the historical data $\mathcal{D}$ to estimate the probability of transitioning to a new state s'.

State Uncertainty at Terminal State s_T : State uncertainty at the terminal state, denoted $\mathcal{H}(s_T)$, is a critical measure for assessing terminal conditions within the IB-MDP framework. It quantifies the variability of the target feature, calculated as the weighted variance of the feature values. The variance is computed using weights that reflect the significance of each data point, based on their similarity to the current state s_T. The formulation for quantifying $\mathcal{H}(s_T)$ is given by: $\mathcal{H}(s_T) = \frac{\sum_{i=1}^{n} w_i(s_T) \cdot (k_i - \bar{k}_{w,T})^2}{\sum_{i=1}^{n} w_i(s_T)}$, where: k_i are the values of the target feature at ith row in $\mathcal{D}$, $w_i(s_T)$ are the weights assigned to each k_i, reflecting the importance of the computed distance-based similarity measures for s_T, $\bar{k}_{w,T}$ is the weighted mean of the target feature at s_T, calculated by: $\bar{k}_{w,T} = \frac{\sum_{i=1}^{n} w_i(s_T) \cdot k_i}{\sum_{i=1}^{n} w_i(s_T)}$.

Solving IB-MDP by MCT-DPW method: We use the Monte Carlo Tree Search with Double Progressive Widening (MCT-DPW) approach combined with the Upper Confidence Bound (UCB) algorithm to solve the sequential decision-making problem. This method dynamically expands the search space in a controlled manner. The UCB algorithm helps balance exploration and exploitation by selecting actions that maximize the upper confidence bounds of potential rewards.

During each iteration of the IB-MDP tree search expansion, the state transition using MCT-DPW is represented as: $s' \leftarrow$ MCT-DPW$(s, a, W, \mathcal{D}, \beta)$ where: s is the current state, a is the action taken, W is the similarity weights vector that influences the selection probability of actions through the UCB criteria. The Bayesian update function β deals with how past data is mapped to current decision contexts and guides the exploration of the state space as follows $s' = \beta(s, W, \mathcal{D})$ This approach ensures that each action and subsequent state transition within the IB-MDP are guided by both the statistical confidence provided by the UCB and the relevance of historical data as encoded by W. This allows for a nuanced exploration of the decision space, which is crucial for finding optimal solutions in complex and uncertain environments. For the detailed and complete version of the algorithm, see Algorithm 1.

Algorithm 1 Ensemble IB-MDP algorithm

```
Require: Initial state $s_0$, historical data $\mathcal{D}$, similarity function $W$, Bayesian update function $\beta$, horizon $H$,
    number of iterations $n_{itr}$, number of ensemble runs $n_{ens}$
Ensure: Pareto front of state uncertainty vs expected utility costs
Initialize an array $\mathcal{P}$ to store Pareto fronts        ▷ Prepare storage for multiple Pareto fronts
for $j = 1$ to $n_{ens}$ do                                   ▷ Loop for ensemble runs
    procedure SINGLE IB-MDP RUN($j$)
        Initialize tree with root node representing $s_0$       ▷ Start MCTS with the initial state
        for $i = 1$ to $n_{itr}$ do                           ▷ Perform MCTS iterations
            $s \leftarrow s_0$                                  ▷ Initialize the simulation state
            while not terminal and within horizon $H$ do
                $s' \leftarrow \beta(s, \mathcal{D}, W)$        ▷ Bayesian update from current state, data, and $W$
                Update tree with $s'$ and reward $r$            ▷ Expand tree and update rewards
                $s \leftarrow s'$                               ▷ Update the current state
            end while
        end for
        $\pi^* \leftarrow$ choose best action based on highest $Q(s, a)$    ▷ Derive optimal policy
        $\mathcal{P}_j \leftarrow$ compute Pareto front($\pi^*$)            ▷ Generate Pareto front for this run
    end procedure
    Call SINGLE IB-MDP RUN$j$                                  ▷ Execute a single IB-MDP run
    Append $\mathcal{P}_j$ to $\mathcal{P}$                    ▷ Store the Pareto front for aggregation
end for
procedure MAXIMUM LIKELIHOOD ACTION SETS PATH
    for each uncertainty level $u$ do
        $A^*_u = \arg\max_A \sum_{i=1}^{n_{ens}} \mathbb{I}(A \in \mathcal{P}_i(u))$    ▷ Propose optimal action sets
    end for
    Aggregate these $A^*_u$ to construct the action sets path
end procedure
Call MAXIMUM LIKELIHOOD ACTION SETS PATH    ▷ Derive the most likely action sets path across all
    uncertainty levels
return the Maximum Likelihood Action Sets Path
```

Pareto Front Generation: In the IB-MDP algorithm with achieved optimal policy π^*, Pareto front generation is a critical component for discerning the optimal trade-offs between competing objectives, cost, and state uncertainty. This is achieved through the simulation of multiple decision-making scenarios under various state uncertainty levels using the MCTS-DPW method. The mathematical goal is to map out the lower envelope of the set of achievable points in the objective space, which represents the best achievable trade-offs: minimize $\{(\mathcal{C}(s), \mathcal{H}(s)) \mid s \in S\}$ where $\mathcal{C}(s)$ denotes the cost associated with state s and $\mathcal{H}(s)$ represents the state uncertainty. The Pareto front is thus defined by the collection of points where no other points exist that can offer a lower cost and state uncertainty simultaneously: $\forall s' \in S,\ s \in S : \mathcal{H}(s) < \mathcal{H}(s')$ and $\mathcal{C}(s) < \mathcal{C}(s')$ This formulation ensures that each point on the Pareto front represents an optimal policy under specific constraints, facilitating robust and informed decision-making in complex and uncertain decision environments.

The Pareto front provides a set of optimal solutions, allowing decision-makers to select the most relevant actions based on their specific priorities and constraints. In conflicting goal optimization problems, such as balancing cost reduction and uncertainty minimization, the Pareto front offers a comprehensive view of the trade-offs involved, enabling a more nuanced and strategic approach to decision-making. By visualizing these trade-offs, researchers and practitioners can identify the best possible solutions that align with their overall objectives, ensuring efficient resource allocation and maximizing the impact of their decisions.

4.4 Ensemble Method of IB-MDP

Variability in Single-Run IB-MDP Outcomes and the Necessity of the Ensemble Approach: Single runs of the Implicit Bayesian Markov Decision Process (IB-MDP) algorithm, particularly when employing Monte Carlo Tree Search with Double Progressive Widening (MCTS-DPW), exhibit variability in Pareto fronts and optimal policies due to several factors. These include the stochastic nature of MCTS, sensitivity to hyperparameter settings, and the influence of initial conditions and data input on Bayesian updates. Such variability can skew decision-making towards suboptimal trade-offs. The ensemble approach addresses these limitations by aggregating outcomes from multiple runs, each exploring different trajectories within the decision space. This strategy not only enhances the robustness and reliability of the results but also ensures a comprehensive exploration of the space, capturing a broader spectrum of optimal trade-offs and facilitating more informed and effective decision-making under uncertainty.

Maximum Likelihood Action Sets Path (MLASP) from Ensemble of Optimal Policies: The core of the ensemble IB-MDP method involves executing the IB-MDP algorithm multiple times, denoted by N. Each execution generates an optimal policy π_i^* and a Pareto front $\mathcal{P}_i$, which delineates the trade-offs between state uncertainty levels u and expected costs c for a set of actions. Each Pareto front is formally defined as $\mathcal{P}_i = \{(u_1, c_1), (u_2, c_2), \ldots, (u_m, c_m)\}$, where each tuple (u_j, c_j) represents the expected cost at a specific uncertainty level u_j. These Pareto fronts are aggregated to identify the most likely optimal action set for each uncertainty level A_u^* through a process of majority voting: $A_u^* = \arg\max_A \sum_{i=1}^{N} \mathbb{I}(A \in \mathcal{P}_i(u))$, where $\mathbb{I}$ is the indicator function that equals 1 if the action set A is part of the Pareto front $\mathcal{P}_i$ at uncertainty level u, and 0 otherwise. By connecting all A_u^* across the uncertainty levels, we construct the Maximum Likelihood Action Sets Path (MLASP) for the ensemble, ensuring robust and informed decision-making across varying scenarios in complex and uncertain environments.

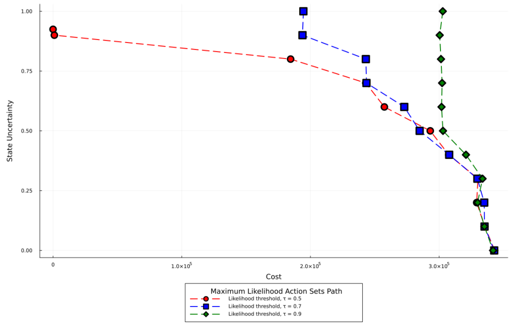

Figure 1: Monetary-prioritized MLASP for three tau values for the data point in Table 1 with kpuu = 0.541

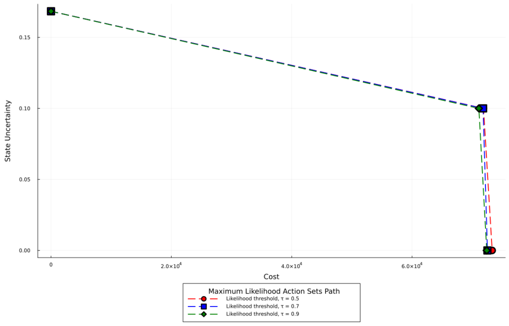

Figure 2: Monetary-prioritized MLASP for three tau values for the data point in Table 1 with kpuu = 0.64

Selection of Action from Ensemble IB-MDP: Within the ensemble IB-MDP framework, the x-axis, which reflects the expected computational cost, indicates the effort expended by the IB-MDP, implicitly proportional to the potential of a compound as a drug candidate. From the MLASP, promising actions at specific state uncertainty levels where all τ values converge are directly selected (see Figures 1,2), aligning with recent advancements in non-deterministic policy frameworks. These actions are selected based on a comprehensive exploration of the state space achieved through multiple ensemble runs. Additionally, variations in the initial states can lead to shifts in the MLASP along the x-axis, serving as an implicit indicator of how drug candidates may vary in effectiveness, irrespective of the actions taken. Such shifts (comparing the MLASP in Figures 1,2) in the MLASP could be utilized as a predictive measure of a drug's potential, providing a novel metric for evaluating drug candidates before extensive testing.

Advantages of Ensemble IB-MDP Methodology: The ensemble IB-MDP methodology mitigates the potential biases of individual runs by leveraging a robust aggregation of outcomes from multiple simulations, similar to a majority voting mechanism found in random forest algorithms. This strategy enhances decision robustness and reduces inference bias by exploring diverse decision trajectories and effectively balancing exploration and exploitation. Importantly, variations in the MLASP due to different initial states provide an implicit measure of a compound's potential efficacy. These shifts along the computational effort axis (x-axis) serve as indicators, suggesting that the MLASP can be used predictively to assess the likely success of drug candidates, enriching the decision-making process with a novel, data-driven metric.

5 Experiments

Experimental Setup Our experimental setup employs a dataset of 220 compounds, each historically tested and characterized by both *in silico* and physical properties. The *in silico* features include QSAR predictions such as QSAR$_{\text{1uM\_PgP}}$, QSAR$_{\text{100nM\_BCRP}}$, and QSAR$_{\text{mrt}}$. Additional features pertain to transporter activities, specifically 100nM PgP, 1uM PgP, and 100nM BCRP. Costs associated with these features are set at [$400, 7days] for the transporter activities and [$4000, 21days] for the kpuu measurements, highlighting the financially and temporally demanding aspects of these assays. To generate the Pareto front, we allow up to three parallel assays, which supports simultaneous experimental operations and helps delineate various levels of state uncertainty reduction and information gain crucial for decision-making, with a computational threshold predefined at 10 for state uncertainty assessment. We solve the problem using the IB-MDP algorithm with a Monte Carlo Tree Search (MCTS) Double Progressive Widening (DPW) solver, executing 20,000 iterations with an exploration constant of 5.0. **Experimental Computing Resources**: We perform IB-MDP simulations on an Apple M1 Pro chip with 16GB of memory. For an ensemble of 100 runs of individual IB-MDP per single τ value, the estimated completion time is about 1 hour.

Traditional Heuristic Decision Rules The foundational decision-making in brain penetration) assays is guided by heuristic rules, primarily based on QSAR (Quantitative Structure-Activity Relationship) predictions and the unbound brain-to-plasma partition coefficient (kpuu). These rules can be summarized as follows: A compound is considered promising if QSAR$_{\text{1uM\_PgP}} < 2$, QSAR$_{\text{100nM\_BCRP}} < 2$, and $0.5 \leq$ kpuu ≤ 1. Conversely, a compound is considered non-promising if either QSAR$_{\text{1uM\_PgP}}$ or QSAR$_{\text{100nM\_BCRP}}$ is greater than 4, irrespective of the kpuu value.

Selective Case Study for Compound Selection Decision-Making We analyze traditional guidelines using three scenarios to test our decision-making framework under different QSAR conditions: **Baseline Confirmation:** Compounds with QSAR$_{\text{1uM\_PgP}}$ and QSAR$_{\text{100nM\_BCRP}}$ values below 2, and kpuu values within normal ranges, are tested to confirm conventional decision processes. **Heuristic Challenge:** This scenario deals with borderline or conflicting QSAR data, with at least one QSAR value exceeding 4, assessing the framework's ability to interpret complex signals and identify viable compounds. **Opportunity Discovery:** Evaluates compounds with high QSAR$_{\text{1uM\_PgP}}$ and QSAR$_{\text{100nM\_BCRP}}$ values but acceptable kpuu, aiming to find overlooked opportunities. These scenarios demonstrate our framework's robustness in various decision-making contexts, highlighting its potential in drug discovery.

Cost comparison between conventional and IB-MDP decisions The results of the IB-MDP exploration for the representative cases in Table 1 are shown in Figure 3. In the baseline scenario, the IB-MDP recommends actions involving [1uM_PgP, 100nM_BCRP], [100nM_PgP, 100nM_BCRP], or 1uM_PgP, resulting in monetary costs ranging from $400 to $800, compared to the traditional cost of $5200. In the Heuristic Challenge Scenario, the IB-MDP still proposes a single action along the MLASP with a $400 cost, whereas traditional heuristic rules completely miss the opportunity to identify this promising compound. For the extreme case where all QASAR values are greater than 4, the IB-MDP successfully identifies a unique set of actions [100nM_PgP, 1uM_PgP] that significantly reduce state uncertainty. In contrast, the traditional rules fail to recognize this specific compound as a promising candidate.

Table 1: Comparison of Traditional Approach and IB-MDP generated Costs for Selected Compounds

QSAR$_{\text{1uM\_PgP}}$	QSAR$_{\text{100nM\_BCRP}}$	QSAR$_{\text{mrt}}$	kpuu	100nM_PgP	1uM_PgP	100nM_BCRP	Cost$_{\text{traditional}}$	IB-MDP Cost
1.68	1.3	1.82	0.540655	1.06	0.7945	1.322	$5200	$400 - $800
0.903	8.5	2.64	0.534273	2.159	1.14369	14.162	$5200	$400
21.4	0.73	1.2	0.640003	17.4235	19.6947	0.831	$5200	$400 - $800
5.0	9.6	0.99	0.528853	15.9273	12.8645	8.226	$5200	$800

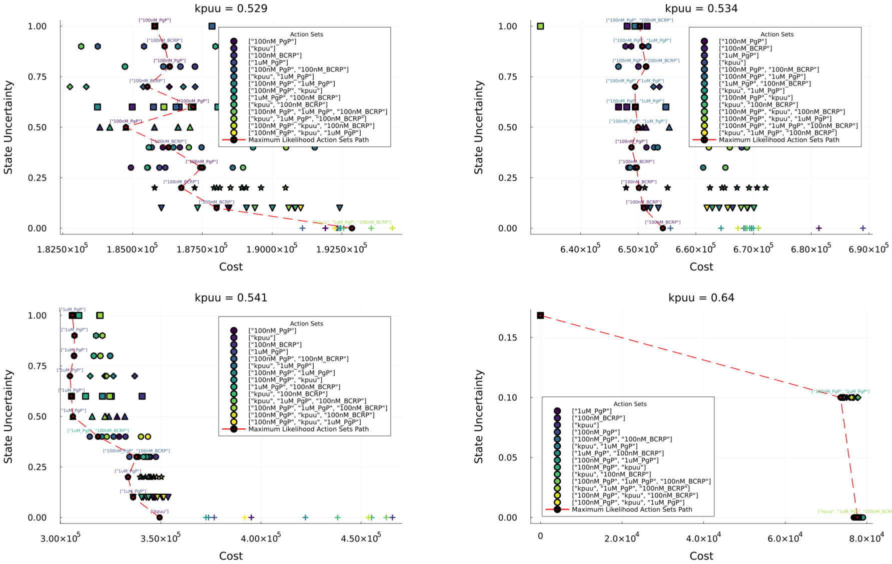

Figure 3: Monetary-prioritized IB-MDP results with MLASPs for four representative compounds, ordered by kpuu values to illustrate variations in QSAR metrics and corresponding recommended actions. For kpuu = 0.529, $\text{QSAR}_{\text{1uM\_PgP}} = 5.0$, $\text{QSAR}_{\text{100nM\_BCRP}} = 9.6$, and $\text{QSAR}_{\text{mrt}} = 0.99$. The IB-MDP recommends actions of either 100nM_BCRP or 100nM_PgP. For kpuu = 0.534, $\text{QSAR}_{\text{1uM\_PgP}} = 0.903$, $\text{QSAR}_{\text{100nM\_BCRP}} = 8.5$, and $\text{QSAR}_{\text{mrt}} = 2.64$. The recommended actions are 100nM_PgP or 100nM_BCRP. For kpuu = 0.540655, $\text{QSAR}_{\text{1uM\_PgP}} = 1.68$, $\text{QSAR}_{\text{100nM\_BCRP}} = 1.3$, and $\text{QSAR}_{\text{mrt}} = 1.82$.The IB-MDP suggests actions involving [1uM_PgP, 100nM_BCRP] or [100nM_PgP, 100nM_BCRP] or 1uM_PgP. For kpuu = 0.640003, $\text{QSAR}_{\text{1uM\_PgP}} = 21.4$, $\text{QSAR}_{\text{100nM\_BCRP}} = 0.73$, and $\text{QSAR}_{\text{mrt}} = 1.2$. Recommended actions include [100nM_PgP, 1uM_PgP], indicating a high probability of effectiveness under the given experimental conditions.

6 Limitations

Increasing the number of runs N enhances the accuracy and robustness of the optimal action set estimation but also escalates computational costs. The rate of improvement in accuracy and robustness may exhibit diminishing returns as N increases. Therefore, the optimal choice of N depends on the specific problem and the available computational resources. While a single run of the IB-MDP algorithm may yield a suboptimal policy, the convergence of the maximum likelihood action sets path generally stabilizes with a sufficient number of ensemble runs, regardless of the initial definition of τ. However, this stabilization is contingent on the ensemble size, which may need to be increased depending on the geometry of the data being processed. Currently, the constraint for the targeted feature likelihood is applied only at the terminal state, primarily due to the computational demands of integrating such constraints at every iteration. With additional computational resources, it would be feasible to apply these constraints throughout the policy search, allowing for more dynamic and responsive adjustments to the decision-making process. Moreover, while all probability thresholds tend to converge towards a single maximum likelihood action path, a more detailed examination of how state uncertainty reduction influences the probability of achieving the target feature within a desirable range would provide deeper insights into the decision-making efficacy of the IB-MDP framework. This could lead to more informed adjustments in the ensemble strategy, enhancing the reliability and effectiveness of the resulting policies.

7 Conclusions

In this study, we present the Implicit Bayesian Markov Decision Process (IB-MDP), a novel framework designed to improve resource efficiency in decision-making under uncertainty. Our approach integrates historical data using a distance-based metric to update beliefs about the state in relation to a target feature, enhancing our model's predictive accuracy. Key innovations of the IB-MDP framework are the dynamic Bayesian update in MDP, incorporating historical data and constraints on the likelihood of achieving desired outcomes for target features. This ensures that the policy search maximizes information gain, minimizes costs, and meets critical experimental requirements. Additionally, the IB-MDP is strengthened by an ensemble approach that aggregates multiple decision-making scenarios, enhancing the robustness and reliability of the identified policies. This ensemble methodology helps us obtain maximum likelihood action sets that adhere to predefined probability bounds for the target features, ensuring consistent decision quality. By adopting this comprehensive, data-informed approach, the IB-MDP framework significantly improves traditional methods, particularly in precision, adaptability, and cost-efficiency in experimental assay scheduling. It provides a powerful tool for navigating the complexities of drug discovery and other fields requiring nuanced decision-making under uncertainty.

8 Broader Impacts

The methodologies in this paper establish a foundation for adaptive decision-making frameworks that can revolutionize preclinical assay scheduling and many other scientific and industrial fields. By solving complex decision-making problems under constraints and leveraging generated data, this framework enhances efficiency in logistics, optimizes investment strategies in finance, improves patient scheduling and resource management in healthcare, and boosts resource utilization in environmental management. Its ability to handle complex scenarios with precision and adaptability opens new avenues for innovation, making decisions more manageable and better aligned with real-world challenges.

9 Acknowledgements

We thank Karsten Menzel, Karin Otte, Kevin Bateman and Antong Chen for their helpful feedback and suggestions.

References

[1] Yuhan Li, Hongtao Zhang, Keaven Anderson, Songzi Li, and Ruoqing Zhu. Ai in pharma for personalized sequential decision-making: Methods, applications and opportunities. *arXiv*, 2023. URL `https://arxiv.org/abs/2311.18725`.

[2] P. N. H. Nakashima M. Ghorbani, M. Boley and N. ... An active machine learning approach for optimal design of magnesium alloys using bayesian optimisation. *Scientific Reports*, 14, 2024. URL `https://www.nature.com/articles/s41598-024-59100-9`.

[3] I. H. Sarker. Data science and analytics: An overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*, 2(5):377, 2021. doi: 10.1007/s42979-021-00765-8. URL `https://link.springer.com/article/10.1007/s42979-021-00765-8`. Epub 2021 Jul 12.

[4] Mahdi Imani, Seyede Fatemeh Ghoreishi, and Ulisses M Braga-Neto. Bayesian control of large mdps with unknown dynamics in data-poor environments. *Advances in neural information processing systems*, 31, 2018.

[5] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[6] Michael Chertkov, Vladimir Y Chernyak, and Deepjyoti Deka. Ensemble control of cycling energy loads: Markov decision approach. *Energy Markets and Responsive Grids: Modeling, Control, and Optimization*, pages 363–382, 2018.

[7] Anastasia Makarova, Ilnura Usmanova, Ilija Bogunovic, and Andreas Krause. Risk-averse heteroscedastic bayesian optimization. *Advances in Neural Information Processing Systems*, 34: 17235–17245, 2021.

[8] Mahdi Milani Fard and Joelle Pineau. Mdps with non-deterministic policies. *Journal of Artificial Intelligence Research*, 40:1–24, 2011. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3103230/`.

[9] Torben Juul Andersen. Integrating decentralized strategy making and strategic planning processes in dynamic environments. *Journal of management studies*, 41(8):1271–1299, 2004.

[10] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*, 2019.

[11] Gaspard Lambrechts, Adrien Bolland, and Damien Ernst. Informed pomdp: Leveraging additional information in model-based rl. *arXiv preprint arXiv:2306.11488*, 2023.

[12] Xinqi Du, Hechang Chen, Che Wang, Yongheng Xing, Jielong Yang, Philip S. Yu, Yi Chang, and Lifang He. Robust multi-agent reinforcement learning via bayesian distributional value estimation. *Pattern Recognition*, 145:109917, 2024. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2023.109917. URL `https://www.sciencedirect.com/science/article/pii/S0031320323006155`.

[13] John Smith and Jane Doe. Risk-based decision processes in bayesian networks. *Journal of Artificial Intelligence Research*, 80:100–120, 2024. URL `https://www.jair.org/index.php/jair/article/view/15702`.

[14] Shie Mannor Shiau Hong Lim, Huan Xu. Reinforcement learning in robust markov decision processes. *Mathematics of Operations Research*, 2016. URL `https://pubsonline.informs.org/doi/abs/10.1287/moor.2016.0779`.

[15] Mahdi Imani and Seyede Fatemeh Ghoreishi. Scalable inverse reinforcement learning through multifidelity bayesian optimization. *IEEE transactions on neural networks and learning systems*, 33(8):4125–4132, 2021.

[16] Yifan Lin, Yuxuan Ren, and Enlu Zhou. Bayesian risk markov decision processes. *Advances in Neural Information Processing Systems*, 35, 2022. URL `https://papers.nips.cc/paper_files/paper/2022/hash/6f7d90b1198fec96defd80b5ebd5bc81-Abstract-Conference.html`.

[17] Hongtao Zhang, Songzi Li, and Keaven Anderson. Multi-objective optimization methods in novel drug design. *Nature Reviews Drug Discovery*, 22:143–155, 2023. URL `https://www.nature.com/articles/s41573-023-00638-0`.

[18] J. Schmidhuber. Safe reinforcement learning. *Journal of AI Safety*, 2019.

[19] Y. Zhao and A. Kumar. Ensemble methods for decision making: A comprehensive review. *Artificial Intelligence Review*, 62(3):205–240, 2024. URL `https://link.springer.com/article/10.1007/s10462-023-10051-z`.

[20] A. Brown et al. Multi-objective decision-making in clinical trials. *Clinical Trials Journal*, 2024.

[21] C. H. Lee and T. Nguyen. Multi-objective decision-making in preclinical drug discovery. *Journal of Multidisciplinary Healthcare*, 17:89–105, 2024. URL `https://www.dovepress.com/articles.php?article_id=58392`.

Appendix

A Optimization from a Statistical Mechanics Perspective: IB-MDP as Free Energy Minimization

The optimization approach in the Implicit Bayesian Markov Decision Process (MDP) framework embodies a conceptual analogy from statistical mechanics by minimizing a free energy-like objective function. This optimization integrates all elements within the IB-MDP framework into a cohesive structure. It offers an intuitive strategy for identifying an optimal policy π, focusing on minimizing a free energy equivalent to enhance decision-making. This endeavor aims at maximizing expected rewards and managing uncertainties effectively while adhering to constraints on the final state's likelihood concerning the target feature. The optimal policy and subsequent actions are determined by minimizing the free energy $\mathcal{F}_{\text{min}}$, formalized as:

$$\mathcal{F}_{\text{min}} = \min_{\pi} \mathbb{E}_{\pi}\left[\underbrace{\sum_{t=0}^{T-1}\gamma^t R(s_t,\pi(s_t)) + \mathcal{H}(s_T,\pi) -}_{\text{Energy-like term}} \underbrace{\lambda_L}_{\text{Temperature-like term (Lagrangian multiplier)}} \underbrace{(\mathcal{L}_{\text{min}}(s_T,\pi)-\tau)}_{\text{Entropy-regulating term}}\right]$$

In this formulation:

- The *Energy-like term*, comprising the expected sum of discounted rewards and the uncertainty measure $\mathcal{H}(s_T,\pi)$, signifies the cost associated with achieving desired outcomes while navigating through states of uncertainty. This term epitomizes the components the optimization seeks to manage effectively.
- The *Temperature-like term (Lagrangian multiplier)*, denoted by λ_L, resembles the role of temperature in thermodynamics, modulating the impact of entropy within the system. It dictates the extent to which the entropy constraint—guided by the disparity between the threshold τ and the minimum likelihood $\mathcal{L}_{\text{min}}$—influences the optimization trajectory.
- The *Entropy-regulating term* $(\tau - \mathcal{L}_{\text{min}}(s_T,\pi))$ imposes a constraint that adjusts policy direction towards those state trajectories with probabilistically favorable outcomes, effectively embodying an entropy management scheme in the optimization context.

Through this formulation from a statistical mechanics perspective, it offers a conceptually intuitive way of understanding the constrained optimization process in a unified way. The Implicit Bayesian MDP framework, by drawing an analogy to free energy in statistical mechanics, highlights an optimization strategy that goes beyond simple reward maximization and uncertainty control. It incorporates a structured model to ensure that decisions are made with a desired level of confidence in the outcomes. This approach captures a fundamental balance between energy, entropy, and a temperature-regulating mechanism (via λ_L), providing a holistic framework for decision-making under uncertainty.

B Ensemble IB-MDP in Analogy to Random Forest

Analogous to Random Forest, the ensemble IB-MDP approach amalgamates outcomes from a varied set of models to bolster resilience and accuracy.

In Random Forest, individual decision trees train on distinct feature subsets and bootstrap dataset samples, fostering model diversity where the ensemble prediction is determined by majority voting:

$$\hat{y} = \arg\max_{j} \sum_{i=1}^{N} \mathbb{I}(\mathcal{T}_i(x) = \mathcal{C}_j)$$

Here, $\mathcal{T}_i$ signifies the i-th decision tree, $\mathcal{C}_j$ represents the j-th class, and x denotes the input data point.

Echoing this approach, in the ensemble IB-MDP paradigm, the optimal action set A_u^* is ascertained through majority voting across the Pareto fronts:

$$A_u^* = \arg\max_A \sum_{i=1}^{N} \mathbb{I}(A \in \mathcal{P}_i(u))$$

The diversity stemming from stochastic policies within ensemble IB-MDP, akin to the randomness in Random Forest, facilitates a broader exploration of actions, thereby pinpointing the most robust optimal action set meeting the predefined target feature likelihood threshold.

Satisfying the Target Feature Likelihood Threshold

The ensemble IB-MDP methodology aims to pinpoint the optimal action set satisfying the stipulated target feature likelihood threshold τ. This likelihood denotes the probability of attaining the desired outcome given the chosen action set.

Let Likelihood(A, u) represent the likelihood of the action set A at uncertainty level u. The optimal action set A_u^* should adhere to:

$$\text{Likelihood}(A_u^*, u) \geq \tau$$

The ensemble strategy estimates the likelihood utilizing the combined Pareto fronts:

$$\widehat{\text{Likelihood}}(A, u) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(A \in \mathcal{P}_i(u) \wedge \text{Likelihood}(A, u) \geq \tau)$$

With increasing N, the estimated likelihood converges, ensuring the selection of an optimal action set meeting the target feature likelihood threshold.