DISTRIBUTED STOCHASTIC CONTEXTUAL BANDITS FOR PROTEIN DRUG INTERACTION

Jiabin Lin[1], *Karuna Anna Sajeevan*[2,3], *Bibek Acharya*[2], *Shana Moothedath*[1], *Ratul Chowdhury*[2,3]

1. Department of Electrical and Computer Engineering
2. Department of Chemical and Biological Engineering
3. The Center for Biorenewable Chemicals
Iowa State University, Ames, IA, USA

ABSTRACT

In recent work [1], we developed a distributed stochastic multi-arm contextual bandit algorithm to learn optimal actions when the contexts are unknown, and M agents work collaboratively under the coordination of a central server to minimize the total regret. In our model, the agents observe only the context distribution and the exact context is unknown to the agents. Such a situation arises, for instance, when the context itself is a noisy measurement or based on a prediction mechanism. By performing a feature vector transformation and by leveraging the UCB algorithm, we proposed a UCB algorithm for stochastic bandits with context distribution. In this paper, we test our algorithm on a real-world dataset and investigate the interactions between drugs and proteins. For this we perform a data pre-processing step to fit the model and we evaluated the performance of our algorithm for the drug-protein interaction study as compared to other benchmark algorithm. Furthermore, we present the results of biological experiments and draw inferences from our findings.

1. INTRODUCTION

Decision-making under uncertainty is an ubiquitous challenge spanning various domains, including control and robotics [2], clinical trials [3, 4, 5], communications [6], and ecology [7]. Learning algorithms have been developed to discern effective policies and strategies for optimal decision-making. Contextual bandits represent one such framework, capturing the sequential decision-making process by leveraging side information, referred to as context [8]. In the typical contextual bandit model, a learner engages with the environment over multiple rounds. During each round, the environment provides a context to the learner, who then selects an action. Following this choice, the learner receives a reward associated with the selected action, with the overarching objective of maximizing cumulative rewards.

While the majority of prior research on Multi-Armed Bandit (MAB) has concentrated on single-agent strategies, the growing significance of distributed learning has spurred exploration into MAB problems involving multiple agents [9]. For instance, in many applications using contextual bandit algorithms, including recommender systems, clinical trials, and control and robotics, collaborative and concurrent learning is desired to expedite the learning process [10, 11]. Additionally, a federated learning framework is often sought after where the agents can collaborate and jointly learn from the data available at multiple agents without sharing the raw data, rather only sharing the estimates thereby maintaining data privacy. This is particularly important in computational biology, where data privacy is a major concern. Another key challenge is that the Contextual Bandit (CB) models considered in the literature typically assume that the contexts are known. However, in many applications there are scenarios where the contexts are noisy or are forecasting measurements (e.g., prediction of physicochemical properties of a drug, weather prediction, or stock market prediction), the exact contexts are unknown, and a distribution on the context is only available [12]. In such cases, the exact context is a sample from this distribution, which is unknown to the learner, and the standard algorithms for contextual bandits are not suitable. In our recent work [1], we studied the distributed stochastic contextual bandit with unknown context and proposed a communication efficient Upper Confidence Bound (UCB) algorithm. The proposed algorithm in [1] achieves an $O(d\sqrt{MT}\log^2 T)$ high probability regret bound, where M, d, T denote the number agents, dimension of feature vector, and number of rounds, when only the context distribution is available and the exact context is unknown.

One particular application where distributed learning under unknown (noisy) data has been encountered is computational and experimental structural biology. In computational and experimental structural biology setups our observations are almost always constrained by our study-objective. In other words, if we seek to measure the binding of a protein with another molecule, we are unlikely to learn any other general property of the protein or the molecule even though that protein or binder molecule will have multiple other chemical properties and functions. One must train focused machine learning models to look at a specific function with data collected on that particular observable. While a gamut of general-purpose latent protein representations of proteins and small molecules have been put forward in the last few years with the advent of language, and physics-informed models (i.e., AlphaFold2 [13], RoseTTAFold [14], OmegaFold [15], AminoBert [16], ProtT5 [17], ProtBert [18], ESM2 [19], and ChemBERT [20]) they are not necessarily guided to understand how a certain drug-protein pair would interact.

Contributions: In this work, we utilize specific binding data from a set of human proteins against a library of drugs (made publicly available by the Harvard Medical School LINCS Center) such that the latent representations of these proteins would likely, selectively reflect how these proteins bind to the said drugs and not any other unrelated biochemical properties. We modeled the protein-drug interaction prediction problem as a bandit learning, where the feature vectors of the drug and the available proteins are extracted from the dataset and then we constructed the latent feature vector corresponding to each drug-protein pair. The goal of the algorithm is to learn the binding relationship between a protein with another molecule and thereby learn to select a protein for a given drug that correspond to maximum binding activity. We implemented and tested the bandit learning algorithm in [1] for the single agent case and for the distributed case (when data is distributed among different labs and the labs collaborate to learn the protein-drug interaction concurrently in a federated manner without sharing the raw data). Subsequently, we thought that would be worthwhile to assess whether the latent representations conform to any structural parameters that define this

drug protein interactions so our d-dimensional feature vector representation of each protein binding to each drug molecule would likely lead us to capture internal structural cues true to each drug-protein pair. One possible intuitive outcome of this experiment would be to check if the d-dimensional vector encoding a protein-drug pair would be able to grasp the top-d possible, docking conformations of the drug against while interacting with its partner protein. In our experiments, we noticed that a reasonably high degree of correlation can be obtained between the normalized latent space representation of the drug-protein pair and the normalized docking energy scores thereby validating the effectiveness of the bandit approach.

1.1. Problem Setting and Notations

Notations: The norm of a vector $z \in \mathbb{R}^d$ with respect to a matrix $V \in \mathbb{R}^{d\times d}$ is defined as $\|z\|_V := \sqrt{z^\top V z}$ and $|z|$ for a vector z denotes element-wise absolute values. Further, $\top$ denotes matrix or vector transpose, $\langle \cdot, \cdot \rangle$ denotes inner product, and $[N] := \{1, 2, \ldots, N\}$, for an integer N.

Problem Setting: Distributed LBs with Unknown Contexts: In this section, we first specify the standard LB problem and then explain the distributed stochastic bandit setting. Let $\mathcal{X}$ be the action set, $\mathcal{C}$ be the context set, and the environment is defined by a fixed and unknown reward function $y : \mathcal{X} \times \mathcal{C} \to \mathbb{R}$. In LBs, at any time $t \in \mathbb{N}$, the agent observes a context $c_t \in \mathcal{C}$ and chooses an action $x_t \in \mathcal{X}$. Each context-action pair (x, c), $x \in \mathcal{X}$ and $c \in \mathcal{C}$, is associated with a feature vector $\phi_{x,c} \in \mathbb{R}^d$, i.e., $\phi_{x_t,c_t} = \phi(x_t, c_t)$. Upon selection of an action x_t, the agent observes a reward $y_t \in \mathbb{R}$ defined by $y_t := \langle \theta^\star, \phi_{x_t,c_t} \rangle + \eta_t$, where $\theta^\star \in \mathbb{R}^d$ is the unknown reward parameter, $\langle \theta^\star, \phi_{x_t,c_t} \rangle = r(x_t, c_t)$ is the expected reward for action x_t at time t, i.e., $r(x_t, c_t) = \mathbb{E}[y_t]$, and η_t is $\sigma-$subGaussian, additive noise. The goal is to choose optimal actions $x_t^\star$ for all $t \in T$ such that the cumulative reward, $\sum_{t=1}^T y_t$, is maximized. This is equivalent to minimizing the cumulative (pseudo)-regret denoted as

$$\mathcal{R}_T = \sum_{t=1}^{T} \langle \theta^\star, \phi^t_{x_t^\star,c_t} \rangle - \sum_{t=1}^{T} \langle \theta^\star, \phi^t_{x_t,c_t} \rangle. \tag{1}$$

Here $x_t^\star$ is the optimal/best action for context c_t and x_t is the action chosen by the agent for context c_t.

In our earlier work [1], we considered *distributed stochastic* LBs with context distribution and unknown contexts. The communication network in [1] consisted of a server and M agents, and the agents can communicate with the server by sending and receiving packets. The context at time t, c_t is *unobservable* rather only a distribution of the context denoted as μ_t is observed by the agents. At round t, the environment chooses a distribution $\mu_t \in \mathcal{P}(\mathcal{C})$ over the context set and samples a context realization $c_t \sim \mu_t$. The agents observe only μ_t and not c_t and each agent selects an action, say action chosen by agent i is $x_{t,i}$, and receive reward $y_{t,i}$, where $y_{t,i} = \langle \theta^\star, \phi_{x_{t,i},c_t} \rangle + \eta_{t,i}$. Our aim is to learn an optimal mapping/policy $\mathcal{P}(\mathcal{C}) \to \mathcal{X}$ of contexts to actions such that the cumulative reward, $\sum_{i=1}^M \sum_{t=1}^T y_{t,i}$ is maximized. Formally, our aim is to minimize the cumulative regret

$$\mathcal{R}(T) = \sum_{i=1}^{M} \sum_{t=1}^{T} \langle \theta^\star, \phi_{x^\star_{t,i},c_t} \rangle - \sum_{i=1}^{M} \sum_{t=1}^{T} \langle \theta^\star, \phi_{x_{t,i},c_t} \rangle. \tag{2}$$

Here, $x_t^\star = \arg\max_{x\in\mathcal{X}} \mathbb{E}_{c\sim\mu_t}[r_{x,c}]$ is the best action provided we know μ_t, but not c_t, and T is the total number of rounds. Such a setting is particularly relevant in scenarios where contexts are noisy and the agents (labs) have limited amount of data and wish to collaborate with other agents (labs) in a federated manner to learn the global objective. Our goal is to develop a distributed multi-armed bandit algorithm with the least possible communication cost to solve this problem. We define the communication cost of a protocol as the number of integers or real numbers communicated between the server and the agents [11]. We make the standard assumptions on the additive noise η_t and the unknown parameter $\theta^\star$ [12] as given below.

Assumptions. Each element η_t of the noise sequence $\{\eta_t\}_{t=1}^\infty$ is conditionally $\sigma-$subGaussian. There exist constants $S, D \geqslant 0$ such that $\|\theta^\star\|_2 \leqslant S$, $\|\phi_{x,c_t}\|_2 \leqslant D$, and $\phi_{x,c_t}^\top \theta^\star \in [0, 1]$, for all t, all $x \in \mathcal{X}$.

1.2. Related Work

Bandit algorithms are well studied in the literature, for a survey see [8] and [21]. The bandit setting closely related to this paper is stochastic contextual bandits, where the learner selects actions based on observed contexts, aiming to learn an optimal mapping from contexts to actions. Linear contextual bandits, a popular variant, has been extensively studied [8] and [21]. and strong theoretical guarantees are established using different solution approaches. In the linear contextual bandit setting, the context is known in each round, making it a special case of the bandit setting addressed in this paper, with μ_t representing a Dirac delta distribution δ_{c_t} for all $t \in T$.

Linear contextual bandits with context uncertainty have been explored in [12, 22, 23]. [22] considered perturbed context scenarios, aiming to compete with an optimal policy using unperturbed feature vectors. Reference [12] tackled contexts that are unobservable, with only a distribution over contexts available, seeking to select the best action based on this distribution. Our prior work [23] considered a single-agent conservative contextual MAB problem where contexts were unknown, and performance constraints were imposed.

In this paper we extend [12] to address a *multi-agent* stochastic contextual MAB problem with unknown context. We transform this into a distributed linear contextual bandit scenario with action-dependent noise, referred to as heteroscedastic bandits, as explored in [24, 25]. Notably, the distinguishing factor between prior heteroscedastic bandit works and our approach is our focus on a multi-agent distributed MAB setting, rather than a single-agent scenario.

Recently, MAB models involving multiple players have garnered increased attention [26]. Our model shares similarities with the distributed bandits examined in [11], where agents encounter the same bandit model and communicate with a central server for collaborative and concurrent learning. Reference [11] considered setting with fixed and time-varying action sets. The time-varying action set setting aligns with our approach, but a key distinction is that, in [11] , contexts are known to the agents, whereas our contexts remain unknown. A related problem variant is addressed in [27], where agents observe the actual context after choosing actions, introducing a delay. For this case, we presented a modified algorithm in [27] utilizing this additional information to achieve a tighter regret bound.

2. DISTRIBUTED UCB ALGORITHM FOR LINEAR BANDITS WITH UNKNOWN CONTEXTS

In this section, we present the algorithm for solving the distributed stochastic contextual bandit when the actual context is not observable rather a distribution is only available (e.g., weather prediction/forecast, stock market prediction, prediction of physicochemical properties of a drug). Our algorithm is built on the works of [11, 12]. The pseudocode of our algorithm is presented in Algorithm 1.

Given the distribution μ_t, we first construct the feature vectors $\Psi_t = \{\psi_{x,\mu_t} : x \in \mathcal{X}\}$, where $\{\psi_{x,\mu_t} := \mathbb{E}_{c\sim\mu_t}[\phi_{x,c}]\}$ is the expected feature vector of action x under μ_t. We use Ψ_t as the feature context set at time t. Our algorithm is based on the *optimism in the face of uncertainty* principle, where at each time $t \in [T]$, each agent $i \in [M]$

maintains a confidence set $\mathcal{B}_{t,i} \subseteq \mathbb{R}^d$ that contains the unknown parameter vector $\theta^\star$ with high probability. Each agent then chooses an optimistic estimate $\tilde{\theta}_{t,i} = \arg\max_{\theta \in \mathcal{B}_{t,i}} (\max_{x \in \mathcal{X}} \ \psi_{x,\mu_t}^\top \theta)$ and chooses an action $x_{t,i} = \arg\max_{x \in \mathcal{X}} \psi_{x,\mu_t}^\top \tilde{\theta}_{t,i}$. Equivalently the agent chooses the pair $(x_{t,i}, \tilde{\theta}_{t,i}) \in \arg \max_{(x,\theta) \in \mathcal{X} \times \mathcal{B}_{t,i}} \psi_{x,\mu_t}^\top \theta$ which jointly maximizes the reward. The agents now play their respective optimistic actions, $x_{t,i}$'s, and receive rewards $y_{t,i}$'s and utilize the reward observations to update their individual confidence set. We note that it is not immediately clear how this is feasible since $y_{t,i}$ is a noisy observation of $\phi_{x_{t,i},c_t}^\top \theta^\star$ and the algorithm expects the reward $\psi_{x_{t,i},\mu_t}^\top \theta^\star$. To address this, we construct a feature set Ψ_t in such a way that $y_{t,i}$ is an unbiased observation for the action choice ψ_t, similar to the technique in [12] for single agent bandits. We denote $\sum_t \psi_{x_{t,i},\mu_t} \psi_{x_{t,i},\mu_t}^\top$ and $\sum_t \psi_{x_{t,i},\mu_t} y_{t,i}$ for each agent $i \in [M]$ as $W_{t,i}$ and $U_{t,i}$, respectively. We construct the confidence set $\mathcal{B}_{t,i}$ using $W_{t,i}$ and $U_{t,i}$ as

$$\mathcal{B}_{t,i} = \Big\{ \theta \in \mathbb{R}^d : \|\hat{\theta}_{t,i} - \theta\|_{\overline{V}_{t,i}} \leqslant \beta_{t,i} \Big\}, \tag{3}$$

where $\beta_{t,i} = \beta_{t,i}(\sigma, \delta) = \sigma \sqrt{2 \log \Big(\frac{\det(\overline{V}_{t,i})^{1/2} \det(\lambda I)^{-1/2}}{\delta} \Big)} + \lambda^{1/2} S$, $\overline{V}_{t,i} = \lambda I + W_{t,i}$, $\hat{\theta}_{t,i} = \overline{V}_{t,i}^{-1} U_{t,i}$.

Without synchronization, agents in our protocol execute Algorithm 1 in [12] separately. In that case the regret will be scaled by a factor M. During synchronization agents share all newly acquired samples with each other. The synchronizations are done at specific time instants. We refer to the timesteps between the two synchronizations as *epochs*. The epochs are designed based on the observation in [28] that the change in the determinant of $\overline{V}_t$ is a good indicator of learning progress. Based on this observation, we only synchronize when agent i finds that the log-determinant of $\overline{V}_{t,i}$ has changed more than a constant factor since the last synchronization, and this reduces the communication cost of the algorithm. The pseudocode of our algorithm is described below. The regret and communication bounds for Algorithm 1 are proved in [1].

Algorithm 1 Distributed UCB for LBs with hidden contexts

Initialization: $B = (\frac{T \log MT}{dM})$, $\lambda = 1$, $W_{\text{syn}} = 0, U_{\text{syn}} = 0, W_{t,i} = 0, U_{t,i} = 0, t_{last} = 0, V_{last} = \lambda I$, for all $i = 1, 2, \ldots, M$
for $t = 1, 2, \ldots, T$ **do**
  Nature chooses $\mu_t \in \mathcal{P}(\mathcal{C})$ and learner observes μ_t
  Set $\Psi_t = \{\psi_{x,\mu_t} : x \in \mathcal{X}\}$ where $\{\psi_{x,\mu_t} := \mathbb{E}_{c \sim \mu_t}[\phi_{x,c}]\}$
  for Agent $i = 1, 2, \ldots, M$, **do**
    $\overline{V}_{t,i} = \lambda I + W_{\text{syn}} + W_{t,i}$, $\hat{\theta}_{t,i} = \overline{V}_{t,i}^{-1}(U_{\text{syn}} + U_{t,i})$
    Construct the confidence ellipsoid $\mathcal{B}_{t,i}$ using $\overline{V}_{t,i}$, $\hat{\theta}_{t,i}$
    $(x_{t,i}, \tilde{\theta}_{t,i}) = \arg\max_{(x,\theta) \in \mathcal{X} \times \mathcal{B}_{t,i}} \langle \psi_{x,\mu_t}, \theta \rangle$
    Play $x_{t,i}$ and get the reward $y_{t,i}$
    Update $W_{t,i} = W_{t,i} + \psi_{x_{t,i},\mu_t} \psi_{x_{t,i},\mu_t}^\top$, $U_{t,i} = U_{t,i} + \psi_{x_{t,i},\mu_t} y_{t,i}$
    $V_{t,i} = \lambda I + W_{\text{syn}} + W_{t,i}$
    if $\log(\det(V_{t,i})/\det(V_{last})) \cdot (t - t_{last}) \geqslant B$ **then**
      Send a synchronization signal to server to start a communication round
    end if
    Synchronization round:
    if a communication round is started **then**
      All agents $i \in [M]$ send $W_{t,i}$ and $U_{t,i}$ to server
      Server computes $W_{\text{syn}} = W_{\text{syn}} + \sum_{i=1}^M W_{t,i}$, $U_{\text{syn}} = U_{\text{syn}} + \sum_{i=1}^M U_{t,i}$
      All agents receive $W_{\text{syn}}, U_{\text{syn}}$ from the server
      Set $W_{t,i} = U_{t,i} = 0$, $t_{last} = t$, for all i, $V_{last} = \lambda I + W_{\text{syn}}$
    end if
  end for
end for

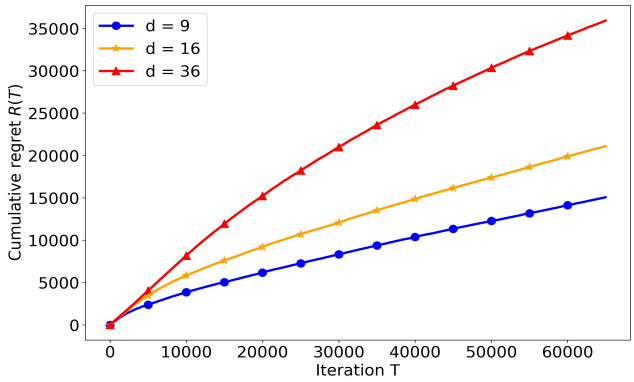
3. EXPERIMENTAL ANALYSIS AND RESULTS

3.1. Bandit Simulations

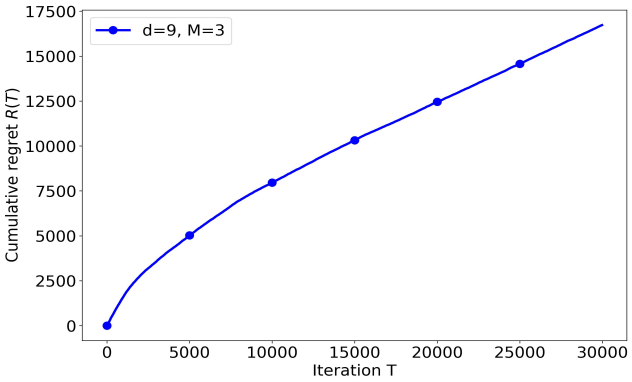
We first describe the data pre-processing process to construct the rating matrix R from the data. The dataset consists of different drugs, proteins, and their corresponding experimental values. In the rating matrix R, drugs are the rows, proteins are the columns, and the experiment value of the i^{th} molecule and j^{th} protein is the (i, j)-th entry of the rating matrix R. In instances where there were multiple experimental values for the same combination of molecule-protein, we chose the lowest value and removed the redundant cases while constructing R. Also in cases with missing experimental values, we set the experimental value as 10001. Our dataset includes 207 molecules and 395 proteins t and the rating matrix $R = [r_{x,c}] \in \mathbb{R}^{207 \times 395}$. Subsequently, we performed a decomposition of R. We then performed the Non-negative Matrix Factorization (NMF) decomposition to decompose R into product of two non-negative matrices $W \in \mathbb{R}^{207 \times \ell}, H \in \mathbb{R}^{\ell \times 395}$.In our simulations, we set $\ell = 3, 4, 6$. In the decomposition, each row of W, $\{W_j^\top\}_{j \in [207]}$, represents a context and each column of H, $\{H_k\}_{k \in [395]}$, represents an action. The feature vector for a given context W_j and action H_k is given by the vectorized form of the matrix $W_j H_k^\top$. Hence the feature vector $\phi(x, c)$ is of dimension ℓ^2. The reward $r(x_t, c_t)$ is bounded above by 1 by normalizing the entries, and the observation noise η_i is set as Gaussian with zero mean and standard deviation 10^{-3}. The plots are shown in Figure 1. Figure 1a presents the regret vs. iteration count (round) plot for $\ell = 3, 4, 6$ (i.e., $d = 9, 16, 36$, respectively) for the single-agent case (i.e., a single agent learning) and Figure 1b presents the distributed setting after distributing the data among 3 agents. For the distributed case, we set $\ell = 3$ (i.e., $d = 9$) and $T = 30,000$.

3.2. Biological Experiments

To facilitate this work, we took the 10-dimensional latent representations of each protein from the previous step. For example, we chose to demonstrate one case where the BRAF protein interacts with a RAF265 drug (Figure 2). We first normalized the 10-dimensional latent vector of BRAF-RAF265 binding and hypothesized it to represent the best binding energy conformations of the drug-protein pair. In parallel, we also performed molecular docking experiments of that same drug with the affiliated protein BRAF (using AutoDock4 program [29]). We gleaned the docking conformations with the objective to identify which normalized binding energies from these conformations are representative of the trend of numbers in the normalized 10-dimensional latent representation of this protein-drug pair. We thus correlated the select ten docking poses which maximized correlation with the 10-dimensional vector of normalized latent representations. We finally validated the rationality of these select drug-binding poses by checking how many of these are actually bound to the known drug-binding pocket of the protein (BRAF, in this case). To facilitate this, we inspected the experimentally determined structure of BRAF co-crystallized with a known inhibitor molecule (CNS292) and reported as PDB accession id: 3Q4C [30]. The location on BRAF where CNS292 is bound represents the expected site of drug binding activity on BRAF. We superimposed our 10 chosen RAF265 docking conformations on BRAF with 3Q4C to check how many of these were at the inhibitor-binding site.



(a) Single agent setting: $M = 1$ and dimension $d = 9, 16, 36$



(b) Multi-agent setting: $M = 3$ and dimension $d = 9$

Fig. 1: The plot showing cumulative regret $R(T)$ with respect to iteration number. We ran 60,000 iterations for Fig. 1a and 30,000 iterations for Fig. 1b.

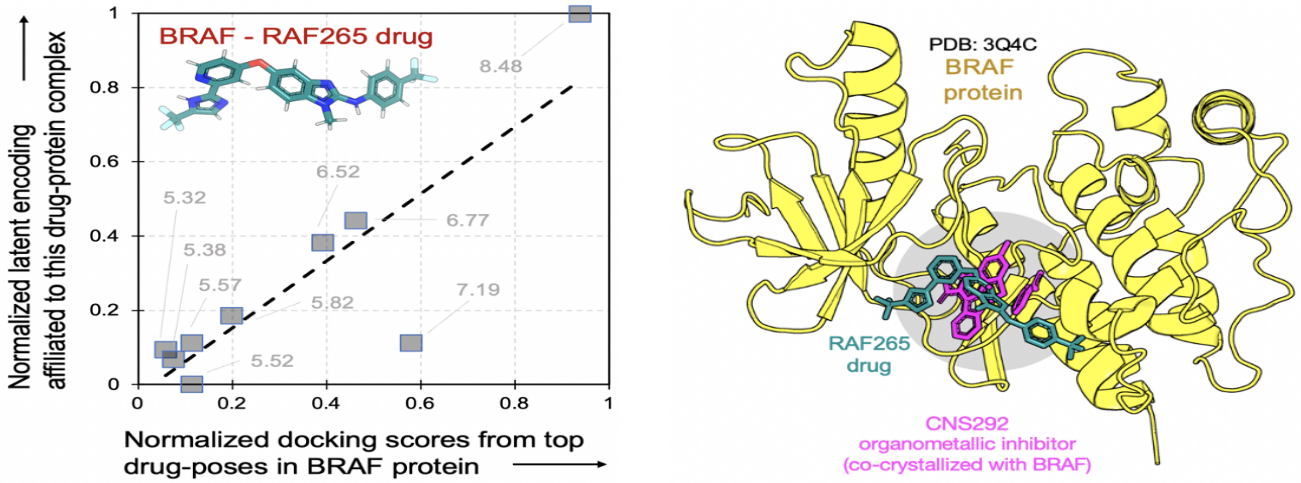


Fig. 2: Normalized drug binding scores from the top AutoDock4-predicted RAF265 drug binding poses with B-Raf proto-oncogene serine/ threonine kinase (abbreviated as BRAF) seem to correlate reasonably well ($R^2 > 0.76$) with the 10-dimensional normalized latent encoding derived from this drug-protein binding data. Individual data points are marked (in gray) with the raw AutoDock4 binding scores (in computational binding units) between BRAF and RAF265. On the right, we show with one representative pose (out of the 10 poses) of RAF265 that the drug bound to the known native drug-binding pocket (as known from experimental BRAF inhibitor binding; co-crystallized PDB 3Q4C).

We observe that a reasonably high degree of correlation (R2 >0.76) can be obtained between the normalized latent space representation of the drug-protein pair (BRAF and RAF265) and the normalized docking energy scores obtained from the ten chosen RAF265 conformations. We further assessed the utility of this correlation by explicitly checking the location of drug binding per conformational pose (recovered from the docking program - AutoDock4). We notice that all these conformations of the RAF265 drug were indeed within the reported pocket of drug activity where the CNS292 inhibitor molecule binds (see Figure 2). While this just a spot check, we believe that a larger campaign of such docking experiments launched against the whole data set would unravel more and more systematic information to map how these latent space dimensions of protein drug finding empirically encodes to drug binding poses at the correct drug-binding location on these proteins.

4. DISCUSSION AND FUTURE WORK

In this paper, we explored the use of bandit learning to study protein-drug interaction. We developed a distributed and federated bandit learning algorithm [1] and showed that it can learn to choose optimal proteins for given drugs to maximize the binding activity. We also performed biological experiments to draw inferences about the feasibility of the bandit learning approach.

While this is an initial analysis for the proof-of-concept of this novel methodology, we in future hope to train separate neural architectures (say, a simple fully connected perceptron) where we would like to create an implicit map between the feature vectors of several hundred thousand of drug-protein combinations. As a start, we will follow up with the dataset which have been used in this study itself, to calculate the structural and biological fidelity of this latent space vector and refine it (if needed). Such efforts would progress the understanding of cryptic binding interactions between different drugs and different proteins and would become a cornerstone for future generative-AI models where one would be able to input a protein molecule's amino acid sequence and the chemical identifier of a drug and internally by predicting the protein structure, generating the latent space vector to predict the probability of a given protein to be affected when an individual (or a living organism) is administered a certain drug. This is consequential in context of disease intervention (where one intends to target a specific protein with a given drug) and side-effect monitoring (where one intends to minimize binding of a drug to any unwanted protein).

5. REFERENCES

[1] Jiabin Lin and Shana Moothedath, "Distributed stochastic bandits with hidden contexts," in *2023 European Control Conference (ECC)*. IEEE, 2023, pp. 1–6.

[2] Vaibhav Srivastava, Paul Reverdy, and Naomi E Leonard, "Surveillance in an abruptly changing world via multiarmed bandits," in *IEEE Conference on Decision and Control (CDC)*, 2014, pp. 692–697.

[3] Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere, "On multi-armed bandit designs for dose-finding clinical trials," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 686–723, 2021.

[4] Yogatheesan Varatharajah and Brent Berry, "A contextual-bandit-based approach for informed decision-making in clinical trials," *Life*, vol. 12, no. 8, pp. 1277, 2022.

[5] Sofía S Villar, Jack Bowden, and James Wason, "Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges," *Statistical science: a review journal of the Institute of Mathematical Statistics*, vol. 30, no. 2, pp. 199, 2015.

[6] Animashree Anandkumar, Nithin Michael, Ao Kevin Tang, and Ananthram Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[7] Vaibhav Srivastava, Paul Reverdy, and Naomi E Leonard, "On optimal foraging and multi-armed bandits," in *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2013, pp. 494–499.

[8] Sébastien Bubeck and Nicolo Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *arXiv preprint arXiv:1204.5721*, 2012.

[9] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard, "Distributed cooperative decision making in multi-agent multi-armed bandits," *Automatica*, vol. 125, pp. 109445, 2021.

[10] Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen, "Federated linear contextual bandits," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[11] Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang, "Distributed bandit learning: Near-optimal regret with efficient communication," *arXiv preprint arXiv:1904.06309*, 2019.

[12] Johannes Kirschner and Andreas Krause, "Stochastic bandits with context distributions," *Advances in Neural Information Processing Systems*, vol. 32, pp. 14113–14122, 2019.

[13] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al., "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[14] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al., "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871–876, 2021.

[15] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al., "High-resolution de novo structure prediction from primary sequence," *BioRxiv*, pp. 2022–07, 2022.

[16] Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, Anant Kharkar, Koushik Roy, Charlotte Rochereau, Gustaf Ahdritz, Joanna Zhang, George M Church, et al., "Single-sequence protein structure prediction using a language model and deep learning," *Nature Biotechnology*, vol. 40, no. 11, pp. 1617–1623, 2022.

[17] Michael Bernhofer, Christian Dallago, Tim Karl, Venkata Satagopam, Michael Heinzinger, Maria Littmann, Tobias Olenyi, Jiajun Qiu, Konstantin Schütze, Guy Yachdav, et al., "Predictprotein-predicting protein structure and function for 29 years," *Nucleic acids research*, vol. 49, no. W1, pp. W535–W540, 2021.

[18] Yaron Geffen, Yanay Ofran, and Ron Unger, "Distilprotbert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts," *Bioinformatics*, vol. 38, no. Supplement_2, pp. ii95–ii98, 2022.

[19] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[20] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar, "Chemberta: large-scale self-supervised pretraining for molecular property prediction," *arXiv:2010.09885*, 2020.

[21] Tor Lattimore and Csaba Szepesvári, *Bandit algorithms*, Cambridge University Press, 2020.

[22] Se-Young Yun, Jun Hyun Nam, Sangwoo Mo, and Jinwoo Shin, "Contextual multi-armed bandits under feature uncertainty," *arXiv preprint arXiv:1703.01347*, 2017.

[23] Jiabin Lin, Xian Yeow Lee, Talukder Jubery, Shana Moothedath, Soumik Sarkar, and Baskar Ganapathysubramanian, "Stochastic conservative contextual linear bandits," *IEEE Conference on Decision and Control*, 2022.

[24] Ping-Chun Hsieh, Xi Liu, Anirban Bhattacharya, and PR Kumar, "Heteroscedastic bandits with reneging.," *arXiv preprint arXiv:1810.12418*, 2018.

[25] Johannes Kirschner and Andreas Krause, "Information directed sampling and bandits with heteroscedastic noise," in *Conference On Learning Theory*, 2018, pp. 358–384.

[26] Xinlei Yi, Xiuxian Li, Tao Yang, Lihua Xie, Tianyou Chai, and Karl Henrik Johansson, "Distributed bandit online convex optimization with time-varying coupled inequality constraints," *IEEE Transactions on Automatic Control*, vol. 66, no. 10, pp. 4620–4635, 2020.

[27] Jiabin Lin and Shana Moothedath, "Distributed stochastic bandit learning with delayed context observatio," *European Control Conference*, 2022.

[28] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári, "Improved algorithms for linear stochastic bandits," *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

[29] Ruth Huey, Garrett M Morris, and Stefano Forli, "Using autodock 4 and autodock vina with autodocktools: a tutorial," *The Scripps Research Institute Molecular Graphics Laboratory*, vol. 10550, no. 92037, pp. 1000, 2012.

[30] Peng Xie, Craig Streu, Jie Qin, Howard Bregman, Nicholas Pagano, Eric Meggers, and Ronen Marmorstein, "The crystal structure of braf in complex with an organoruthenium inhibitor reveals a mechanism for inhibition of an active form of braf kinase," *Biochemistry*, vol. 48, no. 23, pp. 5187–5198, 2009.