

# Bayes Optimal Informer Sets for Drug Discovery\*

Peng Yu, CONTRIBUTING MEMBERS OF UW DRUG DISCOVERY GROUP,  
& Michael A. Newton<sup>†</sup>

June 7, 2020

## Abstract

In early stage drug discovery, there has been significant recent effort in developing compound prioritization methods for a new target from minimal data. One kind of such methods is informer based ranking (IBR) methods, which selects an initial subset of compounds, so-called informer set, to be screened based on original data, and ranks additional compounds for the new target based on this additional screening data and original data. However, the selection of informer set in previous methods is often heuristic, based on chemical diversity or randomized algorithms with no statistical guarantees. In this article, we formalize the informer set ranking procedure as a two-stage decision problem and introduce a new method named as Bayes Optimal Informer SEt (BOISE) to select informer set as well as rank the compounds against the new target. BOISE is a fully Bayesian procedure that selects informer set based on posterior expected loss and ranks compounds based on posterior expectation. It is guaranteed to be Bayes rule under proposed model and loss if computed exactly, and we propose a computationally efficient version of BOISE that achieves balance between ranking performance and computation complexity. Both proposed methods and previous methods are applied on human kinase chemogenomics data set and anti-cancer drug sensitivity data set and evaluated under various ranking criteria. We demonstrate that proposed methods are significantly better than previous methods under all metrics.

**Keywords:** key1, key2, key3

**JEL Codes:** key1, key2, key3

---

\*abc

<sup>†</sup>abc

# 1 Introduction

Informer based ranking (IBR) strategies are a family of methods that arise from early-stage drug discovery (Zhang, Ericksen, and Lee *et al.* 2019). It is designed for computational guidance of drug-discovery experiments. The context is virtual screening (Sliwoski *et al.* 2014). Briefly speaking, a chemical screening facility receives material representing a purified protein target, and is tasked with identifying drugs that may have some inhibitory effect on the bioactivity of the protein. For example, Zhang, Ericksen, and Lee *et al.* (2019) report on drugs that affect certain protein kinases. The facility has specific biochemical assays that test for inhibitory effects of a given drug, and it has access to a very large library of different drug compounds, some small fraction of which are expected to have bioactivity against the new target. Costs limit the number of compounds that may be screened experimentally. The virtual screening task is to prioritize the drugs in advance of experimental screening by some computations so as to improve the hit rate of the experiments performed on the new target.

Virtual screening may operate on different kinds of target and compound data, including possibly chemical structures or the results of experimental assays. For example, when chemical structures are available it is possible to use sophisticated molecular docking programs that aim to find best fits between target and compound structures (e.g. Kitchen *et al.* 2004, Erickson *et al.* 2017) We focus on informer based ranking strategies that need no information on chemical structure, only rely on bioactivity data of many compounds against many targets, and not any bioactivity data from the new target on test. With these initial bioactivity data, informer based ranking method identifies a subset of compounds – the *informer set* – which represent a small fraction of the available compounds and which we will test experimentally for bioactivity against the new target. Subsequently, we use the full set of data (initial data on other targets for many compounds and intermediate data on informer compounds on the new target) in order to prioritize the remaining, non-informer compounds for their likely bioactivity against the new target. If we have done a good job in building

the informer set, then the top of the prioritized list of non-informers will be enriched for truly active compounds. We will have spent computational time building the informer set from initial data, and experimental resources obtaining intermediate data; the payoff is that further experiments will return truly active drug compounds at a higher rate than if we had selected compounds to test by some other scheme.

One common misunderstanding of informer based ranking method is that many researchers tend to select the most frequent hitters in bioactivity matrix as the informer set. We come up with a simple counterexample in Appendix to show that informer set ranking problem is a more complicated problem. In short, most informative compounds are not those that are most likely to be active. Zhang, Ericksen and Lee *et al.* (2019) introduced several informer based ranking strategies and compared them empirically in data on protein kinases. Though differing in detail, the proposed strategies all partition the initial bioactivity data, producing clusters of relatively similar targets; then informer compounds are those that are predictive of the cluster label of a target. Among kinase targets, the proposed bioactivity-based strategies had better predictive performance than commonly used strategies based upon chemical structure. Since effective informer based ranking strategies leverage statistical patterns in the initial bioactivity data, we reason that statistical modeling may provide a useful approach to deriving more effective IBR strategies than are currently available.

In this article, we formalize the informer set ranking procedure as a two-stage decision problem and introduce a new method named as Bayes Optimal Informer SEt (BOISE) to select informer set as well as ranking the compounds against the new target. The proposed method is fully Bayesian and can be applied to other ranking problems where only limited information can be derived from an unknown new target. We show in our analysis that this method is guaranteed to be Bayes rule if computed exactly, and can be efficiently computed with an approximate version to achieves a balance between ranking performance and computation complexity. The proposed methods are applied to human kinase chemogenomics data set PKIS1 and anti-cancer drug sensitivity data set GDSC1. We compare new methods

with previous methods under different informer set sizes and different ranking metrics, such as NEF10, ROCAUC, MCC and F1. We demonstrate that new methods are significantly better than previous methods under these data sets and metrics.

## 2 Methodology

### 2.1 Problem Set Up

We are given an  $m \times n$  binary matrix  $x_0$  containing bioactivity data measured on  $m$  targets with index set  $I = \{1, 2, \dots, m\}$ , and  $n$  compounds with index set  $J = \{1, 2, \dots, n\}$ . Each entry  $x_{ij}$  of  $x_0$  is a summary of experiments, such as inhibition assays, of compound  $j$  on target  $i$ . To be more specific,  $x_{ij} = 1$  means compound  $j$  is active on target  $i$  while  $x_{ij} = 0$  means inactive. Binary outcome is of more interest to us because our task is to rank all active compounds on a given target prioritized above the inactives. Although our method is based on binary matrix, it can be applied to any continuous bioactivity matrices as we can always convert a continuous matrix into a binary matrix. For example, if  $z$  is a continuous matrix, we can use a target-wise threshold criterion (1) to binarize  $z$  based on sample mean and standard deviation of each row:

$$x_{ij} = \begin{cases} 1, & \text{if } z_{ij} \geq \text{mean}[z_{i\cdot}] + 2 \times \text{s.d.}[z_{i\cdot}], \\ 0 & \text{if } z_{ij} < \text{mean}[z_{i\cdot}] + 2 \times \text{s.d.}[z_{i\cdot}]. \end{cases} \quad (1)$$

2SD threshold (1) is widely used to binarize outcomes in biomedical science.

Now let  $i^* \notin I$  index a new protein target of interest, on which we have no initial bioactivity data. We are asked to find a so-called informer set  $A \subset J$  using initial data  $x_0$ . The size of  $A$  is a given small number  $n_A \ll n$ . We will perform a new experiment on new

target  $i^*$  and informer set  $A$  to obtain intermediate data

$$x_A = \{x_{i^*j} : j \in A\}.$$

This  $x_A$  measure the bioactivity of informer set compounds against the new target  $i^*$ . The initial and intermediate data are then used to prioritize compounds for further testing. To formalize this, let us suppose that we must select a final top set  $T$ , of certain cardinality  $n_T$ , with

$$T = T(x_0, A, x_A) \subset J.$$

The notation is intended to emphasize that the proposed top set  $T$  is constructed from initial data, the informer set, and the intermediate data. We imagine that all the compounds  $j$  in  $T$  will be tested against new target  $i^*$ , and we seek this top set  $T$  to have as many active compounds as possible.

In Zhang, Ericksen, and Lee *et al.* (2019), there are 3 novel informer-based-ranking (IBR) methods: Regression Selection (RS), Coding Selection (CS), Adaptive Selection (AS), as well as 2 baseline methods: Baseline Chemometric (BC) and Baseline Frequent-hitters (BF). In this paper, we will introduce a new IBR method named as Bayes Optimal Informer SEt (BOISE) that is optimal in the context of Bayes decision theory.

## 2.2 Statistical Model

Our statistical model is a Bayesian model on both original known data  $x_0$  and new target data  $x_{i^*}$ . To represent similarity among targets, we assume there is a clustering structure  $\mathcal{C}$  on targets  $i \in I$ , and targets within each cluster will have similar interactions with compounds. To be more specific, we assume  $\mathcal{C} = \{c_k, k = 1, 2, \dots\}$  is a partition on all  $m$  targets. Each  $c_k$  labels  $m_k$  targets out of total  $m$  targets. New target  $i^*$  may be in one existing cluster  $c_k$  or form a new cluster by itself, denoted as  $c_0$ . For each entry  $x_{ij}$  of the bioactivity data  $x_0$ ,

we assume that it is a realization of mutually independent Bernoulli trials, determined by both compound  $j$  and the cluster  $c_k$  where corresponding target  $i$  belongs:

$$X_{ij} \sim_{ind} \text{Bernoulli}(\theta_{ij});$$

$$\theta_{ij} = \phi_{kj} \mathbb{1}(i \in c_k).$$

To complete our Bayes framework, we will assume conjugate prior  $\phi_{kj} \sim \text{Beta}(\alpha_0, \beta_0)$  on success rate  $\phi_{kj}$ . Hyperparameters  $(\alpha_0, \beta_0)$  are homogeneous for all compounds  $j$  and clusters  $c_k$ , meaning we have no preference on compounds or clusters. Following the ideas of empirical Bayes methods, we choose hyperparameters so that  $(\alpha_0, \beta_0)$  so that  $\frac{\alpha_0}{\alpha_0 + \beta_0} = \text{mean}(x_0)$ .

In summary, our generative model is:

$$\begin{aligned} \phi_{kj} &\sim \text{Beta}(\alpha_0, \beta_0); \\ \mathcal{C} &\sim \text{CR}(m_0); \\ P(x_{ij} = 1 \mid i \in c_k, \mathcal{C}) &= \phi_{kj}. \end{aligned} \tag{2}$$

**Note:** The reason we use notation  $\mathcal{C}$  is that we consider clustering structure as a random process on all possible partitions of countable targets space. Dirichlet process mixture model (DPMM) is a classical method to handle this kind of clustering structures, and we are using a special case of DPMM called Chinese restaurant (CR) process based on some prior mass  $m_0$  to make the posterior inference. There are two main advantages of using DPMM instead of traditional clustering methods like spectral clustering, hierarchical clustering and k-means: On the one hand, we don't need to specify the number of clusters  $K$  beforehand, or manually cut the tree to get the clustering structure. All we need to set is a prior mass  $m_0$  that describes our belief of divergence and all the rest is data-driven. Thus, DPMM is less subjective than other methods. On the other hand, DPMM is a non-parametric Bayesian clustering method and hence is compatible with our Bayesian framework. Also, since the sample size is not very large ( $\approx 200$  targets), number of targets falling in each cluster can be unbalanced, with some

clusters containing more than 50 targets while other clusters containing less than 5 targets. DPMM is well designed to address this unbalanced impact under Bayesian framework. We will come to details of DPMM later in computation section.

## 2.3 Decision Rule

We are faced with a two-stage decision problem: First, we need a rule to select an informer set  $A$  of size  $n_A$  from  $x_0$ . Then we need to select a top set  $T$  of size  $n_T$  based on initial data  $x_0$ , informer set  $A$  and intermediate data  $x_A$ . Denote these decision rules as  $A = A(x_0)$  and  $T = T(A, x_A, x_0)$  to show their dependency.

Our ultimate goal is to maximize the proportion of active compounds in top set  $T$  with regard to new target  $i^*$ , therefore it is nature to define our loss function on informer-set rule  $A$  and top-set rule  $T$  as:

$$L(A, T) = \sum_{j \in T} (1 - \theta_{i^*j}) \quad (3)$$

where  $\theta_{i^*j}$  is the marginal success rate of new target  $i^*$  on compound  $j$ . Notice that  $\mathbb{E} \sum_{j \in T} \mathbb{1}(x_{i^*j} = 0) = \sum_{j \in T} (1 - \theta_{i^*j}) = L(A, T)$ . Our loss function can be seen as the expected number of inactive compounds in the top set, so we can achieve our goal by minimizing  $L(A, T)$ . We include the informer-set rule  $A$  in the loss, since its value tells us which compounds to use for intermediate data. As presented, this optimization is not empirically resolvable because of the unknown parameters  $\theta_{i^*j}$ , however, by averaging we may try to minimize some kind of expected loss (*i.e.*, risk).

We are thus guided to Bayesian statistical decision theory (Berger, 1985; Robert, 2007; Parmigiani and Inoue, 2009). Recall that after integrating with respect to  $\mathcal{C}$ , each entry has  $X_{ij} \sim_{ind} \text{Bernoulli}(\theta_{ij})$  marginally, we can score possible rules  $A$  and  $T$  by their Bayes risk,

that is, the marginal expected loss over the joint distribution:

$$r(A, T) = \int_{x_0} \int_{x_A} \int_{\Theta} L[A(x_0), T(A, x_A, x_0)] p(x_0|\theta) p(x_A|x_0, \theta) p(\theta) d\theta dx_A dx_0 \quad (4)$$

In a slight abuse of notation, we use integral signs to indicate averaging over data  $x_0$  or  $x_A$ , when really summations over the discrete sample space are being computed. The inner integral is over all components of the parameter matrix  $\theta$ . The sampling probabilities  $p(x_0|\theta)$  and  $p(x_A|x_0, \theta)$  come from the Bernoulli model. A pair of rules that minimize the Bayes risk are called Bayes rules; in the context of the model, the prior and loss, their use is a rational way to design and carry out the experiment.

Since it is a two-stage decision problem, we will also construct our decision rule in two steps: First, we derive the Bayes rule of top set  $T^*(A, x_A, x_0)$  for any fixed informer set  $A$  and intermediate data  $x_A$ ; Then, we will plug-in this  $T^*(A, x_A, x_0)$  to derive decision rule  $A(x_0)$  for informer set.

Consider when we have initial data  $x_0$  as well as intermediate data  $x_A$  on some pre-specified informer set  $A$ , we can get the posterior distribution of  $\theta$ :

$$p(\theta|x_0, x_A) = \frac{p(x_0|\theta)p(x_A|x_0, \theta)p(\theta)}{p(x_0, x_A)}.$$

For all compounds  $j$ , we then construct the posterior means under this distribution as:

$$\hat{\theta}_{i^*j} = \mathbb{E}(\theta_{i^*j}|x_A, x_0)$$

We define a particular top set rule  $T^*$  to select the top  $n_T$  compounds with largest posterior mean:

$$T^*(A, x_A, x_0) = \left\{ j \in J : \sum_{k \in J} \mathbb{1}(\hat{\theta}_{i^*k} > \hat{\theta}_{i^*j}) \leq n_T \right\}. \quad (5)$$



For the loss (3), it is known that  $T^*$  minimizes the Bayes risk (Henderson 2016). We utilize this finding to resolve one part of our compound-decision problem, obtaining:

**Theorem 1** *Suppose  $\hat{\theta}_{i^*j}$  denotes posterior mean  $\mathbb{E}(X_{i^*j}|x_A, x_0)$ . For any rules  $A$  and  $T$ , the Bayes risk  $r(A, T)$  will be bounded from below:*

$$r(A, T) \geq \int_{x_0} p(x_0) \int_{x_A} p(x_A|x_0) \left[ \sum_{j \in T^*(A, x_A, x_0)} (1 - \hat{\theta}_{i^*,j}) \right] dx_A dx_0$$

The lower bound above is in fact the Bayes risk associated with the best possible top-set rule for any given informer-set  $A$ . Essentially, this shows how to score any informer-set rule by profiling out the top-set selection.

To identify the compound Bayes rule, we work conditionally upon  $x_0$ , introducing the first stage posterior expected loss:

$$\text{PEL}_1(x_0, A) = \int_{x_A} p(x_A|x_0) \left[ \sum_{j \in T^*(A, x_A, x_0)} (1 - \hat{\theta}_{i^*,j}) \right] dx_A \quad (6)$$

which is the quantity within the marginal  $x_0$  integral in Theorem 1, taken at some set  $A$ . In other words, Theorem 1 could be restated as

$$r(A, T) \geq \int_{x_0} \text{PEL}_1(x_0, A(x_0)) p(x_0) dx_0.$$

Using a standard result from Bayesian decision theory, the rule  $A$  that minimizes the marginal Bayes risk is obtained by finding the best informer set at each  $x_0$  (Berger, 1985, page 159). Thus, the Bayes optimal informer set  $A^*(x_0)$  is

$$A^*(x_0) = \underset{A \subset J, |A|=n_A}{\text{argmin}} \text{PEL}_1(x_0, A). \quad (7)$$

Then the Bayes optimal top set rule is  $T^*(A, x_A, x_0)$  in (5) applied to intermediate data  $x_{A^*}$  obtained on the informer set  $A^*$ .

Notice that  $\text{PEL}_1(x_0, A)$  can be seen as a loss function on possible informer set  $A$  when  $x_0$  is given. Intuitively, any loss function on  $A$  should be non-increasing when  $A$  gets larger, as we will get more information when we have a larger informer set  $A$  and hence we should make a better decision. The following theorem will show that our  $\text{PEL}_1(x_0, A)$  loss is well-defined to be compatible with this intuition:

**Theorem 2** *Suppose  $\text{PEL}_1(x_0, A)$  is defined as in (6). When  $x_0$  is given, for any  $A \subset B \subset J$ , we have:*

$$\text{PEL}_1(x_0, A) \geq \text{PEL}_1(x_0, B).$$

In preparing for computation, it is also useful to define a second stage posterior expected loss:

$$\text{PEL}_2(x_0, x_A, A) = \sum_{j \in T^*(A, x_A, x_0)} (1 - \hat{\theta}_{i^*, j})$$

and from (6) we have  $\text{PEL}_1(x_0, A) = \int_{x_A} p(x_A | x_0) \text{PEL}_2(x_0, x_A, A, T^*) dx_A$ , or equivalently, we can write  $\text{PEL}_1$  in conditional expectation form:

$$\begin{aligned} \text{PEL}_1(x_0, A) &= \mathbb{E} [\text{PEL}_2(x_0, x_A, A) | x_0] \\ &= \mathbb{E} [\mathbb{E} (\text{PEL}_2(x_0, x_A, A) | \mathcal{C}, x_0) | x_0]. \end{aligned} \tag{8}$$

This condition expectation form (8) will be of great importance in computing  $\text{PEL}_1(A, x_0)$ .

**Note:** In BOISE, we don't require  $T$  and  $A$  to be disjoint. Therefore we need to prevent inactive compounds tested in intermediate data  $x_A$  from reentering top set  $T$  in computation.

### 3 Computations

Mathematically speaking, optimization problems defined in (5) and (7) can be solved exactly, since integral in (6) is actually a sum over fixed length binary array, and feasible set of (7) is a finite set with size determined by  $n_A$ . However, below we will show that exact solution is computationally infeasible for BOISE, and hence we have to turn to approximate solutions in computation:

There are three main computation parts in BOISE: First, we need to apply DPMM clustering on original data  $x_0$ ; Then, for each possible informer set  $A$  we need to calculate the  $\text{PEL}_1(x_0, A)$  loss as in (6), and hence we need integral with respect to the predictive distribution  $p(x_A|x_0)$  and posterior expectation  $\mathbb{E}(\theta|x_0, x_A)$ ; Finally, we need to find the optimal informer set  $A^*$  in optimization problem (7), which is over discrete, size  $n_A$  subsets of compounds  $J = \{1, 2, \dots, n\}$ .

All these three parts are computationally intensive, especially the second and third part. For example, suppose we only consider the scenario where we have fixed cluster assignment. If we have  $m$  targets and  $n$  compounds, then we need  $\mathcal{O}(mn)$  time for evaluating posterior expectation  $\mathbb{E}(\theta|x_0, x_A)$  for each compound, and  $\mathcal{O}(mn \times 2^{n_A})$  time for calculating integral on all  $2^{n_A}$  possible intermediate outcome  $x_A$ . As for optimal informer set selection, we will need to evaluate  $\text{PEL}_1(x_0, A)$  loss on  $\binom{n}{n_A}$  possible informer sets  $A$  of length  $n_A$ . Remember each evaluation requires  $\mathcal{O}(mn \times 2^{n_A})$  time, this implies the total time complexity for BOISE is  $\mathcal{O}(mn^{n_A+1} \times 2^{n_A})$ . Having  $n_A$  in exponential term means exact solution to BOISE is infeasible because of exponential explosion. Therefore, we come up with several approximation computation methods to reduce running time as well as maintain the performance of BOISE.

Briefly speaking, we use Markov Chain Monte Carlo (MCMC) for DPMM clustering and  $\text{PEL}_1(x_0, A)$  integral calculation for possible intermediate  $x_A$ . As for informer set selection, we gradually augment our informer set via greedy search. Details of our algorithm are contained in the following sections.

### 3.1 Dirichlet Process Mixture Model

Dirichlet process mixture model (DPMM) origins from Antoniak (1974) and Ferguson (1983). It is a powerful tool for clustering data that enables the inference of a countably infinite number of mixture components. Using countably infinite mixtures bypasses the need of assumption that we "know" the correct number of clusters, and can be handled in a Bayesian framework by employing Dirichlet process as prior distribution on mixing proportions. DPMM computations have become feasible via MCMC methods for sampling from the posterior distribution of parameters corresponding to distribution of each cluster and of the associations of mixture components with observations. In particular, when conjugate prior distributions are used, algorithms based on Gibbs sampling can easily be implemented for DPMM.

DPMM applies to data  $x_1, \dots, x_n$  which are regarded as exchangeable, or equivalently, as independent samples of some unknown distributions. The distribution from which  $x_i$  are sampled can be parameterized as a mixture of distributions of the form  $F(\theta)$ , and the mixing distribution of  $\theta$  is  $G$ . Typical choice of prior for mixing distribution  $G$  is a Dirichlet process with base distribution  $G_0$  and prior mass  $m_0$ . In summary, a general form of DPMM is:

$$\begin{aligned} x_i | \theta_i &\sim F(\theta_i); \\ \theta_i | G &\sim G; \\ G &\sim DP(G_0, m_0). \end{aligned} \tag{9}$$

In our Bayesian model, each  $x_i$  is a multivariate random vector, and each component  $x_{ij}$  is an independent Bernoulli trial with success rate  $\theta_{ij}$ . Therefore,  $F(\theta_i)$  under BOISE setting is:

$$F(\theta_i) = \prod_{j=1}^n \theta_{ij}^{x_{ij}} (1 - \theta_{ij})^{1-x_{ij}}$$

and our base distribution  $G_0$  is homogeneous Beta( $\alpha_0, \beta_0$ ) on each component  $\theta_{ij}$ .

Ferguson (1973) has shown measures drawn from a Dirichlet process (9) are discrete with probability 1, therefore DPMM can be seen as countably infinite mixtures. Also, we can see

that draws from Dirichlet process are both discrete and exhibit a clustering property through Pólya urn scheme (Blackwell and MacQueen 1973): Suppose  $\theta_1, \theta_2, \dots$  is a sequence of i.i.d random variables distributed with respect to  $G$ , then successive conditional distribution of  $\theta_i$  given previous  $\theta_1, \theta_2, \dots, \theta_{i-1}$  can be derived by integrating  $G$  out in model (9):

$$\theta_i | \theta_1, \dots, \theta_{i-1}, m_0, G_0 \sim \frac{1}{i-1+m_0} \sum_{j=1}^{i-1} \delta(\theta_j) + \frac{m_0}{i-1+m_0} G_0 \quad (10)$$

where  $\delta(\theta_j)$  is the probability measure concentrated on single point  $\theta_j$ .

This conditional distribution shows that  $\theta_i$  has positive probability of being equal to one of the previous draws and there is a clustering tendency: the more one point is drawn in the past, the more often it will be drawn in the future. It can be shown more clear by re-parameterization: Let  $\phi_1, \phi_2, \dots, \phi_K$  be the distinct values in  $\theta_1, \theta_2, \dots, \theta_{i-1}$  and  $m_k = \sum_i \mathbb{1}(\theta_i = \phi_k)$  be the number of  $\theta_i$  that is equal to  $\phi_k$ , then previous conditional distribution (10) can be rewritten as:

$$\theta_i | \theta_1, \dots, \theta_{i-1}, m_0, G_0 \sim \sum_{k=1}^K \frac{m_k}{i-1+m_0} \delta(\phi_k) + \frac{m_0}{i-1+m_0} G_0 \quad (11)$$

and this expression (11) is the well-known **Chinese Restaurant** (CR) process (Aldous 1985), which is a distribution on partitions. We can interpret conditional distribution (11) in terms of a Chinese restaurant with an infinite number of tables: Imagine each  $\phi_k$  is a table and each  $\theta_i$  is the  $i$ th customer. When  $m$ th customer coming into the restaurant, he will sit at table  $k$  with probability  $\frac{m_k}{m_0+m-1}$  and sit at a brand new table with probability  $\frac{m_0}{m_0+m-1}$ . This process will generate a clustering on  $n$  customers and is an efficient representation of Dirichlet process.

The Chinese restaurant process representation (11) implies we can write Dirichlet process mixture model (9) in a limiting form with latent cluster labels  $c$  (Neal 2000):

$$\begin{aligned}
x_i | c, \phi &\sim F(\phi_{c_i}); \\
c_i | p &\sim \text{Discrete}(p_1, \dots, p_K); \\
\phi_c &\sim G_0; \\
p_1, \dots, p_K &\sim \text{Dirichlet}(\frac{m_0}{K}, \dots, \frac{m_0}{K})
\end{aligned} \tag{12}$$

where  $c_i$  is the cluster label that we need in clustering. When  $K \rightarrow \infty$  the conditional distribution  $c_i | c_1, \dots, c_{i-1}$  will approximate a Chinese restaurant process, and model (12) is equivalent to (9). By taking the limit, the posterior distribution of  $c_i$  given all other cluster labels  $c_{-i}$ , observation  $x_i$  and parameters corresponding to existing clusters  $\Phi = (\phi_1, \dots, \phi_K)$  is:

$$P(c_i = c | c_{-i}, x_i, \Phi) \propto \begin{cases} \frac{m_{-i,c}}{m-1+m_0} F(x_i, \phi_c), & \text{if } \exists c_j = c, j \neq i \\ \frac{m_0}{m-1+m_0} \int F(y_i, \phi) dG_0(\phi). & \text{if } \nexists c_j = c, j \neq i. \end{cases} \tag{13}$$

Equation (13) is already eligible for Gibbs sampling on  $(c, \phi_c)$ , however, under conjugate prior setting we can even collapse computation by integrating  $\phi_c$  out: Define  $H_{-i,c}(\phi) = P(\phi | G_0, \{x_j : j \neq i, c_j = c\})$  to be the posterior distribution of  $\phi_c$  given  $G_0$  and all other observations  $x_j$  in cluster  $c$  except  $x_i$ , then equation (13) can be marginally written as:

$$P(c_i = c | c_{-i}, x_i) \propto \begin{cases} \frac{m_{-i,c}}{m-1+m_0} \int F(x_i, \phi) dH_{-i,c}(\phi), & \text{if } \exists c_j = c, j \neq i \\ \frac{m_0}{m-1+m_0} \int F(y_i, \phi) dG_0(\phi). & \text{if } \nexists c_j = c, j \neq i. \end{cases} \tag{14}$$

This conditional distribution (14) leads to the Gibbs sampling method summarized in Neal (2000), which is applied in BOISE computation:

Notice that under BOISE setting, base distribution  $G_0$  is a conjugate prior  $\text{Beta}(\alpha_0, \beta_0)$  on each component, and hence posterior distribution  $H_{-i,c}(\phi)$  can be computed analytically. Therefore, we could have a closed form expression for posterior distribution  $P(c_i | c_{-i}, y_i)$  in

---

**Algorithm 1** DPMM clustering

---

- 1: Set the prior mass  $m_0$  and other hyperparameters;
  - 2: Initialize  $c_1, \dots, c_m$  with regard to Chinese restaurant process  $\text{CR}(m_0)$ ;
  - 3: Set sample size of clustering assignments  $M$ , and gaps between successive draws  $N$
  - 4: **while** Sample size  $< M$  **do**
  - 5:     **for**  $i = 1, \dots, m$  **do**
  - 6:         Update  $i$ th cluster label from posterior distribution  $P(c_i | c_{-i}, y_i)$  in (14)
  - 7:     **end for**
  - 8:     Sample one clustering assignment after every  $N$  label updates.
  - 9: **end while**
- 

(14), and Algorithm 1 will help us sample clustering assignments fast and efficiently.

### 3.2 PEL

In this section, we will focus on how to calculate  $\text{PEL}_1(x_0, A)$  for given  $A$  and  $x_0$ , and this will help evaluate each candidate  $A$  to select the final informer set. As  $x_0$  is always fixed, we will write  $\text{PEL}_1(x_0, A)$  as  $\text{PEL}_1(A)$  for simplicity. Recall the definition of  $\text{PEL}_1(A)$  is:

$$\begin{aligned} \text{PEL}_1(A) &= \mathbb{E}[\text{PEL}_2(x_0, x_A, A) | x_0] \\ &= \int_{x_A} p(x_A | x_0) \left[ \sum_{j \in T^*(A, x_A, x_0)} (1 - \hat{\theta}_{i^*, j}) \right] dx_A \end{aligned}$$

Since exact calculation is infeasible for this task, one natural thought was to sample  $x_A$  from  $p(x_A | x_0)$  and average over all samples. However, this is also inefficient as  $x_A$  and  $x_0$  are not directly correlated in our model (2): their correlation is encoded in a hidden Dirichlet process. This enlightens us to draw samples from  $p(x_A | x_0, \mathcal{C})$  and utilize the expression (8) to get the approximation. Moreover, for each sampled  $x_A$  we need to calculate  $\text{PEL}_2(x_0, x_A, A)$  and get the best top set  $T^*$  before calculating  $\text{PEL}_1(A)$ , and this requires computing and sorting of  $\hat{\theta}_{i^*, j} = \mathbb{E}(\theta_{i^*, j} | x_A, x_0)$  for all compounds  $j$ . Therefore, our computation pipeline is to first sample from  $p(x_A | x_0, \mathcal{C})$ , then calculate  $\hat{\theta}_{i^*, j}$  and sorting  $\hat{\theta}_{i^*, j}$  to get top set  $T^*(A, x_A, x_0)$ , finally we will take eligible average to get  $\text{PEL}_1(A)$ . Details of this part of computation is as following:

From model (2) and conjugacy, we can get the posterior distributions:

$$\begin{aligned} \phi_{kj} | x_0, \mathcal{C} &\sim \text{Beta}(a_{kj}, b_{kj}); \\ x_{i^*j} | i^* \in c_k, \mathcal{C}, x_0 &\sim_{\text{ind}} \text{Bernoulli} \left( \frac{a_{kj}}{a_{kj} + b_{kj}} \right). \end{aligned} \quad (15)$$

where  $a_{kj} = \alpha_0 + \sum_{i \in c_k} x_{ij}$  and  $b_{kj} = \beta_0 + \sum_{i \in c_k} (1 - x_{ij})$  are posterior hits and misses given  $x_0$ , respectively. Notice that when the new target  $i^*$  enters and no response on  $i^*$  is observed, the cluster label of  $i^*$  will follow the rules of CR process with the same prior mass  $m_0$ :

$$P(i^* \in c_k | x_0, \mathcal{C}) = \frac{m_k}{m_0 + m} \quad (16)$$

where  $m_k$  is the number of targets fall in  $k$ th cluster  $c_k$ ,  $m$  is number of total targets in  $x_0$ , and  $c_k = 0$  means new target  $i^*$  forms a new cluster. These posterior distributions tell us how to sample  $x_A$  from  $p(x_A | x_0, \mathcal{C})$ : When  $\mathcal{C}$  and  $x_0$  are given, we first sample the cluster labels of  $i^*$  from (16), then draw  $x_{i^*j}$  from (15) for each  $j \in A$ .

With the sampled intermediate data  $x_A$ , we also need posterior expectation  $\hat{\theta}_{i^*j}$  for top set rule  $T^*$ . After similar calculation, we can get an expression for  $\hat{\theta}_{i^*j}$  without separate MCMC. To be more specific, by conditioning on  $\mathcal{C}$  and  $i^* \in c_k$  we have:

$$\mathbb{E}(\theta_{i^*j} | x_0, x_A, \mathcal{C}, i^* \in c_k) = \frac{a_{kj} + x_{i^*j} \mathbb{1}(j \in A)}{a_{kj} + b_{kj} + \mathbb{1}(j \in A)}; \quad (17)$$

$$p_k = P(i^* \in c_k | \mathcal{C}, x_0, x_A) \propto m_k \prod_{j \in A} \left( \frac{a_{kj}}{a_{kj} + b_{kj}} \right)^{x_{i^*j}} \left( \frac{b_{kj}}{a_{kj} + b_{kj}} \right)^{1-x_{i^*j}}. \quad (18)$$

Combine (17) and (18) will give us the conditional expectation of  $\mathbb{E}(\theta_{i^*j} | \mathcal{C}, x_0, x_A)$ :

$$\hat{\theta}_{i^*j}(\mathcal{C}, x_0, x_A) = \mathbb{E}(\theta_{i^*j} | \mathcal{C}, x_0, x_A) = \sum_{k=0}^K p_k \mathbb{E}(\theta_{i^*j} | x_0, x_A, \mathcal{C}, i^* \in c_k). \quad (19)$$

and the last step for desired  $\hat{\theta}_{i^*j}(x_0, x_A)$  is to take the average of (19) with respect to  $\mathcal{C}$  sampled from Algorithm 1. In summary, our procedure for  $\text{PEI}_1$  calculation is Algorithm 2:



---

**Algorithm 2** PEL Computation

---

**Input:** Initial data  $x_0$ , informer set  $A$ , size of top set  $n_T$ .

**Output:**  $\text{PEL}_1(x_0, A)$  as score of informer set  $A$ .

- 1: Draw clustering samples  $\mathbb{C}$  from  $p(\mathcal{C} | x_0)$  with Algorithm 1.
- 2: **for** each  $\mathcal{C}$  in  $\mathbb{C}$  **do**
- 3:     Sample  $\mathbb{X}_A$  from  $p(x_A | \mathcal{C}, x_0)$  as (15) and (16); ▷ Intermediate data sample
- 4:     **for** each  $x_A$  in  $\mathbb{X}_A$  **do**
- 5:         Calculate conditional expectation  $\mathbb{E}(\theta_{i^*j} | x_0, x_A, \mathcal{C})$  using (19);
- 6:         ▷ Posterior expectation given  $x_0, x_A$  and  $\mathcal{C}$
- 7:         Average over  $\mathcal{C}$  to get  $\hat{\theta}_{i^*j}(x_0, x_A)$  and top set  $T^*(A, x_A, x_0)$
- 8:     **end for**
- 9:     Average over  $x_A$  to get conditional expectation  $\mathbb{E}(\text{PEL}_2(A, x_0, x_A) | \mathcal{C}, x_0)$
- 10: **end for**
- 11: Calculate  $\text{PEL}_1$  loss by averaging over  $\mathcal{C}$  as in (8):

$$\text{PEL}_1(A, x_0) = \mathbb{E}[\mathbb{E}(\text{PEL}_2(A, x_0, x_A) | \mathcal{C}, x_0) | x_0]$$

---

Notice that in Algorithm 2, the sample sizes of  $\mathbb{C}$  and  $\mathbb{X}_A$  are both preset constants, therefore time complexity for computing  $\text{PEL}_1(A)$  is reduced to  $\mathcal{O}(mn)$  instead of  $\mathcal{O}(mn \times 2^{n_A})$  as in exact computation.

### 3.3 Informer Set Selection

Though potentially biased, greedy algorithm seems to be a good fit to our informer set optimization problem (7): The complicated definition of  $\text{PEL}_1(x_0, A)$  in (6) results in an implicit form of objective function, and the discrete nature of informer set  $A$  rules out gradient based optimization methods. Also, Theorem 2 shows that  $\text{PEL}_1(x_0, A)$  is non-decreasing in  $A$  given  $x_0$ , and hence we can do no harm to our informer set when applying greedy algorithm on informer set selection. We come up with 2 kinds of informer selection methods based on greedy algorithm: One is simply following the procedure in Algorithm 2 to get an estimation of  $\text{PEL}_1(x_0, A)$  and make the selection; the other is using the information gain defined by entropy to approximate the  $\text{PEL}_1$  loss of  $A$  and select the best informer set. From empirical analysis, both methods are significantly better than previous informer based ranking methods in Zhang, Ericksen, and Lee *et al.* (2019), and the exact method is slightly

better than approximate method while the approximate method is much faster than exact one. We will describe our methods in this section and show the empirical results later in section 4.

The first method is a straightforward sequential selection. At first we evaluate  $\text{PEL}_1(x_0, A)$  for all  $A = 1, \dots, n$  to select the best  $A$  with least  $\text{PEL}_1$  loss. Then at each iteration we memorize all previous decisions  $A$  and select one best element  $j \in J \setminus A$  that minimizes  $\text{PEL}_1(A \cup \{j\})$ . The procedure can be summarized as following Algorithm 3:

---

**Algorithm 3** Exact Informer Selection

---

**Input:** Initial data  $x_0$ , size of informer set  $n_A$ , size of top set  $n_T$ .

**Output:** Selected informer set of length  $n_A$

- 1: **Initialization:** Evaluate  $\text{PEL}_1(x_0, A)$  for all  $|A| = 1$ . Let  $A^* = \text{argmin}_{|A|=1} \text{PEL}_1(x_0, A)$ .
  - 2: **while**  $|A^*| < n_A$  **do**
  - 3:     Evaluate  $\text{PEL}_1(x_0, A^* \cup \{j\})$  for each  $j \in J \setminus A^*$ .
  - 4:     Let  $A^* \leftarrow A^* \cup \{j^*\}$ , where  $j^* = \text{argmin}_j \text{PEL}_1(x_0, A^* \cup \{j\})$ .
  - 5: **end while**
  - 6: Return  $A^*$  as selected informer set.
- 

The second method is a more heuristic one enlightened by decision tree. Recall that in  $\text{PEL}_1$  calculation procedure, the key step is formula (19) that tells us how to calculate posterior expectation  $\mathbb{E}(\theta_{i^*j} | \mathcal{C}, x_0, x_A)$ , and this helps us think about the role that informer set  $A$  plays in BOISE framework. Notice that (19) consists of two parts: the posterior probability  $p_k = P(i^* \in c_k | \mathcal{C}, x_0, x_A)$ , and another posterior expectation  $E(\theta_{i^*j} | x_0, x_A, \mathcal{C}, i^* \in c_k)$ . We claim that informer set  $A$  has little impact on the second part, since from expression (17) we can see that for  $j \notin A$ , informer set  $A$  and intermediate data  $x_A$  has nothing to do with  $E(\theta_{i^*j} | x_0, x_A, \mathcal{C}, i^* \in c_k)$ , while for  $j \in A$  we already know the interaction of  $j$  on  $i^*$  through intermediate data  $x_A$  and hence we don't really need  $E(\theta_{i^*j} | x_0, x_A, \mathcal{C}, i^* \in c_k)$  in selecting top set  $T^*(A, x_A, x_0)$ . Therefore, informer set  $A$  affects value of objective function  $\text{PEL}_1(x_0, A)$  mainly through posterior probability  $p_k$ . This observation suggests that the optimal informer set  $A^*$  should give us an "optimal" posterior distribution  $\{p_k, k = 1, \dots, K\}$ , and hence we turn to entropy to define a "good" posterior distribution and find an approximate "best" informer set.

Entropy is a measure of uncertainty associated with a random variable and is defined as expected bits required to communicate the value of this random variable:

$$H(Y) = - \int_y p(y) \log_2 p(y) dy.$$

Typically, less entropy means less uncertainty and is exactly what we desired. If we consider the cluster label of  $i^*$  as a random variable  $c(i^*)$ , then the posterior distribution of  $c(i^*)$  given  $\mathcal{C}$ ,  $x_A$  and  $x_0$  is a multinomial distribution with  $\{p_k, k = 1, \dots, K\}$  as event probabilities. In this way, we can define a "good" posterior distribution by its conditional entropy:

$$H(c(i^*) | \mathcal{C}, x_0, x_A) = - \sum_{k=1}^K p_k \log_2 p_k \quad (20)$$

Using (20) we can score an informer set  $A$  by its posterior expected entropy:

$$\begin{aligned} \mathcal{H}(A) &= \mathbb{E}[H(c(i^*) | \mathcal{C}, x_0, x_A) | x_0] \\ &= \int_{\mathcal{C}} p(\mathcal{C} | x_0) \int_{x_A} p(x_A | x_0, \mathcal{C}) H(c(i^*) | \mathcal{C}, x_0, x_A) dx_A d\mathcal{C}. \end{aligned} \quad (21)$$

and this is our approximate objective function for informer set selection. Our selected informer set  $A^*$  should maintain a less posterior entropy  $\mathcal{H}(A)$ , and hence contain less uncertainty.

The intuition for expression (21) is strongly connected with ID3 decision tree developed by Quinlan (1986): If we fix the original data  $x_0$  and clustering assignment  $\mathcal{C}$ , then an ID3 decision tree algorithm aims for a split feature set  $A \subset J$  that maximize the mutual information (or information gain):

$$\text{MI}(x_0, A) = H(c(i^*) | \mathcal{C}, x_0) - \int_{x_A} p(x_A | x_0, \mathcal{C}) H(c(i^*) | \mathcal{C}, x_0, x_A) dx_A. \quad (22)$$

Since  $H(c(i^*) | \mathcal{C}, x_0)$  is a constant when  $\mathcal{C}$  and  $x_0$  are given, maximizing  $\text{MI}(x_0, A)$  in (22) is

equivalent to minimizing the inner integral of  $\mathcal{H}(A)$  in (21). It may worth to say that our approximate informer selection method is different from ID3 decision tree: our goal is not to predict the class label of  $i^*$ , as our clustering assignment is a random process on partitions. What we really need is a less chaotic system that can benefit our top set rule  $T^*(A, x_A, x_0)$  and reduce the final loss  $L(A^*, T^*)$ . However, for the sake of easy understanding, our approximate informer selection method can be regarded as an ID3 decision tree algorithm with the same split feature of each level and maximum depth  $n_A$ . Also, we use ID3 decision tree algorithm instead of C4.5 algorithm (Quinlan 1993) since each candidate split feature  $j \in J$  has the same number of possible outcomes: 0 and 1, so there is no bias towards features with more outcomes. Our second informer selection method is summarized in Algorithm 4:

---

**Algorithm 4** Accelerated Informer Selection

---

**Input:** Initial data  $x_0$ , size of informer set  $n_A$ , size of top set  $n_T$ .  
**Output:** Selected informer set of length  $n_A$

- 1: **Initialization:** Let  $A^* = \{\}$ . Sample clustering assignments  $\mathbb{C}$  from  $p(\mathcal{C} | x_0)$  with Algorithm 1.
- 2: **for** each  $\mathcal{C}$  in  $\mathbb{C}$  **do**
- 3:     Sample  $\mathbb{X}(\mathcal{C})$  from  $p(x_{i^*} | \mathcal{C}, x_0)$  as in (13) and (14).  $\triangleright$  **Auxiliary sample**
- 4: **end for**
- 5: **while**  $|A^*| < n_A$  **do**
- 6:     **for** each  $j \in J \setminus A^*$  **do**
- 7:         **for** each  $\mathcal{C}$  in  $\mathbb{C}$  and corresponding  $\mathbb{X}(\mathcal{C})$  **do**
- 8:             Evaluate  $H(c(i^*) | \mathcal{C}, x_0, x_{A^* \cup j})$  as in (20);
- 9:         **end for**
- 10:         Average over  $x_{A^* \cup j}$  and  $\mathcal{C}$  to get  $\mathcal{H}(A^* \cup \{j\})$  in (21).
- 11:     **end for**
- 12:     Let  $A^* \leftarrow A^* \cup \{j^*\}$ , where  $j^*$  is the minimizer of  $\mathcal{H}(A^* \cup \{j\})$ .
- 13: **end while**
- 14: Return  $A^*$  as selected informer set.

---

As mentioned in the beginning of this section, the major advantage of Algorithm 4 compared with Algorithm 3 is running time. In Algorithm 4 we bypass the computation of  $\mathbb{E}(\theta_{ij} | x_0, x_A)$  and only use  $p_k = P(i^* \in c_k | \mathcal{C}, x_0, x_A)$  to select informer set. This can improve our BOISE running time from days to hours while still maintaining similar performance. Therefore, we name it as "Accelerated-BOISE" in our empirical analysis. Also, although

Accelerated-BOISE uses different informer selection method as BOISE does, they both share the same top set rule  $T^*(A, x_A, x_0)$ , which scores each  $j \in J$  with their posterior expectation and choose the top  $n_T$  as top set.

### 3.4 Missing Data Imputation

As a Bayesian computation procedure, BOISE is naturally compatible with missing data imputations while previous informer based ranking methods are only designed for complete data sets. We will briefly discuss about missing data imputation in this section. There are two procedures in BOISE that require missing data imputation: we need imputed data to update cluster labels in DPMM clustering, and we need a complete  $x_0$  to determine  $a_{kj}$  and  $b_{kj}$  in informer selection. Therefore, we have two different strategies for missing value imputation summarized in Algorithm 5:

---

**Algorithm 5** Missing data imputation

---

- 1: Set the prior mass  $m_0$  and other hyperparameters;
  - 2: Set sample size of clustering assignments  $M$ , and gaps between successive draws  $N$
  - 3: **Initialize:** Let  $N_{ij} = 0$  for each missing entry  $x_{ij}$ . Sample  $c_1, \dots, c_m$  with regard to Chinese restaurant process  $\text{CR}(m_0)$ ;
  - 4: **while** Sample size  $< M$  **do**
  - 5:   For each missing entry  $x_{ij}$  currently in cluster  $c$ , sample a success rate  $\phi_{cj}$  from  $P(\phi | G_0, \{x_j : j \neq i, c_j = c\})$ .
  - 6:   Impute  $x_{ij}$  with Bernoulli( $\phi_{cj}$ ). Let  $N_{ij} \leftarrow N_{ij} + x_{ij}$
  - 7:   **for**  $i = 1, \dots, m$  **do**
  - 8:     Update  $i$ th cluster label from posterior distribution  $P(c_i | c_{-i}, y_i)$  in (14)
  - 9:   **end for**
  - 10:   Sample one clustering assignment after every  $N$  label updates.
  - 11: **end while**
  - 12: **if**  $N_{ij} > M/2$  **then**
  - 13:   Impute  $x_{ij}$  as 1.
  - 14: **else**
  - 15:   Impute  $x_{ij}$  as 0.
  - 16: **end if**
- 

For DPMM clustering, we impute all missing entries after a thorough update of labels, that is, after we update all cluster labels  $c_i$  for  $i = 1, \dots, m$ , we re-sample the value of missing

entries based on posterior distribution  $H_{-i,c}(\phi) = P(\phi | G_0, \{x_j : j \neq i, c_j = c\})$ . When all the clustering assignment samples are drawn, we summarize how many times each missing entry is imputed as 1, say  $N_{ij}$ , and impute  $x_{ij}$  as 1 if  $N_{ij}$  is greater than some threshold, like 50% of clustering sample size. Then we can proceed with this imputed complete data.

## 4 Empirical Results

We compare BOISE and Accelerated-BOISE with previous methods on two data sets: One is PKIS1 data set, which is thoroughly analyzed in Zhang, Ericksen, and Lee *et al.* (2019). The other is GDSC data set, which contains sensitivity data between cancer cell lines and anti-cancer drugs. Both data sets are matrices with continuous values and can be converted to binary matrices meaningfully. The metrics we mainly use to evaluate model performance are ROCAUC and NEF10, which are standard metrics in virtual screening publications and consistent with Zhang, Ericksen, and Lee *et al.* (2019). Some other classification metrics like F1 score and Matthew’s Correlation Coefficient (MCC) are also included in supplementary.

We feel like ROCAUC and NEF10 are best fit to our problem as they measure the extent to which a model prioritizes the active compounds in its ranking. The definition of ROCAUC is simply the area under the ROC curve, and the definition of NEF10 is based on enrichment factor(EF):

$$\text{EF10}_i = \frac{\sum_{j \in B} z_{ij}}{|B|} / \frac{\sum_{j=1}^n z_{ij}}{n} \quad (23)$$

where  $B$  is the top 10% compounds ranked by certain method. Scaling  $\text{EF10}_i$  in (23) with respect to hit rate of each target  $i$  will give us the NEF10 metric:

$$\text{NEF10}_i = (1 + \frac{\text{EF10}_i - \text{EFbase}}{\text{EF10}_{\max} - \text{EFbase}}) / 2 \quad (24)$$

here  $\text{EFbase} = 1$  corresponding to random guess, and  $\text{EF10}_{\max}$  is the maximum theoretical

value of EF10<sub>i</sub>. Similar to ROCAUC, NEF10<sub>i</sub> value is between 0 and 1 with random guess at 0.5. Although ROCAUC is known to provide amplified scores, it is still a good fit to our problem, as the goal of informer set ranking methods is to rank the compounds and maximize the positive proportion in a fixed size top set. It is a ranking problem instead of a classification problem. Actually we are not asked to give any predictions. From this point of view, NEF10 (Enrichment Factor) is the best criteria, however, NEF10 is a measurement fixed at some ratio, say 10% in ranked compounds. If we want to evaluate ranking methods under different ratios of top set, ROCAUC is a good choice to summarize this information. We will show that BOISE and Accelerated-BOISE are better than previous methods on both data sets under these metrics:

#### 4.1 PKIS1 Data Set

PKIS1 data set is a public human kinase chemogenomics data set, originally downloaded from [https://www.ebi.ac.uk/chembl/db/extra/PKIS/PKIS\\_screening\\_data.csv](https://www.ebi.ac.uk/chembl/db/extra/PKIS/PKIS_screening_data.csv). After pre-processing as described in Zhang, Ericksen, and Lee *et al.* (2019), the PKIS1 contains the inhibition values obtained at 1.0  $\mu$ M concentrations. There are 224 rows and 366 columns in PKIS1 data matrix, with each row represents an unique kinase target and each column corresponding to an unique compound. The inhibition values are continuous values, and we derive the binary kinase inhibition matrix by the target-wise 2SD criteria stated in (1). When a new kinase target enters the study, our goal is to rank the compounds in PKIS1 from most likely to be active to least likely to be active, with the help of a fixed length informer set  $A$  of compounds that we can get the test result on this new kinase target. Similar to the previous paper, we conduct the same prospective and retrospective analysis based on PKIS1 data set, and results are as following:

### 4.1.1 Prospective Results

We apply BOISE and Accelerated-BOISE on three novel kinase targets not included in PKIS1 target sets. They are microbial targets and hence phylogenetically distant from most human protein kinase contained in PKIS1 data set. These three new targets are PknB (*Mycobacterium tuberculosis* kinase), BGLF4 (*Epstein-Barr virus* kinase) and ROP18 (*Toxoplasma gondii* kinase).

The task for informer based ranking methods is to prioritize which PKIS1 compounds might be active on these three targets with the inhibition value obtained on an informer set of size 16. These targets are already screened in advance, yet the screening data are held separately from all methods before informer selection. After selecting the informer compounds, screening data only for those selected informer compounds are provided to each method to rank the whole compound set. To evaluate the performance of different methods, all of the available PKIS1 compounds are tested, and different methods are compared with respect to numbers of hits recovered in their top 10% of ranked compounds for all three targets. Based on the screening results, the active thresholds defined by 2SD criteria (1) are: 13.4% for PknB, 20.2% for BGLF4 and 43.8% for ROP18.

However, these experimentally determined thresholds contain information from non-informer compounds and should be kept away from all methods. In order to binarize the inhibition values to get the intermediate data  $x_A$  for BOISE and Accelerated-BOISE, we focus on the sub-matrix of inhibition data  $Z$  on informer compounds, say  $Z_A$ , and keep track of all the possible binary outcomes  $x_l$  of each row of  $Z_A$ . For rows with same binary outcome  $x_l$ , we compute a centroid corresponding to this outcome, say  $c_l$ , then the binary outcome  $x_A$  of new target with inhibition values  $z_A$  is defined as:

$$\begin{aligned} x_A &= x_{l^*} \\ s.t. \quad ||z_A - c_{l^*}|| &= \underset{l}{\operatorname{argmin}} ||z_A - c_l|| \end{aligned} \tag{25}$$



Procedure (25) is exactly the same as CS and AS. There may be better ways to derive intermediate data, however, we follow the same procedure since we want to show our better performance comes from BOISE and Accelerated-BOISE themselves but not from a more accurate intermediate data. The result is summarized in following Table 1:

Table 1: **Recovery counts by different informer based ranking methods on new kinase targets PknB, BGLF4 and ROP18 using PKIS1 data set.**

Targets	Matrix	Baselines			Non-baselines				total
		BC <sub>w</sub>	BF <sub>w</sub>	RS	CS	AS	AC-BOISE	BOISE	
PKnB	PKIS1	1	7	7	2	3	<b>7</b>	<b>7</b>	8
BGLF4	PKIS1	3	9	3	7	10	<b>10</b>	<b>10</b>	11
ROP18	PKIS1	4	7	4	4	2	<b>7</b>	<b>7</b>	16

In Table 1, the total number of experimentally determined active compounds is indicated in the *total* column. The values below each of the methods indicate the number of experimentally determined active compounds that were ranked in the top 10% of predicted active compounds by each method.

It can be shown that RS, CS or AS can be better than baseline methods BF<sub>w</sub> on some targets, but fail to beat the baseline on others, however, BOISE and Accelerated-BOISE are consistently better than baseline and RS, AS and CS methods on all 3 new targets. It may worth to say that baseline method is performing pretty well on these three new targets for two reasons: One is that baselines are allowed to utilize chemical structure similarity information of compounds in ranking, while non-baseline methods are not allowed to use addition information except inhibition matrix. The other is, as stated earlier, these three targets are all microbial targets that are far away from targets in PKIS1 phylogenetically. Therefore, inhibition data in PKIS1 is less informative than chemical structure similarity on these three targets, and (7, 10, 7) recovery seems to be the best informer based ranking methods can do with PKIS1 data set. We can see a more significant difference among baseline methods, RS/CS/AS methods, and (AC-)BOISE in a retrospective analysis still on PKIS1.

### 4.1.2 Retrospective Results

We conduct retrospective leave one out cross validation for each one of 224 PKIS1 targets. Informer size of  $n_A = 8$  and  $n_A = 16$  are applied to compare performance of different methods under different informer size settings. Results from 224 separate experiments are summarized in violin plots.

When informer set size  $n_A = 8$ , Figure 1 summarizes cross validation results with respect to NEF10 and ROCAUC metrics. From pairwise tests, we can see that BOISE and Accelerated-BOISE are both significantly better than other baseline and non-baseline methods, and BOISE performs similarly to Accelerated-BOISE. Although BOISE is still slightly better than Accelerated-BOISE overall, as it is optimal theoretically, their difference is not statically significant. To be more specific, it can be seen from Figure 1 that Adaptive Selection (AS8) is the best method other than (AC-)BOISE8 under NEF10 metric. From paired t-test between AS8 and (AC-)BOISE8, the mean difference between BOISE8 and AS8 is 0.0608, with p-value  $< 2.2 \times 10^{-16}$ ; the mean difference between AC-BOISE8 and AS8 is 0.0520, with p-value  $= 6.466 \times 10^{-14}$ . Similarly, Coding Selection (CS8) is the best method other than (AC-)BOISE8 under ROCAUC. The mean difference between BOISE8 and CS8 is 0.0283, with p-value  $= 3.73 \times 10^{-6}$  and the mean difference between AC-BOISE8 and CS8 is 0.0258, with p-value  $= 3.035 \times 10^{-5}$ .

Same thing happens when we boost informer set size from  $n_A = 8$  to 16. We rely on distributed computing offered by CHTC to complete this cross validation as BOISE will take a lot of time on personal computer. Empirical results summarized in Figure 2 show that BOISE and AC-BOISE are still significantly better than previous methods: Under NEF10 metric, Adaptive Selection (AS16) is the best method other than (AC-)BOISE and the paired t-test shows that the mean difference between BOISE and AS is 0.0421 with p-value  $= 6.115 \times 10^{-7}$ ; Under ROCAUC metric, Regression Selection (RS16) is the best method other than (AC-)BOISE and the paired t-test shows that the mean difference between BOISE and RS is 0.0294 with p-value  $= 1.303 \times 10^{-4}$ .

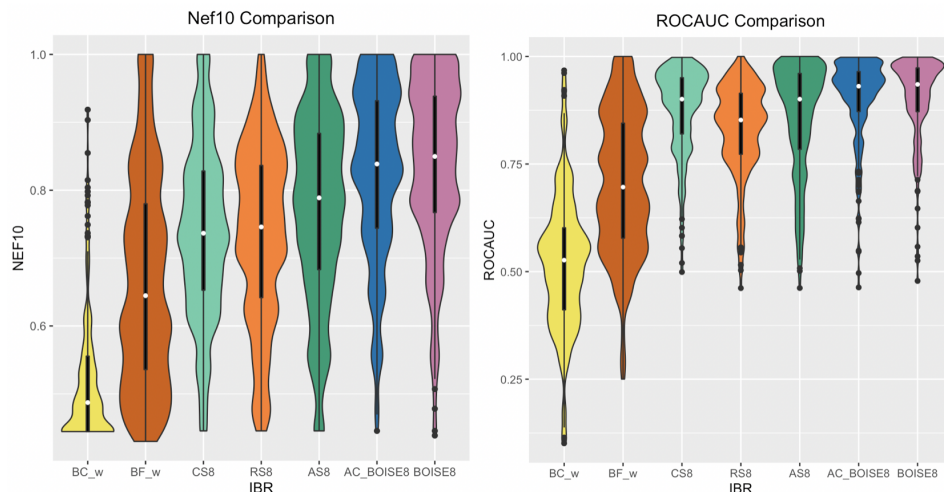


Figure 1: **Violin plot comparison of different methods under informer size  $n_A = 8$  with respect to compound ranking performance assessed by NEF10 and ROCAUC.** Each method is evaluated through leave one target out cross validation of PKIS1 targets. NEF10 is a measurement of active compounds enrichment in top 10% ranked compounds, after normalization with regard to maximum theoretical enrichment that could be achieved at the 10% threshold for the target of interest. NEF10 and ROCAUC of 0.5 indicates a random guess, and of 1.0 represents ideal ranking with all active compounds prioritized above the inactives. The median and interquartile ranges are displayed as a white circle and black bars, respectively.

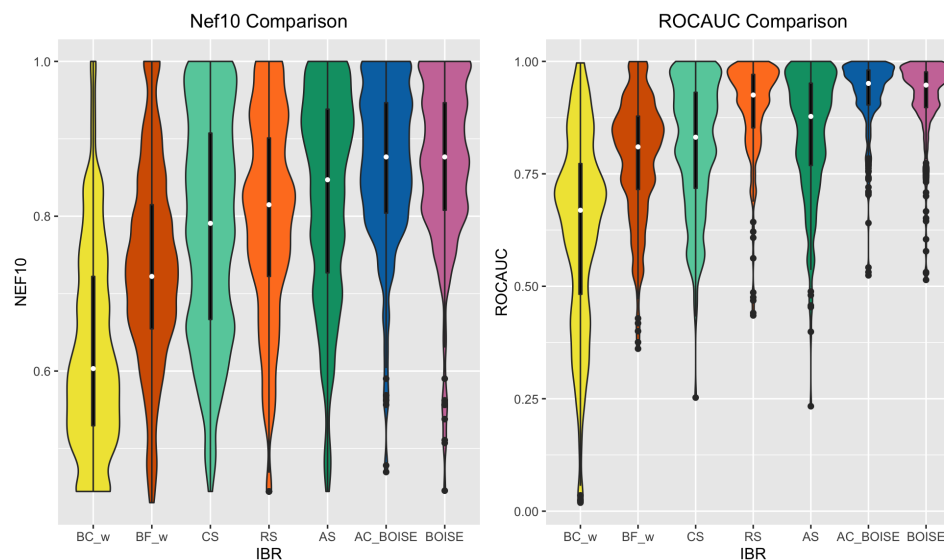


Figure 2: **Violin plot comparison of different methods under informer size  $n_A = 16$  with respect to compound ranking performance assessed by NEF10 and ROCAUC.**

Moreover, there are some concerns about ROCAUC metric, saying that it will amplify the scores. To address these concerns, we include other well-known classification criteria like Matthews Correlation Coefficient (MCC) and F1 Score into our comparison. It turns out different criteria will do no harm to BOISE superiority.

Recall that definition of MCC is:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (26)$$

and definition of F1 Score is:

$$F1 = \frac{2TP}{2TP + FN + FP} \quad (27)$$

where TP, FP, TN, FN come from confusion matrix. MCC takes value in  $[-1, 1]$ , with MCC of 0 indicates a random guess; F1 Score takes value in  $[0, 1]$ , with F1 Score of 0.5 indicates a random guess. Notice that MCC and F1 Score are both for binary classification, yet we are doing ranking for compounds, we will just choose the best split point for our ranked compounds to calculate the classification criteria.

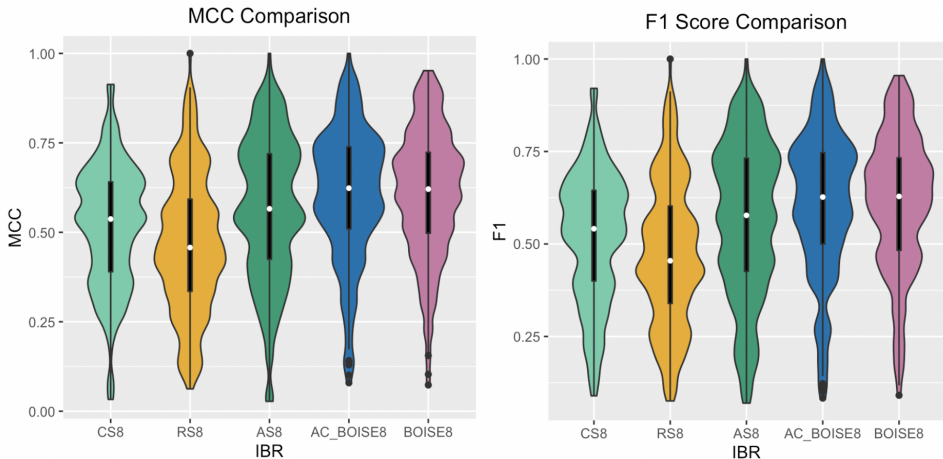


Figure 3: **Violin plot comparison of different methods under informer size  $n_A = 8$  with respect to performance assessed by Matthews Correlation Coefficient (MCC) and F1 Score.**

Figure 3 shows that (AC)-BOISE are still best under these classification criteria: Adaptive Selection (AS8) is the best method other than (AC)-BOISE under MCC and F1 Score. When caring about MCC, the mean difference between AS8 and BOISE8 is 0.0436 with  $p\text{-value} = 3.722 \times 10^{-6}$ ; and mean difference between AS8 and AC-BOISE8 is 0.0401 with  $p\text{-value} = 2.839 \times 10^{-5}$ ; When caring about F1 Score, the mean difference between AS8 and BOISE is 0.0480 with  $p\text{-value} = 1.637 \times 10^{-6}$ ; and mean difference between AS8 and AC-BOISE is 0.0433 with  $p\text{-value} = 1.783 \times 10^{-5}$ .

## 4.2 GDSC Data Set

GDSC data set is a drug sensitivity data set in cancer study. The original data is downloaded from `ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/GDSC1_fitted_dose_response_25Feb20.xlsx`. It contains sensitivity data of each pair between anti-cancer drugs and cancer cell lines. The data set has already been pre-processed, and we care about the Z-score between each drug and each cancer cell line, with  $Z\text{-score} \leq -2$  meaning the cancer cell line is sensitive to the drug. In other words, GDSC data set can be binarized without using 2SD criteria. During data cleaning, we also drop drugs and cell lines with too many missing values and finally we end up with a  $281 \times 207$  GDSC data matrix, with each row as a cancer cell line and each column as an anti-cancer drug. Each entry of the data matrix is continuous sensitivity Z-score, and we derive the binary sensitive matrix by simple threshold  $\mathbb{1}(z_{ij} \leq -2)$ .

Similar to PKIS1 data set, our goal is still to rank the anti-cancer drugs in GDSC when a new cancer cell line comes in, with help of a fixed length informer set  $A$ . As there is no new cancer cell line test data available currently, we only conduct leave-one-out retrospective cross validation on GDSC data set. Also, since Coding Selection (CS) method is too computational extensive and performance of Coding Selection can be bounded by Adaptive Selection (AS) and Regression Selection (RS) from previous analysis, we choose not to include CS in our comparison on GDSC data set.

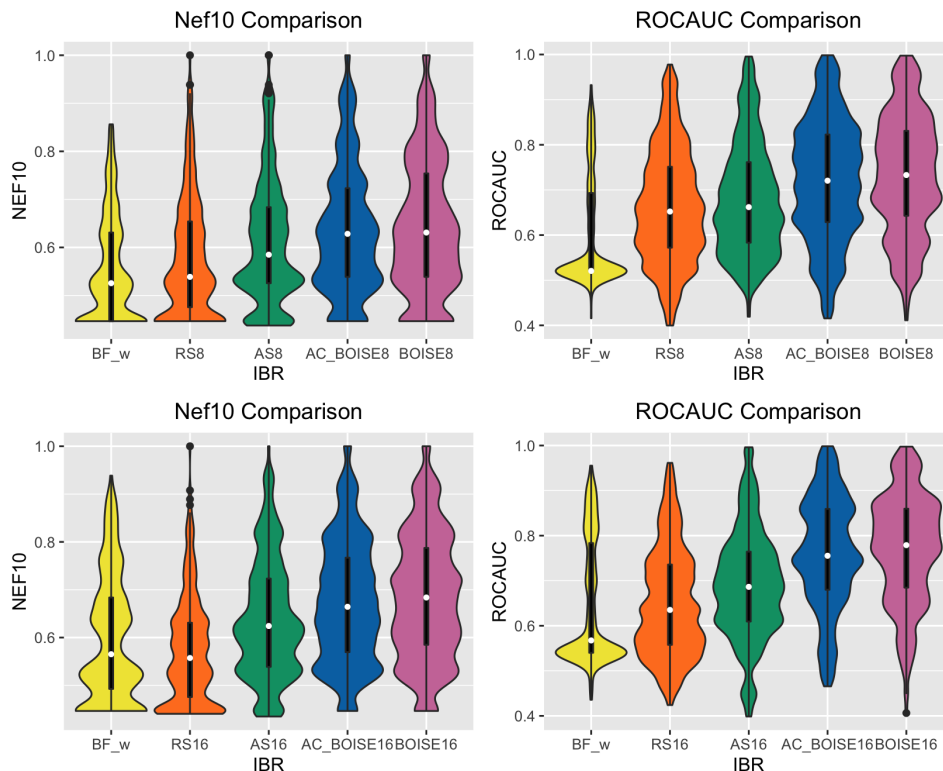


Figure 4: **Violin plot comparison of different methods under informer sizes  $n_A = 8$  and  $n_A = 16$  on GDSC data set with respect to compound ranking performance assessed by NEF10 and ROCAUC.** Each method is evaluated through leave one out cross validation of GDSC cancer cell lines. The median and interquartile ranges are displayed as a white circle and black bars, respectively.

In figures 4 we show that BOISE and AC-BOISE are better than previous methods on the new GDSC data set, with informer size  $n_A = 8$  and  $n_A = 16$ . In particular, Adaptive Selection (AS8 and AS16) is the best method other than (AC-)BOISE. When  $n_A = 8$ , paired t-test shows that the mean difference in NEF10 between BOISE8 and AS8 is 0.0315 with p-value =  $1.743 \times 10^{-4}$ ; the mean difference in ROCAUC between BOISE8 and AS8 is 0.0493 with p-value =  $3.756 \times 10^{-10}$ ; When  $n_A = 16$ , paired t-test shows that the mean difference in NEF10 between BOISE16 and AS16 is 0.0448 with p-value =  $2.837 \times 10^{-8}$ ; the mean difference in ROCAUC between AC-BOISE16 and AS16 is 0.0781 with p-value  $< 2.2 \times 10^{-16}$ .

## 5 Discussion

The idea of BOISE comes from Bayes decision theory (Parmigiani and Inoue 2009) and adaptive clinical trial design (Berry 2006). In section 12.3 of Parmigiani and Inoue (2009), a general procedure is formalized to solve two-stage finite decision problems: The second stage decision problem should be considered first, assuming that action and outcome in previous stage are given. Then we can go back to first stage decision problem with the help of optimal action in second stage. If we formalize the multistage decision problem as a decision tree, as Parmigiani and Inoue did in their book, this procedure seems like a recursion from its leaves to its root. In BOISE, informer set selection is the first stage problem and top set selection (or ranking) is the second stage problem. Therefore, we first look for the optimal top set rule  $T^*(A, x_A, x_0)$ , assuming we know the first stage action  $A$  and outcome  $x_A$ , and then we step back to find best informer set rule with optimal top set rule  $T^*(A, x_A, x_0)$ .

When we seek the best informer set rule in first stage problem, ideas of adaptive design are edifying. Berry (2006) has formalized an adaptive trial design idea based on Bayesian predictive distribution, and this idea finally develops into I-SPY 2 trials (Barker *et al.* 2009). In these clinical trials, drugs (or therapies) will be sequentially evaluated based on their predictive distribution. Every time when new patients come into study, decisions are made to drop, graduate, or continue testing on certain drugs. Evaluating drugs based on predictive distribution is to calculate the loss of all possible outcomes in the future, then take the average with respect to their conditional probability given previous outcomes. In BOISE, this idea is used to find optimal informer set  $A$ , where we pretend we have the future test result on inform set  $A$ , that is  $x_A$ , and take the integral of future loss  $PEL_2(x_0, x_A, A)$ , with respect to conditional distribution  $p(x_A | x_0)$ . An adaptive design idea is proved to be more effective than standard designs in clinical trials, and now is also shown to be promising in informer set selection problems.

One important issue in BOISE computation is MCMC sampling from Dirichlet Process. For conjugate models like in (2), we can collapse the computation by integrating parameters

out and MCMC will mix well. One way to check for mixing properties of the Markov chain in DPMM is to check for the chain over numbers of clusters. Figure 5 shows the plot and estimated auto-correlation of this chain over cluster numbers on PKIS1 and GDSC.

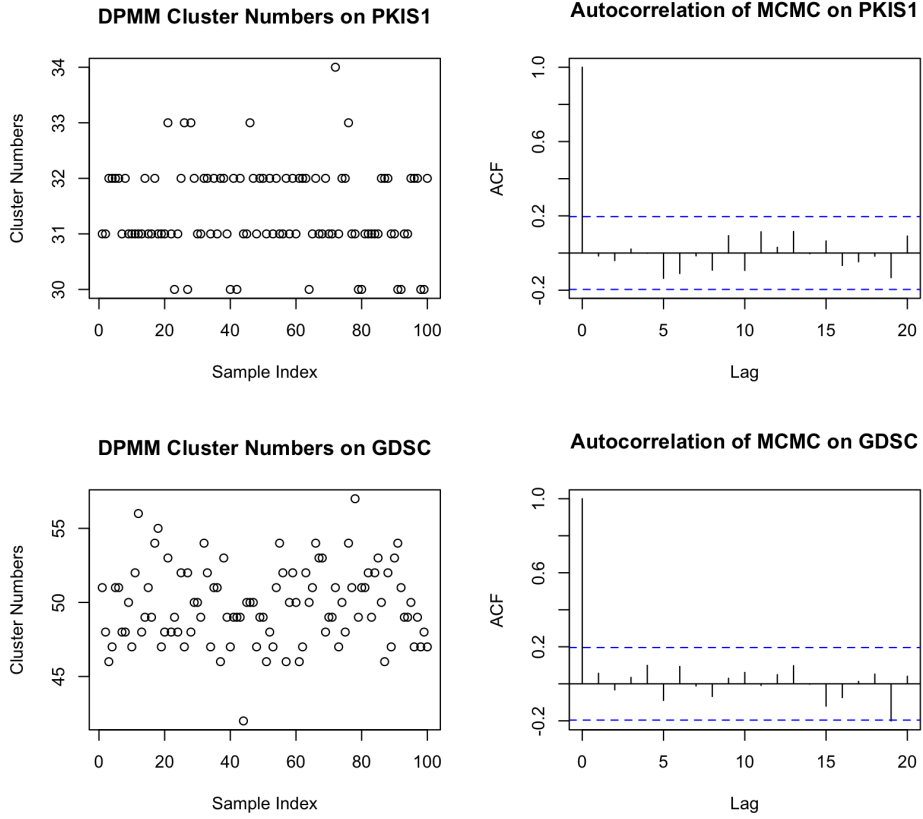


Figure 5: **Mixing properties of MCMC sampling over clusters on PKIS1 and GDSC** The plots are based on the chain over numbers of clusters in each DPMM sample.

From the ACF plots in Figure 5, Markov chains on PKIS1 and GDSC both mix well as there is tiny auto-correlation among consecutive samples. However, from left two plots of cluster numbers in each sample in Figure 5, Markov chain on PKIS1 seems to be more stable than it on GDSC, and we believe this is caused by the nature of the GDSC data set since the large volatility still remains when we run a longer MCMC or change the prior mass. The volatility also corresponds to the difference between performances of BOISE on PKIS1 and GDSC: By comparing Figures 1, 2 with Figure 4, we can see that overall BOISE perform better on PKIS1 data set than on GDSC. This difference between PKIS1 and GDSC



seems to be caused by some overwhelming data points that can either be assigned to some existing clusters or form a new cluster. In the future, we may allow for multiple cluster assignments for a single data point instead of forcing it to stay in only one cluster, and there is a corresponding extension of Chinese Restaurant process, named as Indian Buffet process (Griffiths and Ghahramani 2011), that admits multiple labels in DPMM.

## A Proofs

### A.1 Proof of Theorem 1

**Proof.**

Theorem 1 claims that for any given informer  $A$  and intermediate data  $x_A$ , there exists a top-set rule

$$T^*(A, x_A, x_0) = \left\{ j \in J : \sum_{k \in J} \mathbb{1}(\hat{\theta}_{i^*k} > \hat{\theta}_{i^*j}) \leq n_T \right\}$$

such that  $r(A, T) \geq r(A, T^*)$ . In other words, given  $A$  and  $x_A$ , the best top-set rule is to select top  $n_T$  compounds with largest posterior mean. Recall the expression of Bayes risk:

$$\begin{aligned} r(A, T) &= \int_{x_0} \int_{x_A} \int_{\Theta} L(A, T) p(\theta, x_A, x_0) d\theta dx_A dx_0 \\ &= \int_{x_0} \int_{x_A} \int_{\Theta} L(A, T) p(\theta|x_A, x_0)p(x_A|x_0)p(x_0) d\theta dx_A dx_0 \end{aligned}$$

To prove the Theorem, it is sufficient to show that for any possible fixed  $x_A, x_0$ , inequality

$$\int_{\Theta} L(A, T) p(\theta|x_A, x_0) d\theta \geq \int_{\Theta} L(A, T^*) p(\theta|x_A, x_0) d\theta \quad (28)$$

holds, and this is obvious if we write loss function  $L(A, T)$  explicitly:

$$\begin{aligned}
\int_{\Theta} L(A, T) p(\theta|x_A, x_0) d\theta &= \int_{\Theta} (n_T - \sum_{j \in T} \theta_{i^*j}) p(\theta|x_A, x_0) d\theta \\
&= n_T - \sum_{j \in T} \hat{\theta}_{i^*j} \\
&\geq n_T - \sum_{j \in T^*} \hat{\theta}_{i^*j} \\
&= \int_{\Theta} L(A, T^*) p(\theta|x_A, x_0) d\theta
\end{aligned}$$

The equality comes from previous definition that  $\hat{\theta}_{i^*j} = \mathbb{E}(\theta_{i^*j}|x_A, x_0)$ . ■

## A.2 Proof of Theorem 2

We first prove a lemma that will be used in proof of Theorem 2:

**Proposition 1** *In BOISE model (2), if we use  $x_A$  and  $x_k$  to denote subsets of outcomes on the new target:  $x_{i^*A}$  and  $x_{i^*k}$ , and  $x_{0A}$  and  $x_{0k}$  to denote subsets of outcomes on the original data set  $x_0$ , then we have:*

$$x_A \perp\!\!\!\perp x_k \mid x_0. \quad (29)$$

**Proof of Proposition 1.** Notice that in model (2) we have:

$$\begin{aligned}
(x_A, x_k) &\perp\!\!\!\perp x_0 \mid (\theta_{i^*A}, \theta_{i^*k}); \\
x_A &\perp\!\!\!\perp x_k \mid (\theta_{i^*A}, \theta_{i^*k}); \\
x_{0A} &\perp\!\!\!\perp x_{0k} \mid (\theta_{i^*A}, \theta_{i^*k}); \\
\theta_{i^*A} &\perp\!\!\!\perp \theta_{i^*k}; \\
x_{0A} &\perp\!\!\!\perp x_{0k}.
\end{aligned} \quad (30)$$

From independence and conditional independence in (30), we can show that  $\theta_{i^*A} \perp\!\!\!\perp \theta_{i^*k} \mid x_0$

as following:

$$\begin{aligned}
p(\theta_{i^*A}, \theta_{i^*k} | x_0) &= \frac{p(x_0 | \theta_{i^*A}, \theta_{i^*k})p(\theta_{i^*A}, \theta_{i^*k})}{p(x_0)} \\
&= \frac{p(x_{0A}, x_{0k} | \theta_{i^*A}, \theta_{i^*k})p(\theta_{i^*A}, \theta_{i^*k})}{p(x_{0A}, x_{0k})} \\
&= p(\theta_{i^*A} | x_{0A})p(\theta_{i^*k} | x_{0k}) \\
&= p(\theta_{i^*A} | x_0)p(\theta_{i^*k} | x_0).
\end{aligned}$$

Therefore, for desired  $p(x_A, x_k | x_0)$  we have:

$$\begin{aligned}
p(x_A, x_k | x_0) &= \int_{\Theta} p(x_A, x_k | \theta_{i^*A}, \theta_{i^*k}, x_0)p(\theta_{i^*A}, \theta_{i^*k} | x_0)d\theta_{i^*A}d\theta_{i^*k} \\
&= \int_{\Theta} p(x_A | \theta_{i^*A})p(x_k | \theta_{i^*k})p(\theta_{i^*A} | x_0)p(\theta_{i^*k} | x_0)d\theta_{i^*A}d\theta_{i^*k} \quad (31) \\
&= p(x_A | x_0)p(x_k | x_0).
\end{aligned}$$

and this is equivalent to  $x_A \perp\!\!\!\perp x_k | x_0$ . ■

With the help of Proposition 1 we can now prove Theorem 2:

### Proof of Theorem 2.

When  $x_0$  is given, we can write  $\text{PEL}_1(x_0, A)$  as  $\text{PEL}_1(A)$  for simplicity. To prove Theorem 2 we only need to show that, for any set  $A \subset J$ , and any  $k \notin A$ ,  $\text{PEL}_1(A \cup \{k\}) \leq \text{PEL}_1(A)$ .

For simplicity of notation, we denote top set rule as  $T_1^*(x_A, x_k) = T^*(A \cup \{k\}, x_{A \cup \{k\}}, x_0)$  and  $T_2^*(x_A) = T^*(A, x_A, x_0)$ . Recall that:

$$\text{PEL}_1(A \cup \{k\}) = \int_{x_A} \int_{x_k} p(x_A, x_k | x_0) \left[ \sum_{j \in T_1^*(x_A, x_k)} 1 - \mathbb{E}(\theta_{i^*,j} | x_0, x_A, x_k) \right] dx_A dx_k$$

Here we still use  $x_A$  and  $x_k$  to denote  $x_{i^*A}$  and  $x_{i^*k}$ , a subset of outcomes on the new target  $i^*$ .

From the fact that  $T_1^*(x_A, x_k)$  is the Bayes rule when  $A \cup \{k\}$ ,  $x_{A \cup \{k\}}$  and  $x_0$  are given, it

should be no worse than  $T_2^*(x_A)$  with regard to posterior expected loss, and hence we have:

$$\begin{aligned} & \int_{x_A} \int_{x_k} p(x_A, x_k | x_0) \left[ \sum_{j \in T_1^*(x_A, x_k)} 1 - \mathbb{E}(\theta_{i^*,j} | x_0, x_A, x_k) \right] dx_A dx_k \\ & \leq \int_{x_A} \int_{x_k} p(x_A, x_k | x_0) \left[ \sum_{j \in T_2^*(x_A)} 1 - \mathbb{E}(\theta_{i^*,j} | x_0, x_A, x_k) \right] dx_A dx_k \end{aligned}$$

From Proposition 1,  $x_A \perp\!\!\!\perp x_k | x_0$ , and hence we have:

$$\begin{aligned} & \int_{x_A} \int_{x_k} p(x_A, x_k | x_0) \left[ \sum_{j \in T_2^*(x_A)} 1 - \mathbb{E}(\theta_{i^*,j} | x_0, x_A, x_k) \right] dx_A dx_k \\ & = \int_{x_A} p(x_A | x_0) \sum_{j \in T_2^*(x_A)} \left[ \int_{x_k} p(x_k | x_0) (1 - \mathbb{E}(\theta_{i^*,j} | x_0, x_A, x_k)) dx_k \right] dx_A \\ & = \int_{x_A} p(x_A | x_0) \sum_{j \in T_2^*(x_A)} [1 - \mathbb{E}(\theta_{i^*,j} | x_0, x_A)] dx_A. \end{aligned}$$

The first equality is because  $T_2^*(x_A)$  is a function only of  $x_A$ .

This implies  $\text{PEL}_1(A \cup \{k\}) \leq \text{PEL}_1(A)$ , and the equality holds if and only if  $T_1^*(x_A, x_k) = T_2^*(x_A)$ , which means top sets are the same for any possible outcomes  $x_A$  and  $x_k$ .

■

### A.3 Frequent hitters rule

There are some researchers feeling that informer set is simply the most frequent hitters in compounds. We now show that even under the simplest setting frequent hitters may still not be the best informer set.

Under our model (2), if we eliminate clustering structure for simplicity, that is:

$$\begin{aligned} x_{i,j}|\theta_j &\sim_{ind} \text{Bernoulli}(\theta_j); \\ \beta_j &\sim_{i.i.d} \text{Beta}(\alpha, \beta) \end{aligned}$$

Let  $s_k = \sum_{i=1}^n x_{ik}$  be column sums. We can assume that  $s_1 \leq s_2 \leq \dots \leq s_m$ . The posterior distribution is:

$$\begin{aligned} \theta_j|x_0 &\sim \text{Beta}(\alpha + s_j, \beta + n - s_j) \\ \theta_j|x_0, x_A &\sim \begin{cases} \text{Beta}(\alpha + s_j, \beta + n - s_j), & \text{if } j \notin A, \\ \text{Beta}(\alpha + s_j + x_j^*, \beta + n + 1 - s_j - x_j^*), & \text{if } j \in A, \end{cases} \end{aligned} \quad (32)$$

and the posterior expectation is:

$$E(\theta_j|x_0, x_A) = \begin{cases} (\alpha + s_j)/(\alpha + \beta + n), & \text{if } j \notin A, \\ (\alpha + s_j + x_j^*)/(\alpha + \beta + n + 1), & \text{if } j \in A, \end{cases} \quad (33)$$

We first prove a general proposition:

**Proposition 2** *Given  $n_A, n_T$  and the setting above, if for different informer sets  $A_1$  and  $A_2$ , for any 0-1 sequence  $\nu$ ,  $T^*(A_1, x_{A_1} = \nu, x_0) = \hat{T}(A_2, x_{A_2} = \nu, x_0) \equiv T$ , i.e. Top set is constant w.r.t. any results of two different informer sets. Then  $\text{PEL}_1(x_0, A_1) = \text{PEL}_1(x_0, A_2)$ .*

**Proof.** Let  $\Omega = \{\text{All the 0-1 sequences of length } n_A\}$  be the sample space of  $x_{A_1}$ . Then

from Fubini we have:

$$\begin{aligned}
n_A - \text{PEL}_1(x_0, A_1) &= \int_{\Omega} P(x_{A_1} = s | x_0) \sum_{j \in T} E(\theta_j | x_0, x_{A_1} = s) \, ds \\
&= \sum_{j \in T} \int_{\Omega} \int_{[0,1]} \theta_j \times \frac{P(\theta_j, x_0, x_{A_1} = s)}{P(x_0, x_{A_1} = s)} \times \frac{P(x_{A_1} = s, x_0)}{P(x_0)} \, d\theta_j \, ds \\
&= \sum_{j \in T} \int_{[0,1]} \int_{\Omega} \theta_j P(\theta_j, x_{A_1} = s | x_0) \, ds \, d\theta_j \\
&= \sum_{j \in T} E(\theta_j | x_0)
\end{aligned}$$

Since the value of  $\text{PEL}_1(x_0, A_1)$  has nothing to do with informer set  $A_1$ ,  $\text{PEL}_1(x_0, A) \equiv \text{Constant}$  for any informer sets. ■

When there are no ties in  $s_j$ 's, we can show that  $\text{PEL}_1(x_0, A)$  is constant risk for any  $A$  under this simple setting:

Since

$$E(\theta_i | x_0, x_i^* = 0, x_k^* = 1) \geq E(\theta_k | x_0, x_i^* = 0, x_k^* = 1), \text{ for } k < i;$$

and

$$E(\theta_i | x_0, x_j^*, x_k^* = 1) > E(\theta_k | x_0, x_j^*, x_k^* = 1), \text{ for } k < i, j \neq i, k.$$

This means: When there is no ties,  $E(\beta_i | x_0, x_A) \geq E(\beta_k | x_0, x_A)$  for any  $k < i$  and  $x_A$ . Therefore, we will choose a constant top set  $T = \{m - n_T + 1, \dots, m - 1, m\}$  w.r.t any informer set  $A$  and any results of  $x_A$ .

From our previous Proposition, when there is no ties, Top set is constant w.r.t. any results of different informer sets, and hence  $\text{PEL}_1(x_0, A)$  is a constant for any  $A$ .

We now show that: when  $n_A = 1$ ,  $n_T = 2$ , and  $s_m > s_{m-1} = s_{m-2}$ ,  $A^* = \{m - 1\}$  is better than the Frequent Hitters Rule  $\hat{A} = \{m\}$  in our framework.

First calculate  $P(x_A|x_0)$ :

$$\begin{aligned}
P(x_A|x_0) &= \int_{\Theta} P(x_A|\theta_1, \dots, \theta_m) P(\theta_1, \dots, \theta_m|x_0) d\theta \\
&= \int_{\Theta_A} P(x_A|\theta_A) P(\theta_A|x_0) d\beta_A \\
&= \frac{1}{\alpha + \beta + n} \times \frac{\Gamma(\alpha + s_A + x_A) \Gamma(\beta + n + 1 - x_A - s_A)}{\Gamma(\alpha + s_A) \Gamma(\beta + n - s_A)}
\end{aligned}$$

Next calculate  $\text{PEL}_1(x_0, A)$ :

Recall  $\hat{A} = \{m\}$  and  $A^* = \{m-1\}$ . Notice that whenever  $x_{\hat{A}} = 0$  or  $1$ ,  $T^*(x_0, x_{\hat{A}}, x_0) = \{m-1, m\}$ .

$$\begin{aligned}
n_T - \text{PEL}_1(x_0, \hat{A}) &= P(x_m^* = 1|x_0) [E(\theta_m|x_0, x_m^* = 1) + E(\theta_{m-1}|x_0, x_m^* = 1)] + \\
&\quad P(x_m^* = 0|x_0) [E(\theta_m|x_0, x_m^* = 0) + E(\theta_{m-1}|x_0, x_m^* = 0)] \\
&= \frac{(\alpha + s_m) + (\alpha + s_{m-1})}{\alpha + \beta + n}
\end{aligned}$$

When  $x_{A^*} = 1$ ,  $T^*(x_0, x_{A^*}, x_0) = \{m-1, m\}$ ; When  $x_{A^*} = 0$ , we have  $T^*(x_0, x_{A^*}, x_0) = \{m, m-2\}$ .

$$\begin{aligned}
n_T - \text{PEL}_1(x_0, A^*) &= P(x_{m-1}^* = 1|x_0) [E(\theta_m|x_0, x_{m-1}^* = 1) + E(\theta_{m-1}|x_0, x_{m-1}^* = 1)] + \\
&\quad P(x_{m-1}^* = 0|x_0) [E(\theta_m|x_0, x_{m-1}^* = 0) + E(\theta_{m-2}|x_0, x_{m-1}^* = 0)] \\
&= \frac{\alpha + s_m}{\alpha + \beta + n} + \frac{\alpha + s_{m-1}}{\alpha + \beta + n} \left( \frac{\alpha + s_{m-1} + 1}{\alpha + \beta + n + 1} + \frac{\beta + n - s_{m-1}}{\alpha + \beta + n} \right)
\end{aligned}$$

Since:

$$\frac{\alpha + s_{m-1} + 1}{\alpha + \beta + n + 1} > \frac{\alpha + s_{m-1}}{\alpha + \beta + n}.$$

We have  $n_T - \text{PEL}_1(x_0, \hat{A}) < n_T - \text{PEL}_1(x_0, A^*)$ , so  $\hat{A}$  is not the Bayes rule.

## References

- Aldous, D. (1985), "Exchangeability and Related Topics," *École d'Été de Probabilités de Saint-Flour XIII-1983*, Springer, Berlin, 1-198.
- Antoniak, C.E. (1974), "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2(6), 1152-1174.
- Barker, A.D., Sigman, C.C., Kelloff, G.J., Hylton, N.M., Berry, D.A., and Esserman, L.J. (2009), "I-SPY2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy", *Clinical Pharmacology & Therapeutics*, 86(1), 97-100.
- Berger, J.O. (1985), *Statistical decision theory and Bayesian analysis*. 2nd ed. Springer-Verlag, New York.
- Berry, D.A. (2006), "Bayesian Clinical Trials," *Nature Reviews Drug Discovery*, 5, 27-36.
- Blackwell, D., and MacQueen, J.B. (1973), "Ferguson Distributions via Pólya Urn Schemes", *The Annals of Statistics*, 1, 353-355.
- Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017), "Stan: A probabilistic programming language," *Journal of statistical software*, 76(1).
- Eriksen, S.S., Wu, H., Zhang, H., Michael, L.A., Newton, M.A., Hoffmann, F.M., and Wildman, S.A. (2017), "Machine Learning Consensus Scoring Improves Performance Across Targets in Structure-Based Virtual Screening," *Journal of Chemical Information and Modeling*, 57(7), 1579-1590.
- Ferguson, T.S. (1983), "Bayesian Density Estimation by Mixtures of Normal Distributions," *Recent Advances in Statistics*, 24, 287-302.
- Ferguson, T.S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209-230.
- Griffiths, T.L., and Ghahramani, Z. (2011), "The Indian Buffet Process: An Introduction and Review," *Journal of Machine Learning Research*, 12, 1185-1224.
- Henderson, N.C., Newton, M.A. (2016), "Making the Cut: Improved Ranking and Selection for Large-scale Inference", *Journal of the Royal Statistical Methodology Series B*, 78(4), 781-804.



- Kitchen, D.B., Decornez, H., Furr, J.R., and Bajorath J. (2004), "Docking and scoring in virtual screening for drug discovery: methods and applications," *Nature Reviews Drug Discovery*, 3(11), 935-949.
- Li, H. and Hong, F., (2001), "Cluster-Rasch models for microarray gene expression data," *Genome biology*, 2(8), research0031-1.
- Neal, R.M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249-265.
- Parmigiani, G. and Inoue, L. (2009), *Decision theory: principles and approaches* (Vol. 812), West Sussex, England: John Wiley & Sons.
- Quinlan, J.R. (1986), "Induction of Decision Trees," *Machine Learning*, 1, 81-106.
- Quinlan, J.R. (1993), *C4.5: Programs for Machine Learning*, San Francisco, CA: Morgan Kaufmann Publishers Inc.
- Reker, D and Schneider, G (2015), "Active-learning strategies in computer-assisted drug design," *Drug Discovery Today*, 20, 458-465.
- Robert, C. (2007), *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.
- Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E.W. (2014), "Computational methods in drug discovery," *Pharmacological Reviews*, 66(1), pp.334-395.
- Zhang, H., Ericksen, S.S., Lee, C., Ananiev, G.E., Wlodarchak, N., Yu, P., Mitchell, J.C., Gitter, A., Wright, S.J., Hoffmann, F.M., Wildman, S.A., and Newton, M.A. (2019), "Predicting kinase inhibitors using bioactivity matrix derived informer sets," *PLoS Computational Biology*, 15(8): e1006813.