Advanced Bayes informer sets for various user scenarios

Peng Yu, AJ Fagan, Spencer S. Ericksen, Scott Wildman, Anthony Gitter and Michael A. Newton

September 30, 2024

Abstract

Virtual screening is an efficient way to search the chemical space in early stage drug discovery, in which the key problem is to prioritize the available compounds for a novel target. Informer-based ranking (IBR) methods solve the problem with the help of potentially relevant targets and a small set of compounds — the so-called “informer set”. So far the best IBR method is Bayes optimal informer set (BOISE), which selects informers and prioritizes the compounds by solving a two-stage decision problem with a flexible model and relevant loss function. However, the scalability of BOISE is constrained by its computation complexity. Considering the trade-off between scalability and accuracy, we introduce a variant of BOISE, fast BOISE, that solves the scaling problem to some degree. We evaluate this BOISE variant and compare it to original BOISE both retrospectively and prospectively, obtaining comparable or improved performance with significant reduction in computation cost. We also apply fast BOISE to a real-world drug discovery data set, PCBA. It scales up smoothly and exhibits better predictive performance than naive informer selection methods.

1 Introduction

The key interest of drug discovery is to discover compounds that interact with specific drug targets. Virtual screening (VS) allows computational methods to operate on available data, guiding the set of compounds tested experimentally in order to reduce the burden of negative experimental results. Established VS approaches depend on structural or bioactivity data that may be unavailable for novel protein targets. An interesting challenge in this domain is how to screen compounds when all that is known about the new target is that it is related to some other targets for which bioactivity has been already measured on some common set of compounds.

Clemons et al. (2021) provides a review of approaches to the aforementioned challenge that involve so-called *informer sets*, which are relatively small subsets of compounds utilized in the first stage of experimentation on a novel target. Aspects of the biological and chemical context certainly guide the choice of methodology for a particular protein target, and the definition of informer set can vary under different contexts. In Zhang et al. (2019), a specific VS framework using informer sets is proposed and named as Informer-Based Ranking (IBR). Figure 1 summarizes the IBR framework: it starts with a targets by compounds matrix of initial bioactivity data, and requires the selection of a small compound subset (the informers) for which data will be collected on the new target. It also prioritizes non-informer compounds for bioactivity predictions against the new target, based upon the aggregate of initial and intermediate data. IBR framework assumes no additional information on targets other than the observed bioactivity matrix and hence is a good fit for novel targets.

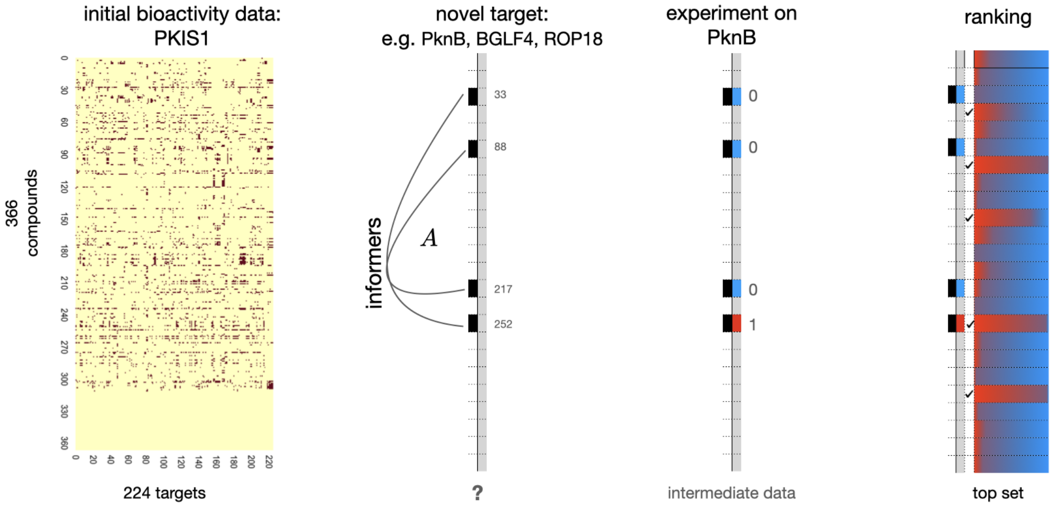

Figure 1: Informer-based-ranking problem: A matrix of binary bioactivity data is available (left; red active, yellow inactive). The problem is to first identify a subset of the n compounds as informer compounds that will be evaluated experimentally on new target, and then to prioritize all the compounds for further testing after intermediate data is obtained.

In Yu et al. (2022), the heuristic formulation is characterized as a statistical decision problem, from which a sophisticated and effective IBR method, the Bayes Optimal Informer SEt (BOISE), emerges. BOISE aims to find an informer set that minimizes the average loss computed on hypothetical intermediate data, where the loss function is designed to penalize the inactive probabilities among top-ranked compounds and the hypothetical data is generated from a predictive distribution given the initial bioactivity data. Such a method achieved a significant improvement in performance criteria when compared to the heuristic methods in the first generation IBR approaches of Zhang et al. (2019).

Despite the superiority over predecessors, it remains a question whether existing BOISE method is applicable to realistic scenarios — the largest data set that BOISE has been tested on is 224 × 366, while in reality thousands of candidate compounds is common in initial bioactivity data. When scaling up to larger data sets, a series of problems could emerge, such as the trade-off between predictive performance and computation cost, the efficiency of model sampling, and possible failures caused by floating-point error. In this paper, we present a BOISE variant, "fast BOISE", aimed to resolve these computational problems. This method permits a decrease in computation time by orders of magnitude with no substantial loss in performance.

2 Methodology for BOISE variants

It may help to first recall the formulation of original BOISE. Suppose $x_0 = (x_{i,j})$ is the $m \times n$ binary initial bioactivity data matrix, and $x_A = \{x_{i^*,j} : j \in A\}$ is the intermediate data derived between novel target i^* and informer set A. BOISE assumes $x_{i,j}$ and $x_{i^*,j}$ are Bernoulli trials governed by parameters $\theta_{i,j}$ and $\theta_{i^*,j}$, respectively. The key assumption of IBR framework is the similarity between targets, and hence a Chinese Restaurant Process (CRP) is applied on targets to enforce information sharing among targets while still retaining flexibility. BOISE models assume, therefore, that there is a partition $\mathcal{C} = \{c_k\}$ of the m initial targets, wherein each cluster c_k contains identically distributed targets in the sense that $\theta_{i,j} = \sum_k \phi_{k,j} \mathbb{1}(i \in c_k)$ for a reduced set of cluster/target parameters $\phi = \{\phi_{k,j}\}$. The proposed modeling is thus:

$$\begin{aligned} \mathcal{C} &\sim \mathrm{CRP}_m(m_0), \\ \phi_{k,j} &\sim \mathrm{Beta}(\alpha_0, \beta_0), \quad k = 1, \cdots, K,\ j = 1, \cdots, n \\ x_{i,j} \mid \mathcal{C}, \phi &\sim \mathrm{Bernoulli}\left\{ \sum_k \phi_{k,j} \mathbb{1}(i \in c_k) \right\}, \quad i = 1, \cdots, m,\ j = 1, \cdots, n. \end{aligned} \tag{1}$$

where Beta(α_0, β_0) is the base distribution and m_0 is the prior mass of CRP.

With generative model in (1), BOISE searches for the informer set A that minimizes the posterior expected loss given initial x_0,

$$L_0(A, T; \theta) = \sum_{j \in T} (1 - \theta_{i^*,j}), \tag{2}$$

where A is the informer set and T is the *top set*, the set of compounds on which we will perform further experimentation in order to identify as many active compounds as possible against novel target i^*. The size of informer and top set, n_A, and n_T, respectively, are predetermined as compounds are purchased and screened in batches in VS. In this work we consider, also, a modified version of this loss function,

$$L_1(A, T; \theta) = \sum_{j \in T} (1 - x_{i^*,j}) \,. \tag{3}$$

The original loss function used in BOISE, L_0, permits odd circumstances where "misses", compounds observed to be inactive in the informer set, are selected for the top set. By replacing the $\theta_{i^*,j}$ with $x_{i^*,j}$, hits in the informer set will always be ranked first, misses will always be ranked last, and all compounds not in the informer set are ranked exactly as they were according to L_0.

The concept of top set origins comes from chemical screening procedure, yet there is an argue around its necessity. The major concern is about the top set size n_T, as when n_T changes, previous informer set will no longer be optimal and another round of training and selection is needed. Although original BOISE is shown to be quite robust for mismatched n_T's in simulation, there is no theoretical guarantee about this robustness. Besides, the desired output for IBR methods is a ranking of all candidate compounds, and top-set-defined losses like (3) only select the top n_T compounds without regard to the ordering of either the top set or the "bottom" set. With such gaps between practice and decision theory, we may abandon top set T in loss functions and reframe it to match the ranking procedure.

Due to the binary nature of bioactivity data (active/inactive), the unknown vector of interactions, x_{i^*}, automatically gives the true ranking of the compounds. The question now is how to define a loss between estimated and true ranking. There are a handful of ranking losses available, as summarized in **?**. Considering the significant number of ties in x_{i^*}, ranking losses based on pairwise comparison are more applicable. Specifically, if an estimated ranking of compounds is given from most likely to least likely to be active on the novel target, it can be represented as a $n \times n$ matrix R, where n is the number of compounds:

$$R(i,j) = \begin{cases} 1, & rank(i) < rank(j) \\ 0 & rank(i) == rank(j) \\ -1, & rank(i) > rank(j) \end{cases} \tag{4}$$

and a new loss function for BOISE could be defined as:

$$L_2(A,R;x_{i^*}) = \sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\{-R(j_1,j_2)(x_{i^*,j_1}-x_{i^*,j_2})\}_+ \tag{5}$$

where function $\phi(z)=(z)_+$ means the positive part of z. Despite the complicated form as it may look like, $L_2(A,R;x_{i^*})$ has a simple interpretation: it is the number of discordant pairs between estimated ranking and true ranking. Therefore, there will be many optimal rankings R^* which all achieve the same minimum of $L_2(A,R;x_{i^*})$.

2.1 Block BOISE

When scaling up BOISE to a real-world screening, the major concern is on the dramatic increase in candidate compounds: typically there is a large number of compounds tested against a relatively small number of targets, resulting in a "wide" bioactivity matrix x_0. In practice, the sampling of clustering structures $\mathcal{C}$ can be a challenge for huge number of compounds. Therefore, the constraint that all compounds share the same target clustering may be eschewed in the presence of a "wide" bioactivity matrix x_0, and replaced with a compound-level *grouping*, where each group has an individual target-level *clustering* structure shared among its members. Suppose $\mathcal{G}=\{g_k\}$ is a partition of n compounds (columns) into K groups, where for each group of compounds $\{j\in g_k\}$, another partition $\mathcal{C}_k=\{c_{k,r}\}$ is applied on m targets into R_k clusters. For simplicity of notations, we use x_k to denote the kth sub-matrix defined by g_k, where x_0 is the whole bioactivity data matrix. A generalization of model (1) can be made as:

$$\begin{aligned}
\mathcal{G} &= \{g_k\} \\
\mathcal{C}_k &\sim \mathrm{CRP}_m(m_0), \quad k=1,\cdots,K \\
\phi_{r,j}\mid j\in g_k &\sim \mathrm{Beta}(\alpha_k,\beta_k), \quad k=1,\cdots,K,\ r=1,\cdots,R_k \\
x_{i,j}\mid j\in g_k,\mathcal{C}_k,\phi &\sim \mathrm{Bernoulli}\left\{\sum_r \phi_{r,j}\mathbb{1}(i\in c_{k,r})\right\}, \quad i=1,\cdots,m,\ j=1,\cdots,n.
\end{aligned} \tag{6}$$

We use the same m_0 in the model but it can be different prior masses for different sub-matrices. It is easy to see that model (6) and (1) are equivalent when $\mathcal{G}=\{\{1,2,\cdots,n\}\}$ is trivial. Since the bioactivity matrix generated from (6) is divided into blocks, the BOISE variant based on it is named as "block BOISE".

2.2 Fast BOISE

An exhaustive search for the optimal size n_A informer set is intractable. BOISE, therefore, utilizes sequential selection by selecting, first, the optimal size 1 informer set, then iteratively adding the next best compound given the current informer set until the desired quantity is reached. This sequential selection becomes problematic as the size of the informer set increases. For $n_A \approx 100$ or 1000, sequential selection can take a huge amount of time. Even if the chemical screening researchers are willing to wait for the selection, it is better to have a fast version that provides a rough estimation of the performance when n_A increases, and that can serve as a reference for the design of experiments. To meet this end, we introduce a class of loss functions that are capable of non-sequential selection:

$$L_f(A) = \sum_{j \notin A} f(j) \tag{7}$$

In the discussion of Yu et al. (2022), an inspection into the posterior expected loss computation shows that the posterior probability $p_k = P(i^* \in c_k \,|\, \mathcal{C}, x_0, x_A)$ may effectively score the informer sets, since $(p_0, p_1, \cdots, p_K)$ constitute the conditional distribution of the cluster label for target i^* given $\mathcal{C}$, x_A and x_0, and posterior expectations $E(\theta_{i^*,j} \,|\, x_0, x_A, \mathcal{C}, i^* \in c_k)$ have a limited role in selecting top set $T^*(A, x_A, x_0)$. An approximate version of BOISE is then developed using the entropy of p, guided by ID3 decision tree algorithm in Quinlan (1986). Suppose random variable $L_{i^*}(\mathcal{C})$ denotes the cluster label for target i^* given clustering structure $\mathcal{C}$, and $X_{i^*,j}$ denotes the interaction of jth compound against target i^*. A reasonable choice of loss is $L_{\tilde{f}}(A)$ where:

$$\tilde{f}(j) = D_{KL}\left[P\left\{L_{i^*}(\mathcal{C}) \;\mid\; X_{i^*,j}\right\} \| P\{L_{i^*}(\mathcal{C})\}\right] \tag{8}$$

where $D_{KL}(\cdot)$ is the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951), $P\{L_{i^*}(\mathcal{C})\}$ is the prior distribution of cluster label for target i^* (that is the CRP prior) and $P\left\{L_{i^*}(\mathcal{C}) \;\mid\; X_{i^*,j}\right\}$ is the conditional distribution of cluster label given the outcome $X_{i^*,j}$. Notice that this conditional distribution is different from previous $p = (p_0, p_1, \cdots, p_K)$, which is conditioned on the whole intermediate data x_A. The KL divergence measures the information gain of $L_{i^*}(\mathcal{C})$ when $X_{i^*,j}$ is known, and the posterior expected loss of $L_{\tilde{f}}(A)$ with regard to (8) can be computed via double expectation:

$$\begin{aligned}
\mathrm{PEL}_1(A) &= E\left[E\left\{\sum_{j\notin A}^{n}\tilde{f}(j)\middle| x_0, x_A\right\}\middle| x_0\right] \\
&= \sum_{j\notin A} E[D_{KL}\left[P\left\{L_{i^*}(\mathcal{C}) \mid X_{i^*,j}\right\}\|P\{L_{i^*}(\mathcal{C})\}\right] \mid x_0] \\
&= \sum_{j\notin A} E\left[E\left[D_{KL}\left(\cdot\right) \mid \mathcal{C}, x_0\right] \mid x_0\right]
\end{aligned}$$

and the inner part of last equation, $E\{D_{KL}(\cdot) \mid \mathcal{C}, x_0\}$, is the mutual information (MI) between $L_{i^*}(\mathcal{C})$ and $X_{i^*,j}$ conditional on $\mathcal{C}$, which is in line with selection criteria in ID3 algorithm when clustering assignment $\mathcal{C}$ is given.

The Bayes rule of revised $L_{\tilde{f}}(A)$ would be to select top n_A compounds j with largest expected mutual information between cluster label $L_{i^*}(\mathcal{C})$ and $X_{i^*,j}$, the potential outcome on j. Although integrated fast BOISE method is not the same as ID3 algorithm which is also a sequential selection with greedy search, the selected informers are most likely to cover the real "key" compounds for large informer size n_A, and that is exactly the scenario under which fast BOISE is developed. In case studies, fast BOISE is shown to be closely related to original BOISE with an example target, and it achieves comparable performance considering its reduction in computation time by orders of magnitude.

3 Empirical studies

3.1 FDA data set

The FDA data set is a newly constructed data set between FDA approved drugs and Pubchem bio-assays (Wang et al., 2016). It contains bioactivity information on 688 targets and 2002 compounds, with a missing rate of 75.8% and active rate of 5.2%. We select compounds from the Selleck FDA set with at least 1 active and inactive label on the frequently tested Pubchem Bioassay targets (tested on $\geq$ 1000 compounds). After that, targets are selected from those frequently tested assays with at least 2 active and inactive labels on the selected compounds. FDA approved drugs are most common starting points in drug discovery process, and we compare variants of BOISE in both retrospective and prospective ways on this FDA data set.

3.1.1 Retrospective analysis

To evaluate and compare BOISE variants retrospectively, 60 targets are randomly selected from FDA data set and tested in a way that the selected target plays the role of i^*, while the other 687 targets provide data x_0. Due to the high missing rate of FDA data set, each sampled target is required to have at least 400 complete responses to become a test target. Informer selection is restricted to non-missing compounds of each test target, as incomplete x_A could lead to unexpected results.

The next experiment evaluates the performance of block BOISE and fast BOISE on FDA data set. To get compound clustering $\mathcal{G}$ in block BOISE, the divisive analysis clustering (DIANA) is applied on Morgan (ECFP) fingerprints of compounds with 2048 bit length and radius$=3$ (Rogers and Hahn, 2010). DIANA is a top-down approach of hierarchical clustering where all data points are initially assigned a single cluster and then split into groups (Kaufman and Rousseeuw, 2005). The chemical fingerprint of a compound is a fixed length binary vector (2048 in this case) where each position indicates the presence/absence of certain molecule. We apply Jaccard distance on chemical fingerprints to define similarity between two compounds in this experiment, and Silhouette score (Rousseeuw, 1987) is used to determine the number of clusters. After evaluating on $K = 2 \cdots 25$ for DIANA clustering, $K = 23$ achieves highest Silhouette score, thus 2002 FDA compounds are partitioned into 23 groups.

With this partition $\mathcal{G}$, block BOISE and fast BOISE are applied on the same 60 test targets sampled from FDA data set. In this experiment, $L_1(A, T; x_{i^*})$ is used for block BOISE. The left panel of Fig 2 shows average performances of block and fast BOISE without partition $\mathcal{G}$ for $n_A = 1 \cdots 20$, while the right panel shows the average performances with partition $\mathcal{G}$ for $n_A = 1 \cdots 30$.

In Fig 2, "orig-BOISE" refers to the traditional BOISE with loss function $L_1(A, T; x_{i^*})$ because block BOISE and traditional BOISE are equivalent when $\mathcal{G}$ is trivial, and original BOISE stops at informer size 20 due to unacceptable running time and stagnant performances. Such stagnation seems not to happen for block BOISE, even when n_A increases to 30. "Rand-BOISE" is a baseline method where informers are randomly sampled from available compounds, and each point is the average performance of 25 random samples for each test target. BOISE with $L_1(A, T; x_{i^*})$ loss is still the best method under NEF10 metric, while block BOISE steadily achieves better ROCAUC after $n_A = 20$. As for fast BOISE, it performs slightly better without

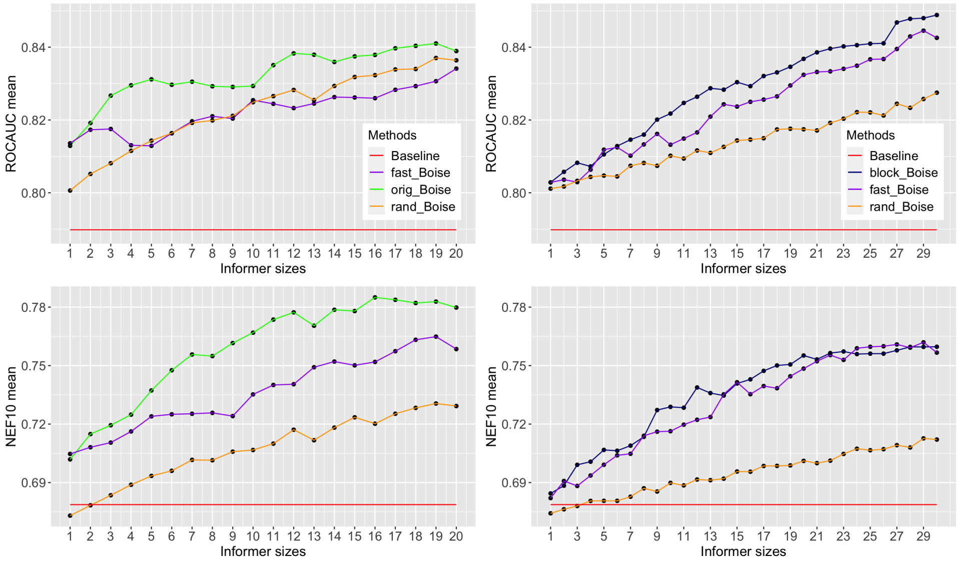

Figure 2: Comparison of block BOISE, fast BOISE and baseline randomly selected informers when n_A increases.

the restriction of compound partition $\mathcal{G}$. Although fast BOISE is not so good as traditional BOISE or block BOISE, it is quite close to block BOISE and can significantly reduce the informer search time, as shown in Table 1.

Table 1: Computation time of posterior expected loss evaluation in different BOISE variants for 20th informer selection. The computation time below refers to the evaluation time on HTCondor for one candidate compound during the search of 20th informer.

BOISE variants	Average time (min)	Max time (min)
orig-BOISE	180	240
block-BOISE	30	45
fast-BOISE	< 1	< 1

In the upper left plot of Fig 2, the ROCAUC of randomly selected informers is sometimes better than fast BOISE without compound clustering $\mathcal{G}$, and this is by accident, as in Fig 3 fast BOISE shows a consistent superiority over randomly selected informers for a larger range of informer sizes. Fig 3 also shows that fast BOISE will achieve satisfying performances when informer size becomes larger.

3.1.2 Prospective analysis

Following the retrospective analysis on FDA data set, we further test the BOISE variants on a few more targets in a prospective way. A prospective analysis can help prevent the look-ahead bias and give us an estimate of BOISE variants performances in a real-world scenario. There are 6 chemical screening labs who test FDA approved drugs on their interested targets, and

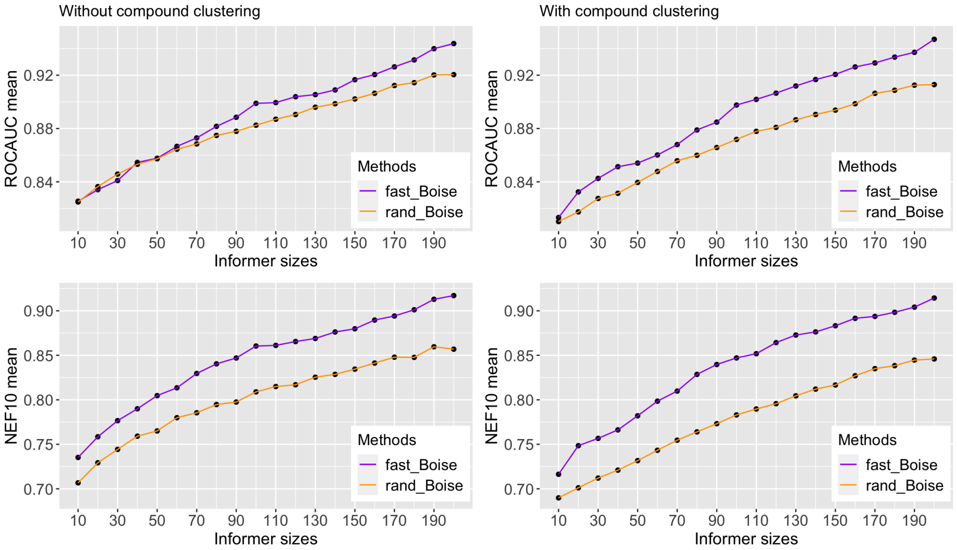

Figure 3: Comparison of fast BOISE and randomly selected informers with and without compound clustering $\mathcal{G}$ when n_A ranging from 10 to 200.

they kindly made 18 novel targets available to us. A brief summary of these novel targets is in Table 2.

Table 2: Summary of 18 novel targets in prospective analysis

Lab name	Target IDs	Target types	Protocol description
Ahlquist	1 to 4	Cell	Yeast growth inhibition assay, OD600 measurement;
Hoskins	5	Cell	AFUMSF3B1 with FDA library, OD600 measurement;
Hull	6 to 10	Spores and fungal cell	10h and 22h Germination with FDA plates; 12h yeast with FDA plates;
Keck	11 to 16	AlphaScreen assay, FP assay and cell	Selleck FDA library screen using AlphaScreen assay of Biotin-SSB-Ct peptide interaction and fluorescence polarization assay of FAM-SSB-Ct peptide interaction with 6x-His Klebsiella pneumoniae DnaG (KpnDnaG) primase and PriA (KpnPriA) helicase;
Senes	17	Protein	Raw alpha screen
Xing	18	Protein	MTDH-SND1 for FDA screen; His-CFP-SND1(16-336), 0.8 uM His-YFP-MTDH(386-407), 12.8 uM compound, 60uM

The prospective test on these 18 targets follows exactly the same procedures as in Figure 1: at the very beginning, nothing is known about the targets except for which FDA compounds will be tested against them. Various informer sets are then selected for each target using different BOISE variants, with informer sizes the same as in retrospective analysis. We provide these informer sets to collaborators and they give us the actual bioactivity responses on each target. After that, rankings of all tested FDA compounds are given based on initial x_0 and intermediate x_A and we evaluate these rankings on the true outcomes.

The performances of BOISE variants in prospective analysis are similar to those in retrospective analysis: the mean and median of NEF10 and ROCAUC are almost the same for original BOISE with $L_1(\cdot)$ loss. Block BOISE performs better prospectively than retrospectively, and original BOISE with $L_2(\cdot)$ loss is slightly worse, probably due to its sensitivity to violation of model assumptions. Block BOISE achieves best average performance on both NEF10 and ROCAUC in this prospective analysis. All three variants mentioned above are better than fast BOISE, although the difference is not significant. After having the full bioactivity data on novel targets, we also test the baseline rand-BOISE where 30 informers are randomly sampled, and it is still worse than all the BOISE variants. Table 3 below shows the number of hits recovered within top 100 ranked compounds under different methods for each of 18 novel targets, and Figure 4 is a summary of ROCAUC and NEF10 in this prospective analysis.

Table 3: Number of hits recovered by different BOISE variants in their top 100 ranked compounds on 18 novel targets using FDA data set. The novel targets are sorted based on their total number of hits in FDA compounds. The recovery counts of rand-BOISE is the average of 25 randomly selected informer sets.

ID	rand-BOISE	fast BOISE	orig BOISE(L2)	block BOISE	orig BOISE(L1)	total
18	4.60	4	4	5	4	6
15	6.04	6	6	6	6	7
14	6.96	7	7	7	7	8
11	5.76	7	7	6	7	9
3	5.76	6	7	8	7	14
12	8.04	10	6	8	11	15
17	8.00	10	10	9	10	16
4	8.08	8	12	12	12	17
16	9.92	11	9	12	11	18
13	7.80	7	9	12	10	20
5	11.84	8	13	17	12	23
1	12.52	17	15	16	18	27
6	11.80	13	13	17	9	28
2	16.40	23	19	21	21	35
10	16.12	18	21	21	20	36
9	13.92	17	16	20	16	38
7	12.28	14	14	17	16	48
8	16.28	14	19	22	19	53

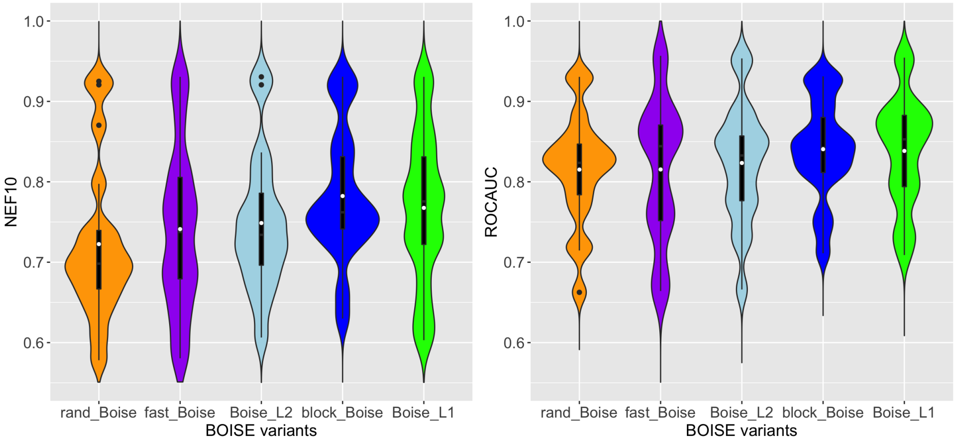

Figure 4: Summary of performances of BOISE variants in prospective analysis. All methods are applied on 18 novel targets, with FDA data set as initial bioactivity data x_0. The informer set sizes are the same as in retrospective analysis, with $n_A = 20$ for original BOISE methods and $n_A = 30$ for block BOISE, fast BOISE and baseline rand-BOISE.

3.2 PCBA data set

The PCBA data set is a bioacitivity data matrix used in real drug discovery research, which is first introduced by Ramsundar et al. (2015) from PubChem library. In this experiment, a condensed version of PCBA is used with most missing values removed due to the limitation of computer memory. The condensed PCBA data contains 102 targets and 134264 compounds, with a missing rate of 5.4% and active rate of 0.7%. We apply fast BOISE method on this PCBA data to further test its scalability and compare it with naive informers that are frequently used by biochemistry researchers. The experiment is conducted in the same retrospective and prospective way as on FDA data set, with 30 targets randomly selected from PCBA data set as test targets, and one novel target used to test fast BOISE prospectively. The results on fast BOISE, as aforementioned, provide the guidance of how other BOISE variants perform and how many informers are enough for PCBA.

Previous experiments on FDA data show that fast BOISE performs slightly better without compound clustering $\mathcal{G}$, thus a target grouping on the whole bioactivity matrix is in favor. However, floating point error due to large number of compounds makes sampling algorithm of CRP extremely inefficient. On that account, we approximate the posterior sampling of CRP by randomizing the distance matrix as suggested in the appendix of Ma et al. (2021). The distance matrix is defined by Jaccard distance on complete responses between two targets, and the number of clusters for a randomized distance matrix is selected by validity score with threshold of 0.9, following the same procedure in the original paper. With this approximate sampling of CRP, we compare fast BOISE with randomly selected informers for $n_A = 100, 200, 500$ and 1000 on 30 test targets. The ROCAUC and NEF1 results are illustrated with scatterplots in Fig 5 and Fig 6, respectively, and the majority of points ($\approx 22/30$) are above the diagonal line, assuring the superiority of fast BOISE over random selections.

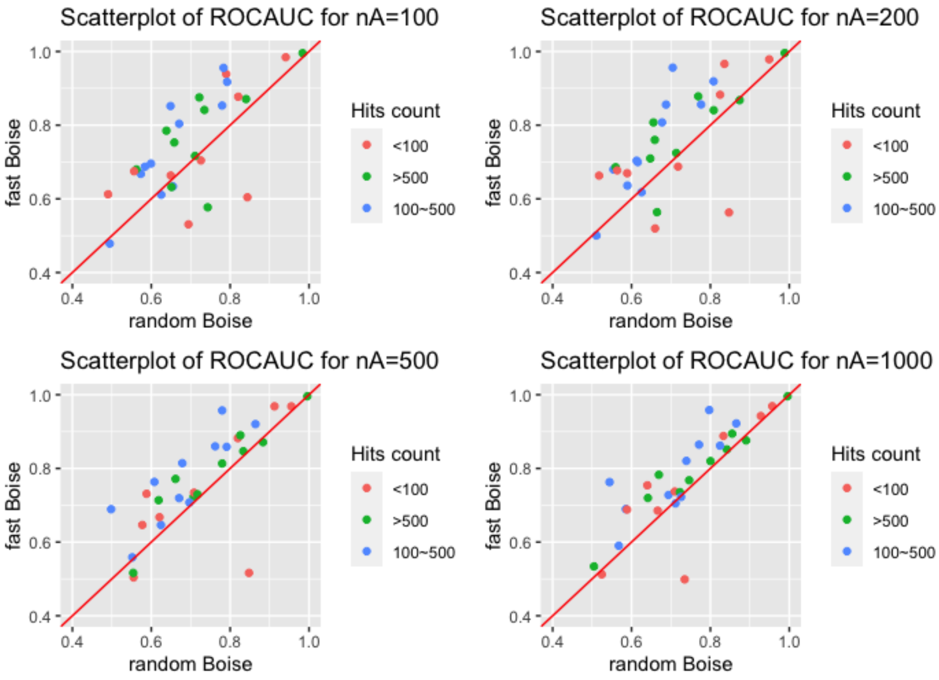

Figure 5: ROCAUC comparison of fast BOISE and randomly selected informers on 30 test targets from PCBA data set for $n_A = 100, 200, 500$ and 1000.

In practice, biochemistry researchers often use FDA approved drugs as the starting point when assessing a novel target. These drugs serve a similar role as the informer set in IBR. In condensed PCBA data set, 324 compounds overlap with the FDA data. Fig 7 is the comparison between fast BOISE and these FDA approved informers on 30 test targets, with informer size of fast BOISE kept the same as the number of available FDA approved compounds for each target. Among 30 test targets, fast BOISE achieves a better ROCAUC on 26, a better NEF1 on 23, and the same NEF1 on 3 targets, implying a potential boost in ranking accuracy if the role of FDA approved drugs is replaced with informers selected by BOISE methods.

When evaluated prospectively, a novel target, PriA-SSB, is tested on most of the compounds in PCBA and is predicted by fast BOISE method. PriA-SSB is a bacteria protein-protein interaction target that has been tested on 114081 out of 134264 PCBA compounds (Alnammi et al., 2021). As for FDA approved drugs, there are 253 drugs that are tested on both PriA-SSB and the other 102 targets in PCBA data set. In this prospective analysis, an approximate CRP sampling is similarly applied on the whole PCBA data set for fast BOISE method, which is

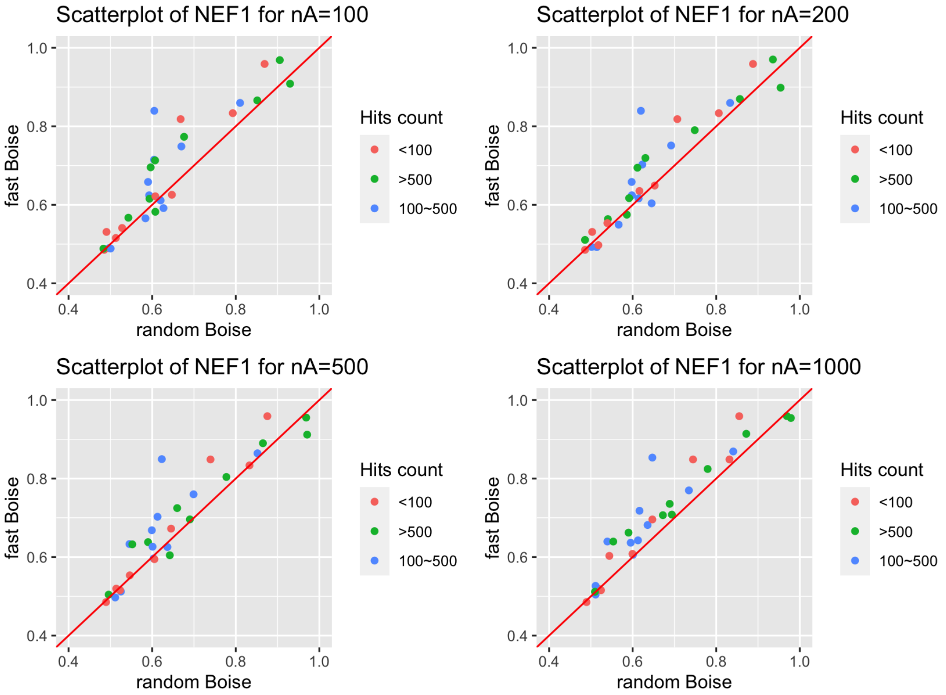

Figure 6: NEF1 comparison of fast BOISE and randomly selected informers on 30 **test targets from PCBA data set for** $n_A = 100, 200, 500$ **and** 1000**.**

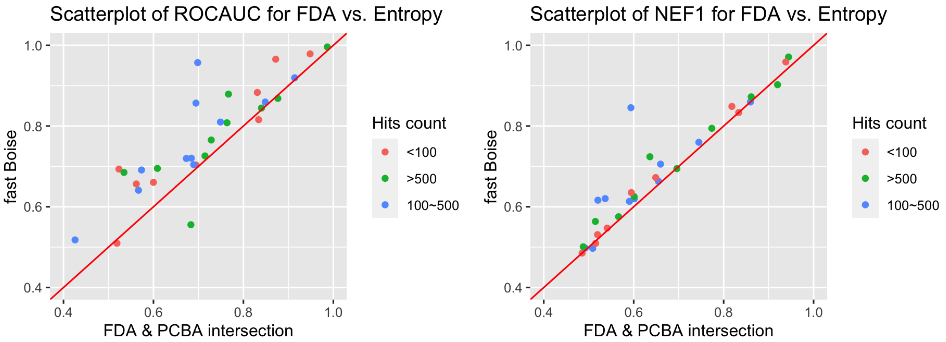

Figure 7: Comparison of fast BOISE and informers using FDA approved drugs on 30 **test targets from PCBA data set.** The informer size of fast BOISE is the same as the number of FDA approved compounds with non-missing values on each test target.

later compared with naive informer sets like randomly selected informers and FDA-approved informers on PriA-SSB. The NEF1 and ROCAUC results are summarized in following Table 4. The result is not so impressive as in retrospective analysis, where fast BOISE is better than FDA approved drugs on both NEF1 and ROCAUC, and achieves a better NEF1 but slightly worse ROCAUC than randomly selected informers.

A further investigation explains the mediocre performance of fast BOISE on PriA-SSB. After clustering the 102 targets from PCBA through CRP, a 103×103 matrix could be used to illustrate the frequency at which each pair of targets are grouped together, where the last target (the 103rd column) is a "fake" target that represents the unknown new cluster that potentially

Table 4: Prospective comparison between fast BOISE and naive informer set on novel target PriA-SSB. The informer size $n_A = 253$ is for informer set that consists of FDA-approved drugs, and $n_A = 500, 1000$ is for randomly selected informer sets.

informer size n_A	253	500	1000
NEF1-fast	0.553	0.605	0.622
NEF1-random	0.564	0.542	0.533
NEF1-FDA	0.536	–	–
ROCAUC-fast	0.862	0.883	0.890
ROCAUC-random	0.890	0.891	0.893
ROCAUC-FDA	0.838	–	–

exists but no one belongs to. The left panel in Figure 8 shows this "adjacency frequency" from our approximate CRP clustering after rearrangement, where darker color indicates that pair of targets are more frequently grouped together. There are a few blocks in the plot, meaning that some targets are highly correlated and hence always clustered together in CRP.

After selecting informers through fast BOISE and getting the intermediate data x_A on PriA-SSB, we can compute the posterior probability on which each of 103 targets will be potentially grouped with PriA-SSB. The posterior probabilities act like the similarity between each target and the novel target. The right panel in Figure 8 plots a weighted adjacency frequency of 103 targets, where weights are those posterior probabilities. This plot aims to find groups of targets (i.e. dark blocks) that are most similar to the novel target based on CRP clustering, and the conclusion is astonishing: it says PriA-SSB is most similar to Target ID 103, the "fake" target representing targets outside PCBA. Therefore, fast BOISE claims that PriA-SSB is not similar to any existing target clusters in PCBA and we should not use PCBA data set to predict results on PriA-SSB. That explains why the performance is not so impressive as in retrospective analysis.

For a novel target not in existing target clusters, if we are forced to give a prediction with current data, the best we can do is to give a simple prediction with the aggregated data, and it turns out the simplest ranking method is the best on PriA-SSB. For example, if we forget about CRP or informer selection and simply rank all the compounds with their average active rates in PCBA data set, this ranking achieves 0.657 on NEF1 and 0.955 on ROCAUC for PriA-SSB, both better than any method in Table 4.

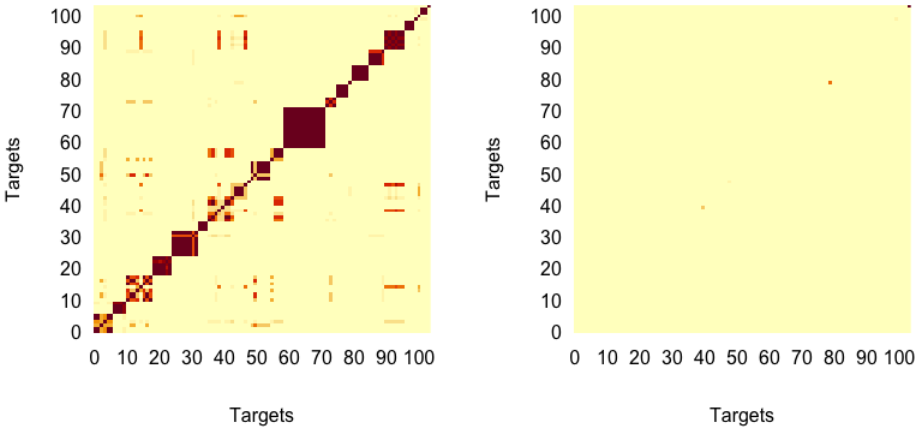

Figure 8: Adjacency frequencies with approximate CRP clustering on PCBA (left), and a weighted version of the same clustering with weights representing similarity to PriA-SSB (right). The left plot shows the adjacency frequency from the approximate CRP clustering after rearrangement, where darker color indicates that pair of targets are more frequently grouped together and blocks indicate groups of targets that are highly correlated. The right plot shows a weighted adjacency frequency of the same clustering, where weights are posterior probabilities of each target being grouped with PriA-SSB. The last "target" (the 103rd column) is a fake target that represents the unknown new cluster that potentially exists but no one belongs to.

4 Discussion

BOISE is an effective IBR method that characterize the virtual screening procedure as a statistical decision problem. The advantage of BOISE comes from a combination of Bayes decision theory and flexible statistical model. Despite being the most effective IBR method, a limitation of BOISE is the computation complexity, which impedes its application to real-world large scale virtual screening. Scalability is crucial in virtual screening, especially in early stage drug discovery where millions of compounds are to be screened. There is a trade-off between accuracy and scalability, however, the fact that BOISE performs better than heuristic IBR methods suggests potential improvement over naíve informer selection when scaling up BOISE.

The proposed BOISE variants, block BOISE and fast BOISE, solve the scaling problem to various degrees: block BOISE aims for enormous chemical space but small informer set, while fast BOISE is recommended for larger informer sets. The idea of block BOISE is to generalize traditional BOISE, where initial bioactivity data is broken into separate blocks. A degenerate case of block BOISE is exactly the original BOISE. Therefore, block BOISE achieves a balance between computation cost and prediction accuracy. For example, it is the best method on ROCAUC in both retrospective and prospective analysis, while only costing 1/6 as much time as needed for original BOISE method. Fast BOISE, on the other hand, origins from the analysis of informers selected. It is related to original BOISE but operates on a totally different family of loss functions. The most significant improvement of fast BOISE is its non-sequential selection of informers, which can save a lot of time when the informer set size is large. It is not a good idea to use non-sequential selection in regard to prediction accuracy, as correlation among informers is not taken into account. However, fast BOISE achieves comparable predictive performances in both retrospective and prospective analyses, indicating that it could serve as a lower bound of the performance for BOISE-like methods and a reference for experimental designs.

There are also a few gaps found between theory and practice for original BOISE. Being the Bayes rule only guarantees the optimal decision rule under given loss function, and that doesn't necessarily lead to a decent IBR method, as shown for frequent-hitters rule in Section 2.1.4. The proposed new loss function $L_1(A, T; x_{i^*})$ resolves the issue where inactive compounds tested in intermediate data can still be ranked top in the final ranking. The improvement is consistent in both synthetic data and case studies, and persists even when model assumptions are violated to different degrees. We then upgrade original BOISE with $L_1(A, T; x_{i^*})$ throughout the case studies. Another proposed loss function, $L_2(A, R; x_{i^*})$, utilizes a pairwise ranking loss for justification of the ranking procedure. BOISE with $L_2(A, R; x_{i^*})$ loss is expected to have best overall ranking performance in theory, and it is true when model assumptions are mostly satisfied. However, the performance is quite sensitive to the structure of initial bioactivity data x_0, as is shown in synthetic data. The sensitivity is also proved in FDA case study, where it is the best method on ROCAUC retrospectively, but merely above fast BOISE prospectively. For the reason of potential violations of target-clustering model (1), loss function $L_1(A, T; x_{i^*})$ may be preferred over $L_2(A, R; x_{i^*})$.

In prospective analysis on FDA data, an apparent performance discrepancy appears among target types, although the computation does not include such information. In average, most BOISE variants perform better on molecule-level targets like AlphaScreen / FP assays and protein targets; the performance degrades on cell-level targets like spores and fungal cells. It is less related to chemical screening labs: for example, targets 11 to 16 are all from Keck's lab, and target 13 is a cell target while the others are molecule-level targets. Although other targets are always among the best performed targets throughout the prospective analysis, performance on target 13 is usually just around the average and inferior to performances on protein targets from other labs. The discrepancy among target types indicates that there is a more direct relationship between compound structure and bioactivity for molecule-level assays. For cellular assays, where the readout is often OD600, chemical screening researchers are basically just looking at how well visible light passes through the medium, where liquid gets more clear on cell death. For cell death, there might be many molecular mechanisms that produce a positive readout. A careful selection of targets may be helpful in constructing the initial bioactivity matrix x_0, as it may be confusing to mix molecular targets and cellular targets in a single bioactivity matrix.

References

Alnammi, M., Liu, S., Ericksen, S. S., Ananiev, G. E., Voter, A. F., Guo, S., Keck, J. L., Hoffmann, F. M., Wildman, S. A., and Gitter, A. (2021). Evaluating scalable supervised learning for synthesize-on-demand chemical libraries. 10.26434/chemrxiv-2021-fg8z9.

Clemons, P. A., Bittker, J. A., Wagner, F. F., Hands, A., Dančík, V., Schreiber, S. L., Choudhary, A., and Wagner, B. K. (2021). The Use of Informer Sets in Screening: Perspectives on an Efficient Strategy to Identify New Probes. *SLAS DISCOVERY: Advancing the Science of Drug Discovery* **26,** 855–861.

Kaufman, L. and Rousseeuw, P. J. (2005). *Finding groups in data: an introduction to cluster analysis.* Wiley series in probability and mathematical statistics. Wiley, Hoboken, N.J.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics* **22,** 79 – 86.

Ma, X., Korthauer, K., Kendziorski, C., and Newton, M. A. (2021). A compositional model to assess expression changes from single-cell RNA-seq data. *The Annals of Applied Statistics* **15,** 880 – 901.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning* **1,** 81–106.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. (2015). Massively multitask networks for drug discovery.

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **50,** 742–754. PMID: 20426451.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20,** 53–65.

Wang, Y., Bryant, S. H., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B. A., Thiessen, P. A., He, S., and Zhang, J. (2016). PubChem BioAssay: 2017 update. *Nucleic Acids Research* **45,** D955–D963.

Yu, P., Ericksen, S., Gitter, A., and Newton, M. A. (2022). Bayes optimal informer sets for early-stage drug discovery. *Biometrics* **1-13,** https://doi.org/10.1111/biom.13637.

Zhang, H., Ericksen, S. S., Lee, C.-p., 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,** e1006813.