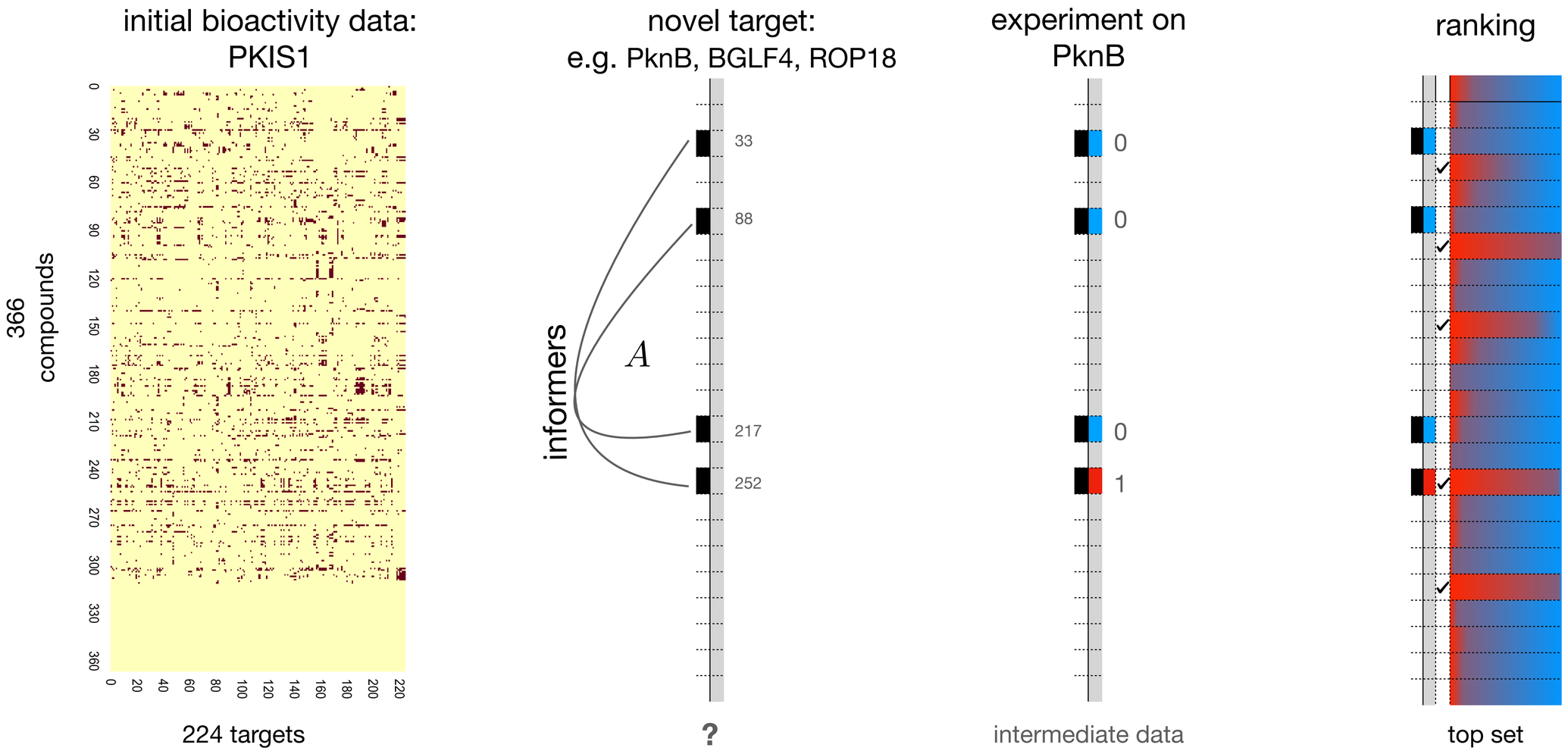
BOISE with blocks

Peng Yu, Michael A. Newton and Drug discovery group
Feb. 17, 2022

Informer-based ranking (IBR)



Original BOISE: Setting

- Data: $x_0 = \{x_{i,j}\}$, a binary bioactivity matrix.
- Sampling model: $P(x_{i,j} = 1 \mid \theta_{i,j}) = \theta_{i,j}$
- Stage 1: informer set selection $A \subseteq \{\text{compounds } j\}$
- Stage 2: top set selection $T \subseteq \{\text{compounds } j\}$
- Loss function: $L(A, T, \theta) = \sum_{j \in T} (1 - \theta_{i^*,j})$
- Prior model for $\theta = \{\theta_{i,j}\}$
- Objective: minimize posterior expected loss

$$(\hat{A}, \hat{T}) = \arg\min_{A,T} \mathbb{E}\{L(A, T, \theta) \mid x_0\}$$

Original BOISE: model for θ

- Rational: new target is similar to some previous targets in initial data.
- clustering on targets: $\mathcal{C} = \{c_k\} \sim \mathrm{CRP}_m(m_0)$

$$p(\mathcal{C}) = \frac{m_0^K \Gamma(m_0)}{\Gamma(m + m_0)} \prod_{k=1}^{K} \Gamma(m_k)$$

- identical targets within each cluster:

$$\theta_{i,j} \mid \mathcal{C}, \phi = \phi_{k,j} 1(i \in c_k)$$

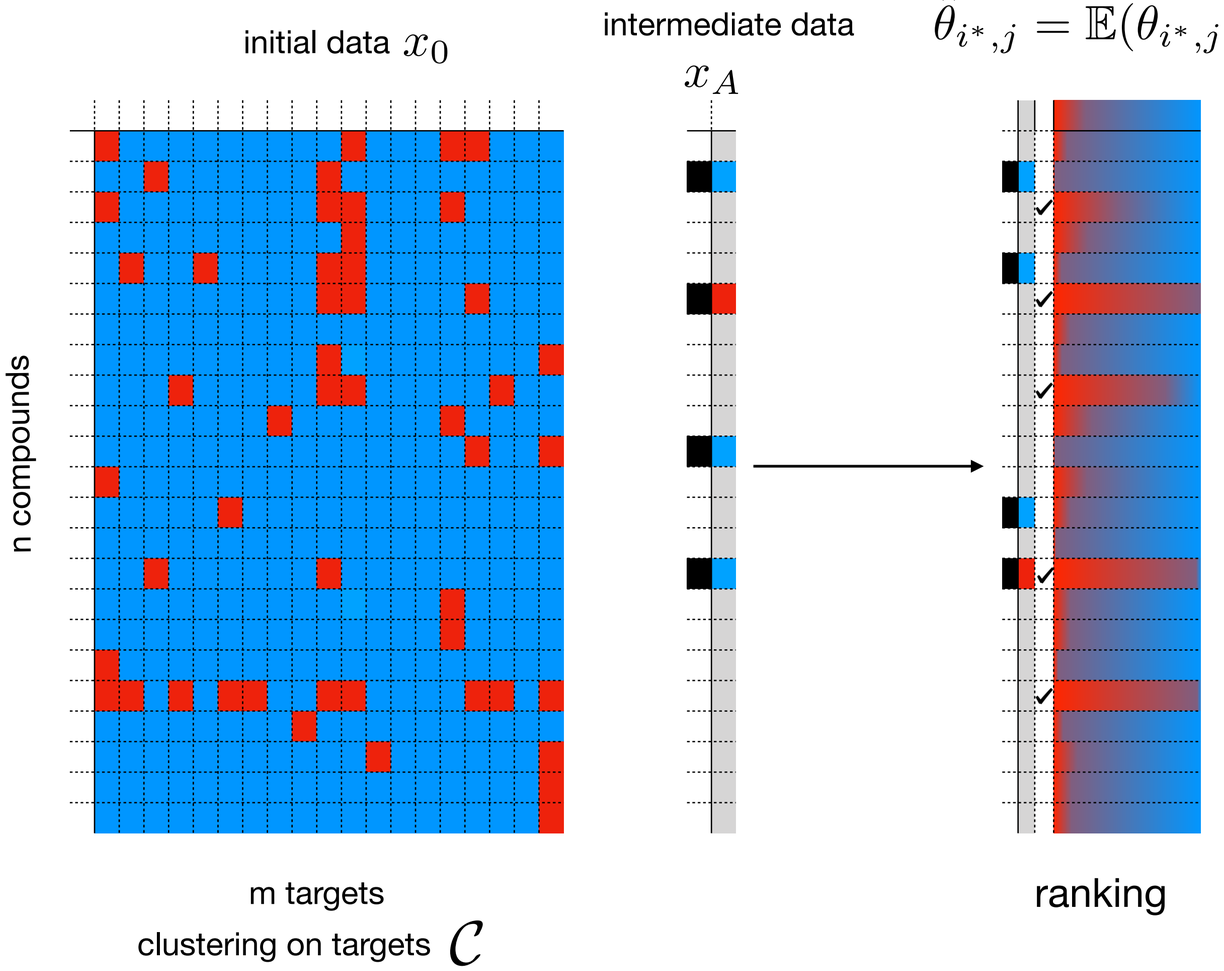
$$x_{i,j} \mid \mathcal{C}, \phi \sim \mathrm{Bernoulli}\{\phi_{\mathrm{k,j}} 1(\mathrm{i} \in \mathrm{c_k})\}$$

- independent compounds:

$$\phi_{k,j} \sim \mathrm{Beta}(\alpha_0, \beta_0),\ \mathrm{k} = 1, \cdots, \mathrm{K},\ \mathrm{j} = 1, \cdots, \mathrm{n}$$

Original BOISE

PEL2 computation / ranking



approach is to re-use the sampled clusterings $\mathcal{C}$, noting by iterated expectations that $\hat{\theta}_{i^*,j} = E\left\{\tilde{\theta}_{i^*,j}|x_0, x_A\right\}$, where $\tilde{\theta}_{i^*,j} = E(\theta_{i^*,j}|\mathcal{C}, x_0, x_A)$. This inner expectation is an average over ways the new target i^* may (or may not) cluster with the existing targets, and we find:

$$\tilde{\theta}_{i^*,j} = \sum_{k=0}^{K} p_k \left\{ \frac{a_{k,j} + x_{i^*,j}\mathbb{1}(j \in A)}{a_{k,j} + b_{k,j} + \mathbb{1}(j \in A)} \right\} \tag{11}$$

where p_k is the conditional probability (given x_0, x_A, and $\mathcal{C}$) that i^* links to cluster c_k:

$$p_k \propto m_k \prod_{j \in A} \left(\frac{a_{k,j}}{a_{k,j} + b_{k,j}} \right)^{x_{i^*,j}} \left(\frac{b_{k,j}}{a_{k,j} + b_{k,j}} \right)^{1 - x_{i^*,j}}$$

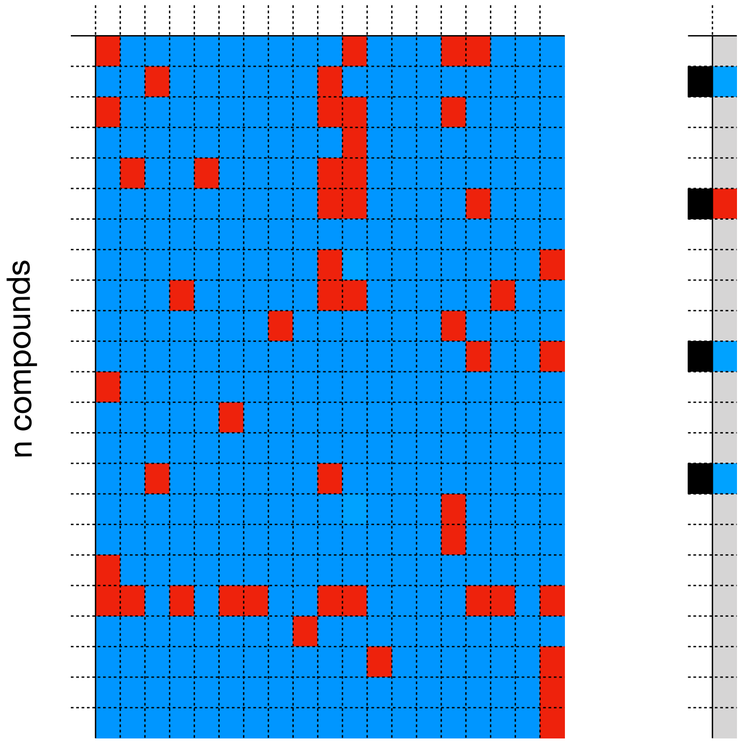
and where proportionality is resolved by $\sum_{k=0}^{K} p_k = 1$. The sought-after $\hat{\theta}_{i^*,j}$ is marginal to uncertainty in clusterings but conditional on intermediate data x_A. The generic solution

Original BOISE

PEL1 computation

initial data x_0

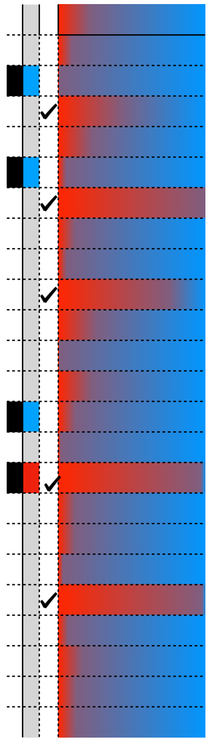
Sample: 100,000 possible x_A



$x_A \sim p(x_A|x_0)$

For each sampled x_A

PEL2



Average PEL_2 over all sampled x_A

$$\mathrm{PEL}_1(x_0, A) = \mathbb{E}\left\{\mathrm{PEL}_2(x_0, A, x_A) \,|\, x_0\right\}$$

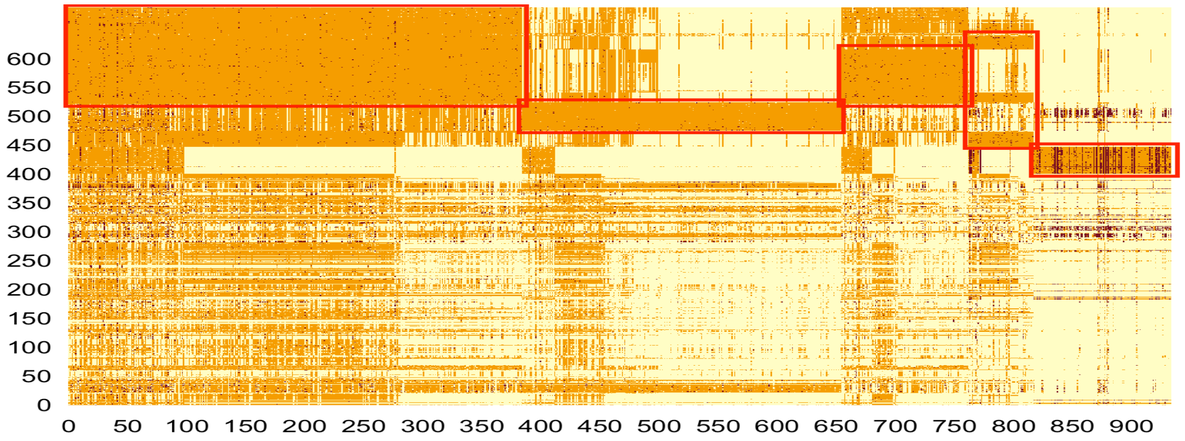
$$\mathrm{PEL}_2(x_0, A, x_A) = \sum_{j \in T^*(x_0, A, x_A)} (1 - \hat{\theta}_{i^*,j})$$

New data, new challenge

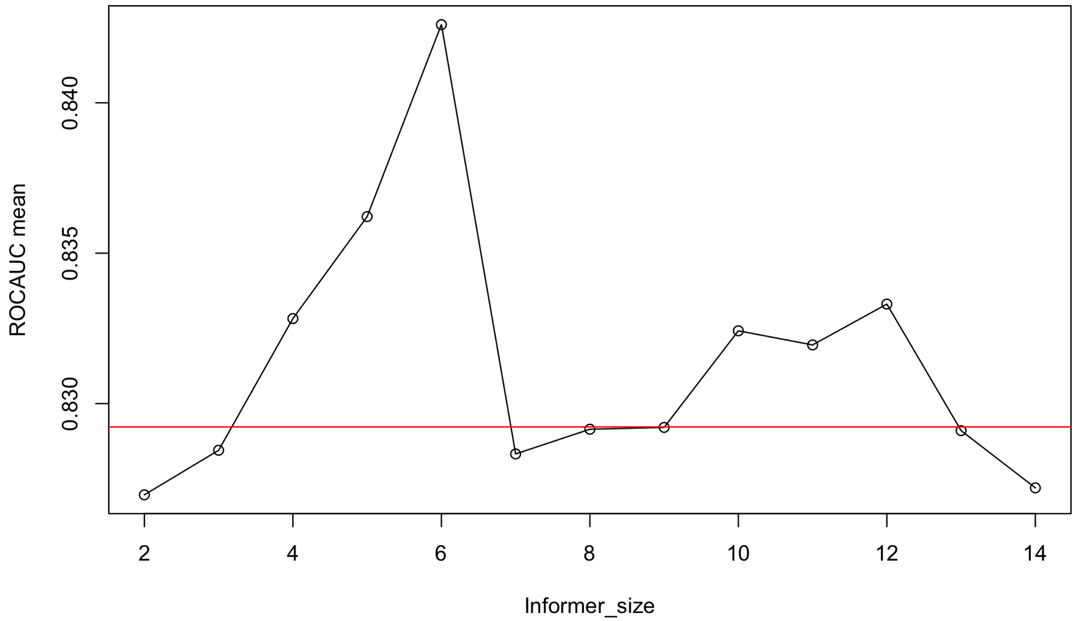
- FDA data: 688 targets x 2002 compounds. 75.8% missing. 5.17% active on non-missing.
- How to evaluate performance?
  - A test set (around 100) that are as complete as possible.
- Initial attempts on FDA data:
  - iteratively find largest **complete** sub-matrices with size > 1000, drop the unselected compounds:
    - 5 blocks, with 688 targets x 933 compounds; 54.6% missing rate, 4.37% active rate.
    - top 100 complete targets get > 2/3 complete responses.
    - average missing rate on dropped columns is 92.9% — non-informative compounds

New data, new challenge

- preprocessed data:

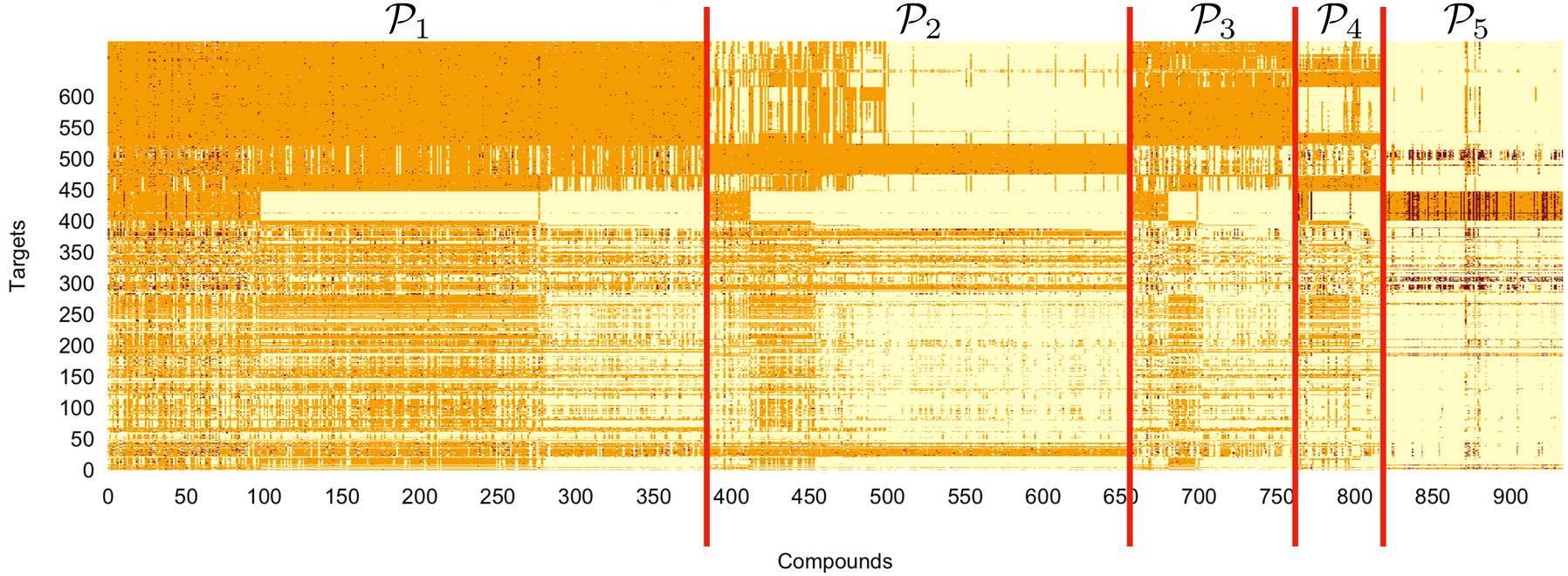


- Original BOISE on this data:
  - horizontal red line is the performance of basic ranking --
  - rank the compounds w.r.t. average active rate. No informers needed



Will more flexible model help?

- Original BOISE assumes all compounds share the same target clustering structure. This could lead to all singleton clusters with high probability (especially for “wide” matrix).
- One idea is to partition compounds into different groups and fit a separate model for each group of compounds.
- Posterior predictive check shows promising p-value on preprocessed FDA data: (0.132 for separate models vs. 0.003 for single model.)



Block BOISE: model for θ

- Idea: partition compounds into groups; fit separate BOISE models.
- partition of compounds: $\mathcal{G} = \{g_k\}$
- clustering on targets: $\mathcal{C}_k \sim \mathrm{CR}_m(m_0), k = 1, \cdots, K$
- identical targets within each cluster:

$$\theta_{i,j} \mid j \in g_k, \mathcal{C}_k, \phi = \phi_{r,j} 1(i \in c_{k,r})$$

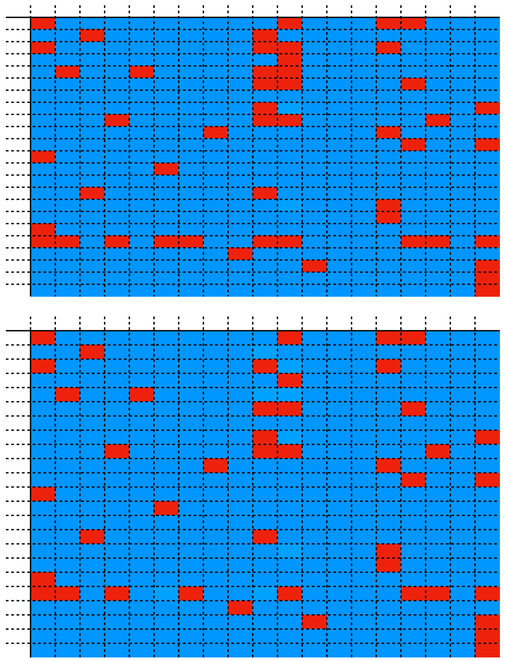
- independent compounds:

$$\phi_{r,j} \mid j \in g_k \sim \mathrm{Beta}(\alpha_k, \beta_k), r = 1, \cdots, R_k$$

Block BOISE

PEL2 computation

initial data x_0



x_1

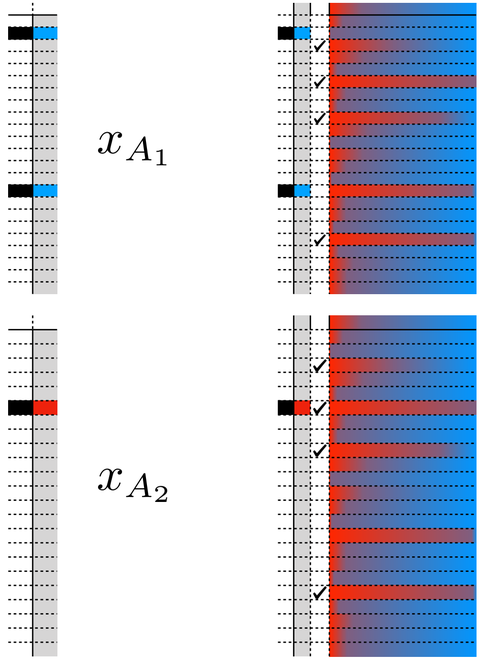
x_2

n compounds in groups

m targets

clustering on targets $\mathcal{C}$

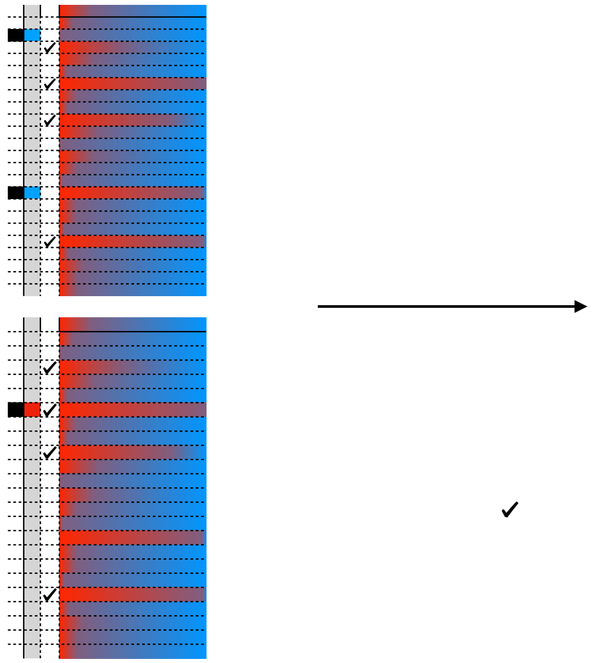
Intermediate data



x_{A_1}

x_{A_2}

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



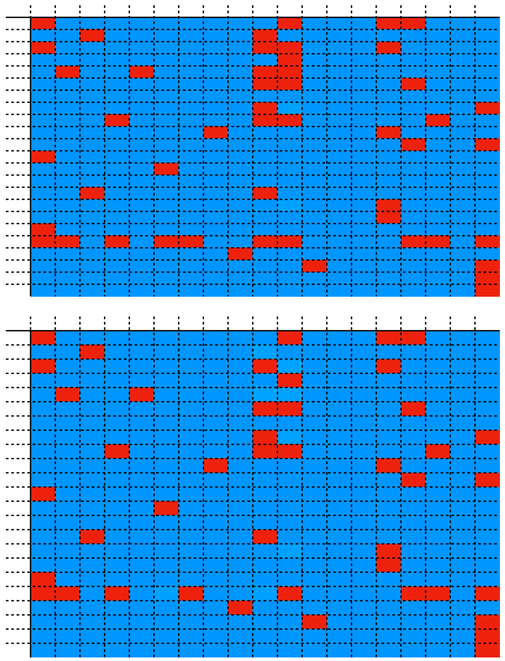
✓

Combine together to select Top set T and calculate PEL2

Block BOISE

PEL1 computation

initial data x_0



x_1

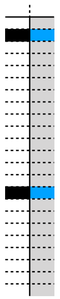
x_2

n compounds in groups

m targets

clustering on targets $\mathcal{C}$

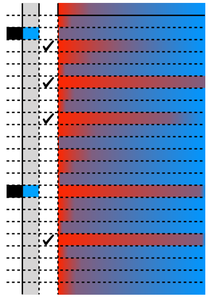
Sample

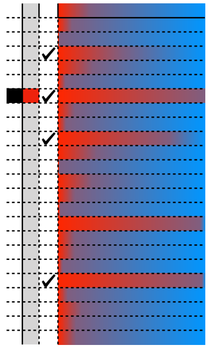


$$x_{A_1} \sim p(x_{A_1} \mid x_1)$$

$$x_{A_2} \sim p(x_{A_2} \mid x_2)$$

$$\hat{\theta}_{i^*,j} = \mathbb{E}(\theta_{i^*,j} \mid x_k, x_{A_k})$$





Final samples of x_A is:

$$x_{A_1} \otimes x_{A_2}$$

Example: 5 cpds 2 blocks

Samples of x_{A_1}

x_{A_1}	**prob.**	$\hat{\theta}_{i^*,j}$
"00"	0.85	[0.01,0.01,0.02]
"01"	0.049	[0.01,0.98,0.53]
"10"	0.10	[0.97,0.01,0.02]
"11"	0.001	[0.95,0.99,0.6]

Samples of x_{A_2}

x_{A_2}	**prob.**	$\hat{\theta}_{i^*,j}$
"0"	0.95	[0.01,0.13]
"1"	0.05	[0.99,0.47]

$$x_{A_1} \otimes x_{A_2}$$

Final samples of x_A

x_A	**prob.**	$\hat{\theta}_{i^*,j}$
"000"	0.8075	[0.01,0.01,0.02,0.01,0.13]
"010"	0.047	[0.01,0.98,0.53,0.01,0.13]
"100"	0.095	[0.97,0.01,0.02,0.01,0.13]
"110"	0.001	[0.95,0.99,0.6,0.01,0.13]
"001"	0.0425	[0.01,0.01,0.02,0.99,0.47]
"011"	0.00245	[0.01,0.98,0.53,0.99,0.47]
"101"	0.005	[0.97,0.01,0.02,0.99,0.47]
"111"	5E-05	[0.95,0.99,0.6,0.99,0.47]

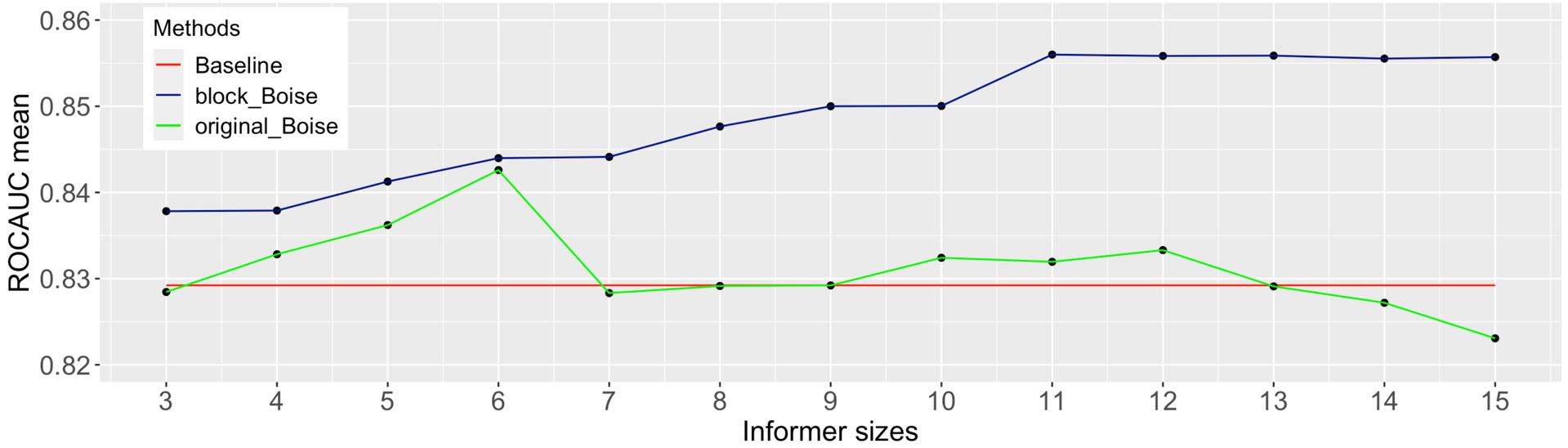
Calculate PEL1 from the final samples!

Block BOISE: fast informer search

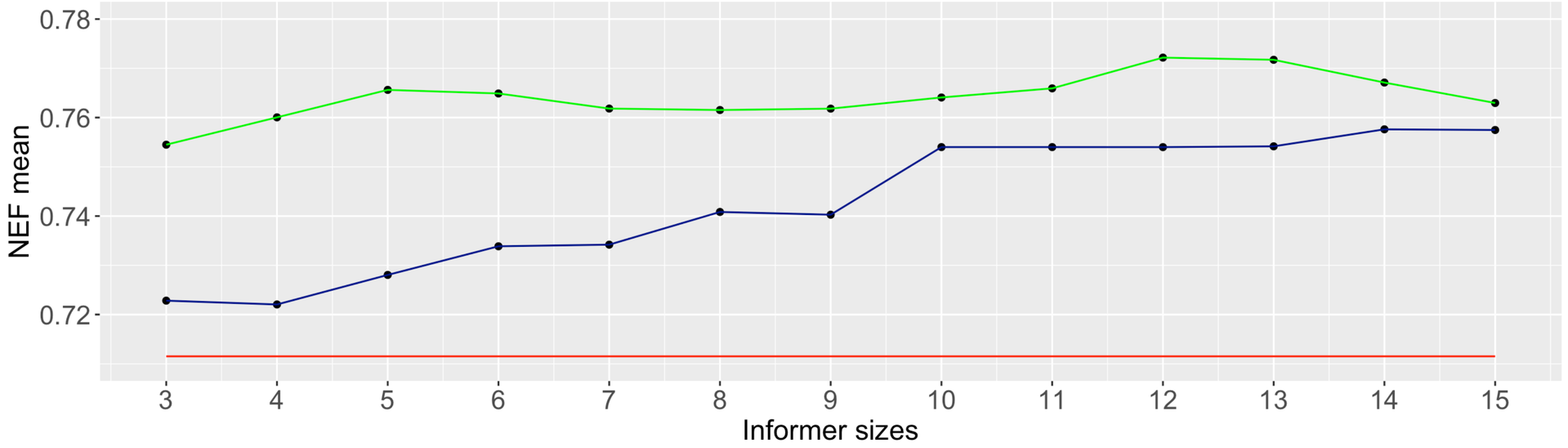
- The computation is reduced -- $\mathcal{O}(2^{n_A}) \to \sum_k \mathcal{O}(2^{n_{A_k}})$
- In the previous example, original BOISE needs to evaluate posterior mean and probability for 8 times; block BOISE only needs 2 + 4 = 6 evaluations.
- Approximate time to select 15 informers on FDA data:
  - original BOISE: 2 weeks.
  - block BOISE (random column cluster): 5-7 days.
  - block BOISE (fixed column cluster): < 24h

Empirical results on FDA data

ROCAUC

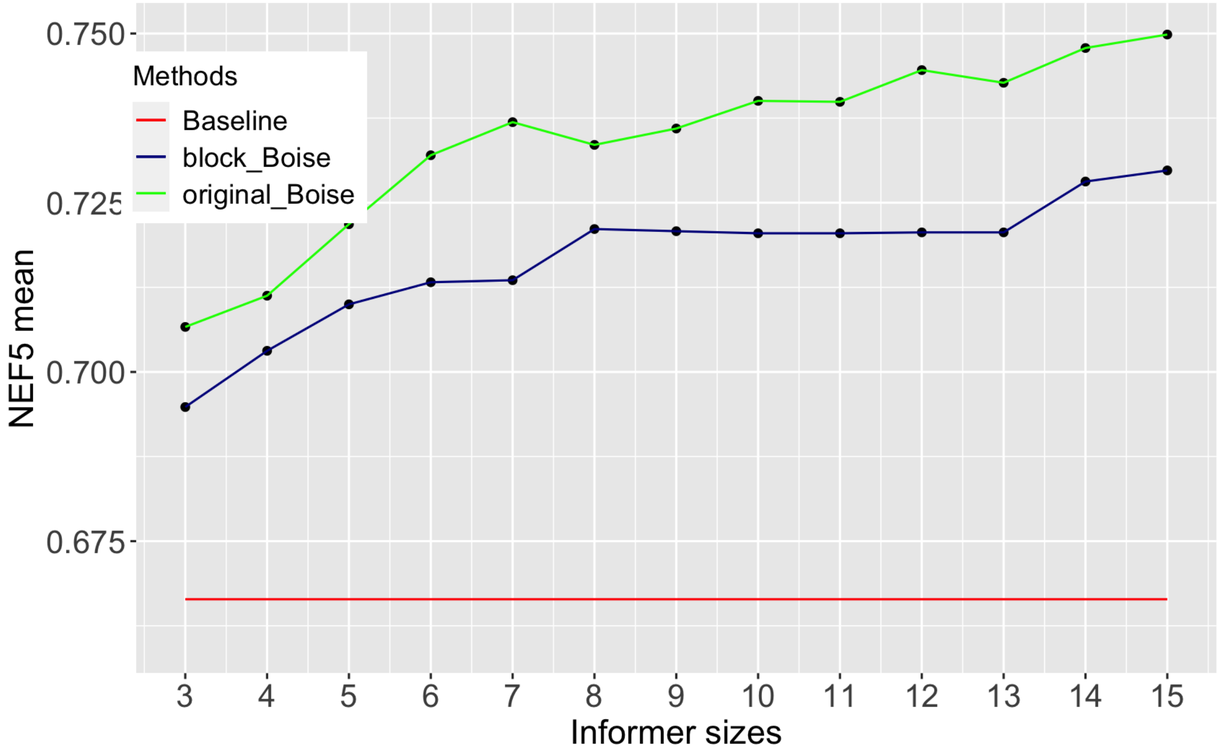


NEF10

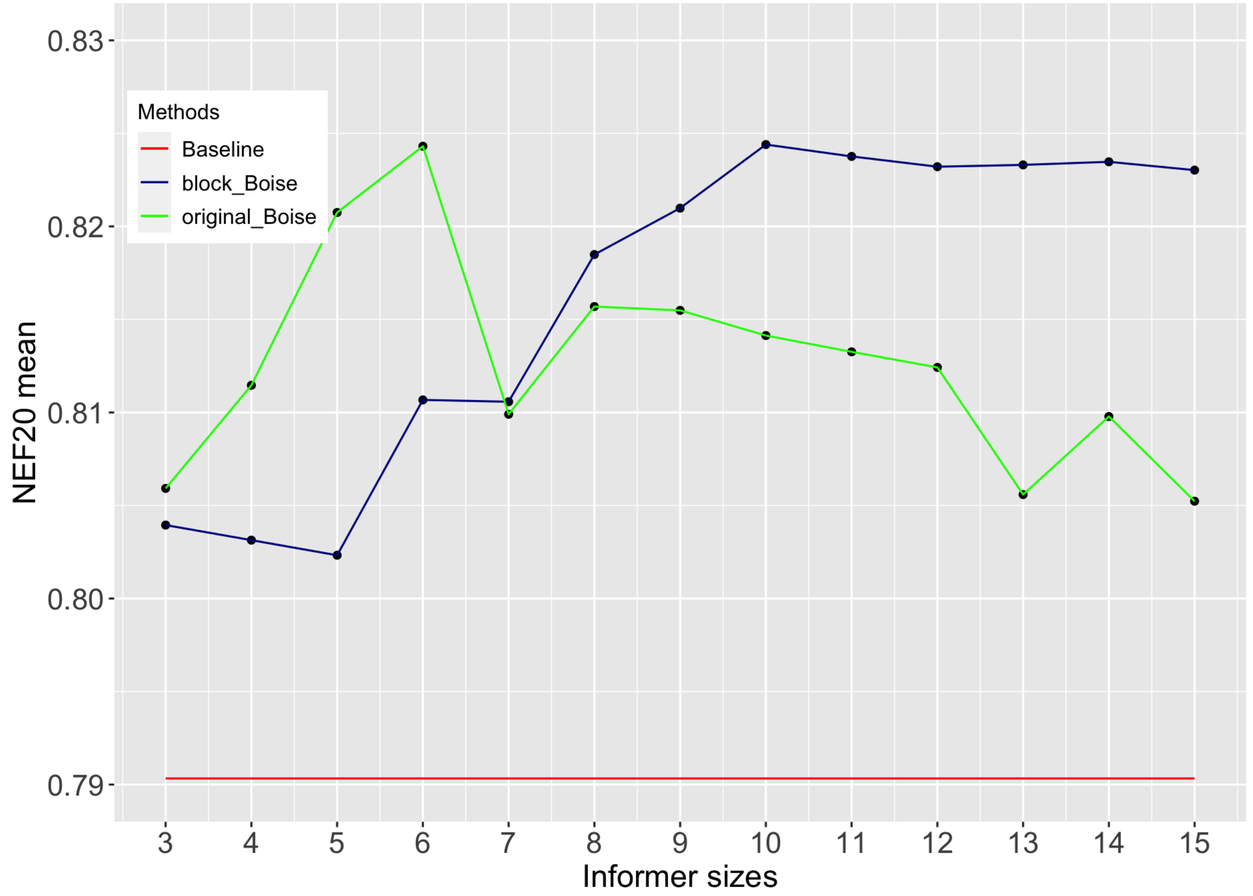


NEF5 vs NEF20

- NEF 5

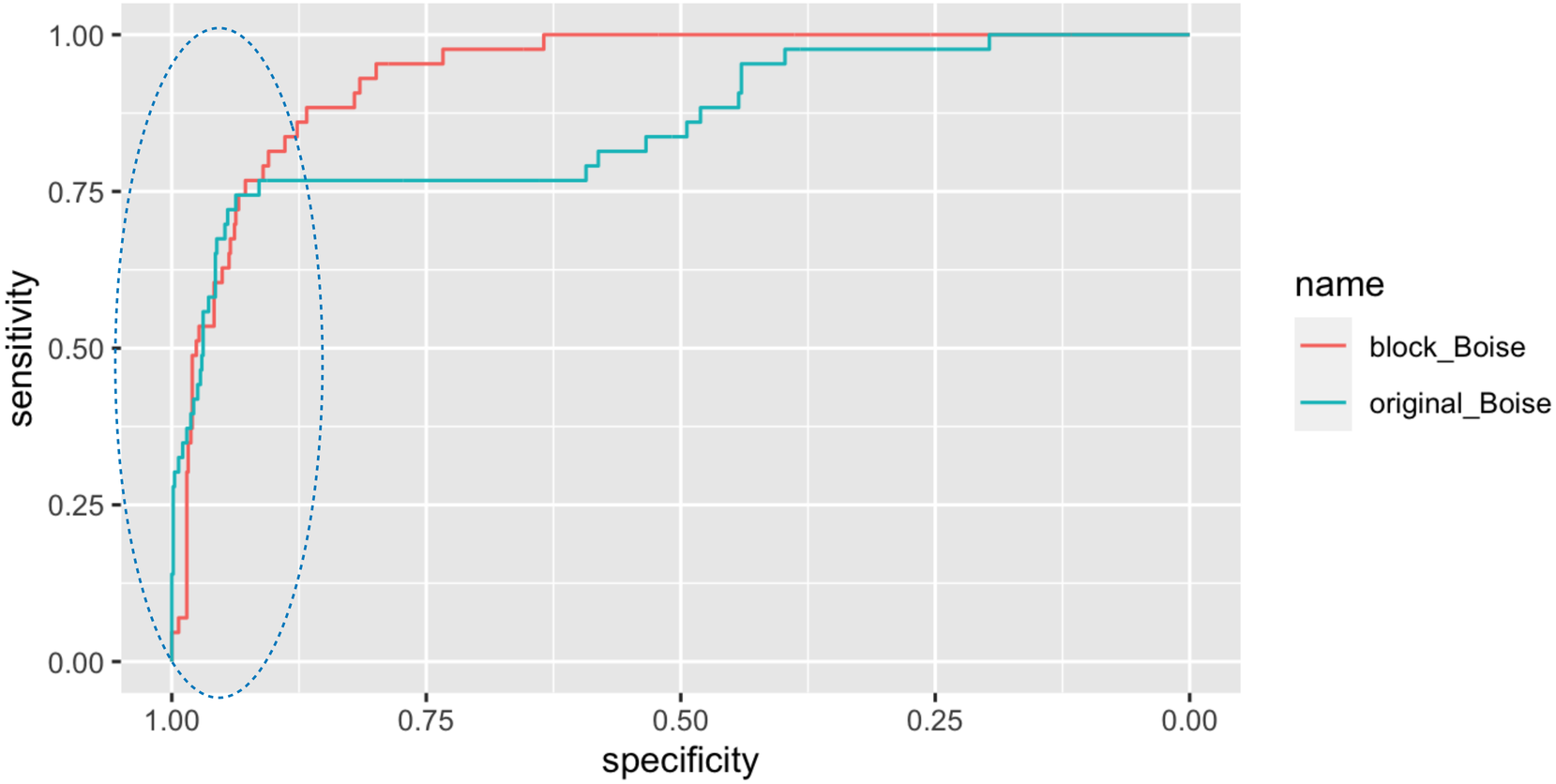


- NEF 20

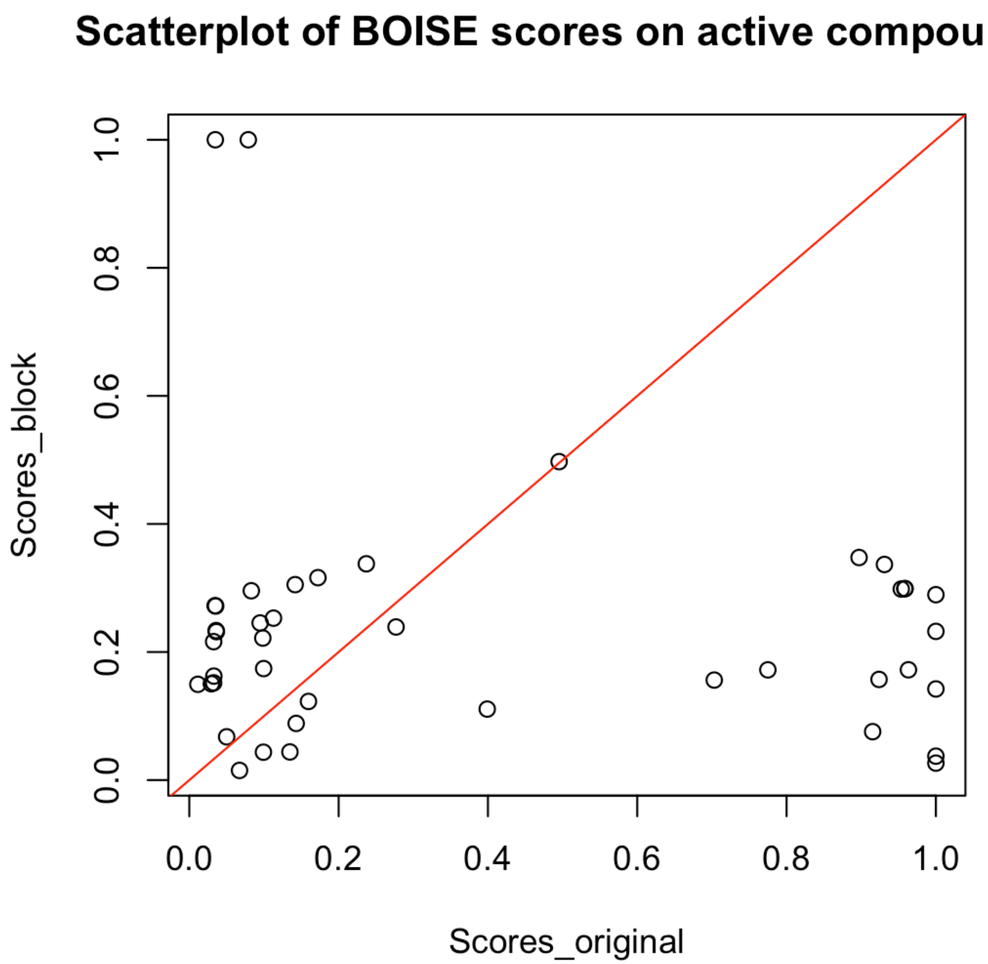


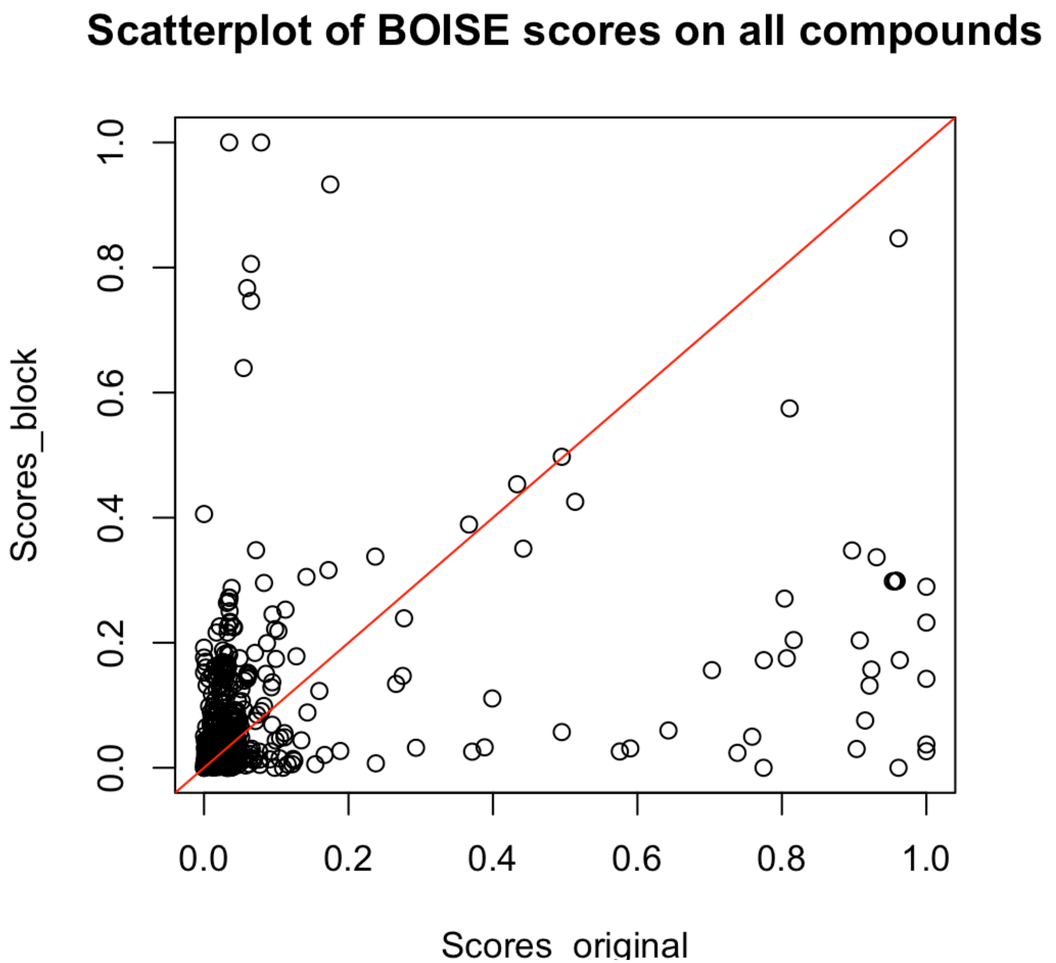
AID_1296009: ROC plot

- nA = 13, original BOISE ROCAUC = 0.858, NEF10 = 0.858, NEF5 = 0.722, NEF20 = 0.855.
- Block BOISE ROCAUC = 0.938, NEF10 = 0.845, NEF5 = 0.762, NEF20 = 0.927



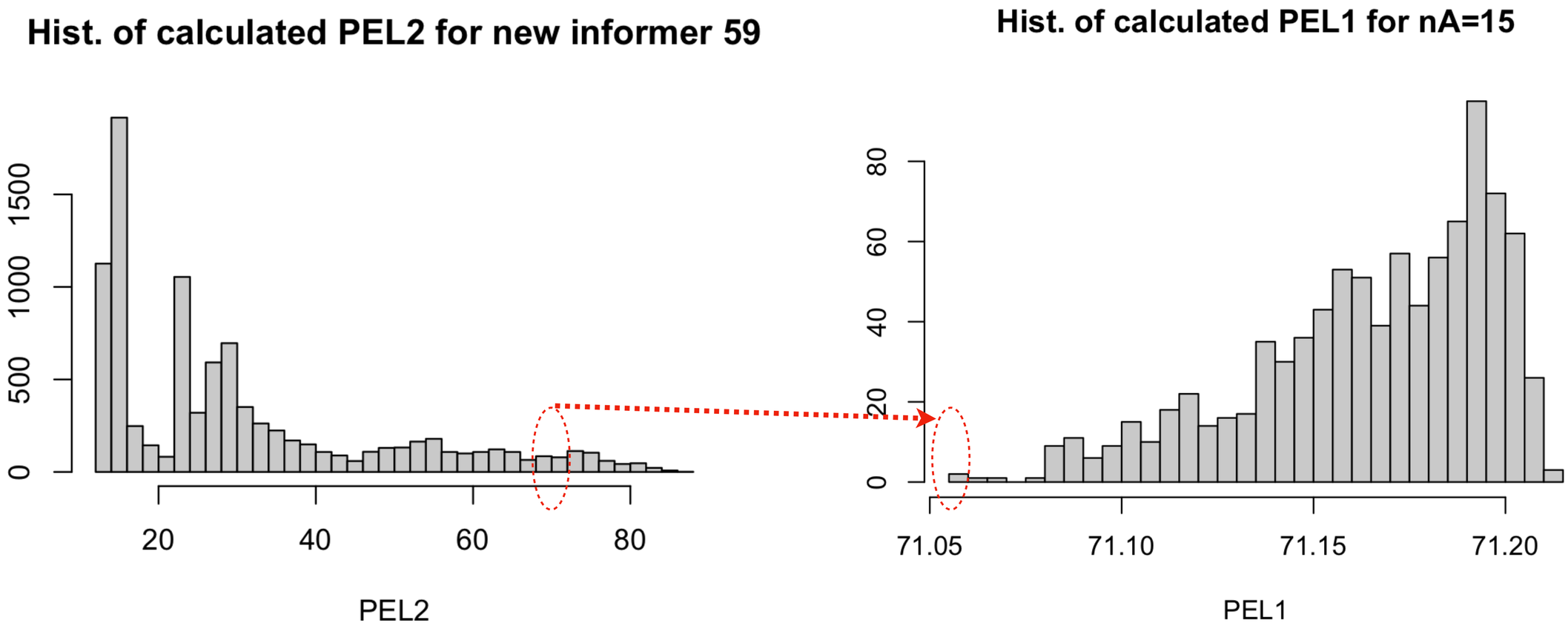
AID_1296009: Compound scores





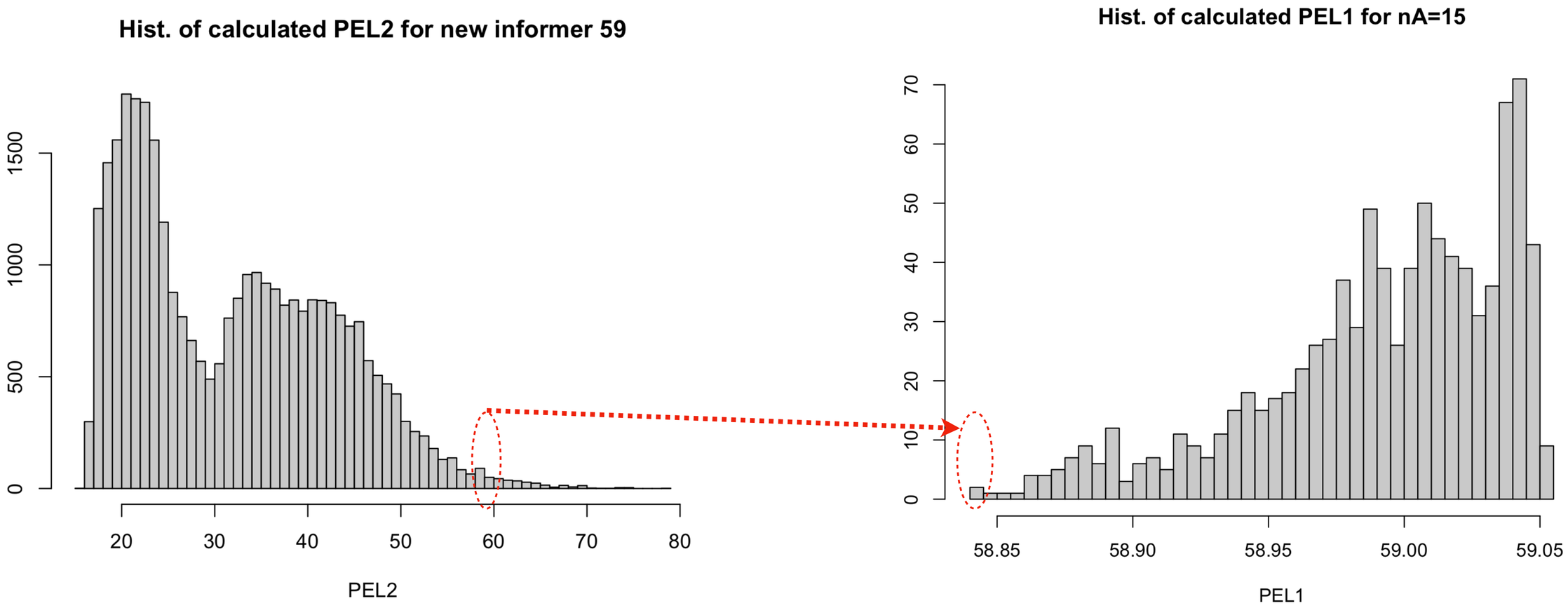
Original BOISE: PELs

- Original BOISE: going from 14 to 15 informers



Block BOISE: PELs

- Block BOISE:



Higher missing rates on informers

- Missing rates in informers are higher than average missing rate in the test set:

NA rate	original BOISE	block BOISE	Average in dataset
Test	0.389	0.422	**0.278**
Train	0.555	0.599	0.592

- In preprocessing data we hope that at least informers are not missing in the test set, while the truth is around 40% informers are useless in evaluation…