Confidence Bands and Hypothesis Tests for Hit Enrichment Curves

Jeremy R Ash (✉ jrash@ncsu.edu)

North Carolina State University

Jacqueline M Hughes-Oliver

North Carolina State University

Method Article

Keywords: Virtual Screening, Enrichment Factor, Lift Curve, Early Enrichment, Ranking Algorithm, Empirical Process

Posted Date: February 21st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1374910/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Confidence Bands and Hypothesis Tests for Hit Enrichment Curves

Jeremy R Ash^{1,2*} and Jacqueline M Hughes-Oliver¹

*Correspondence: jrash@ncsu.edu

¹Department of Statistics,
Bioinformatics Research Center,
North Carolina State University,
Raleigh, United States
Full list of author information is
available at the end of the article

Abstract

In virtual screening for drug discovery, hit enrichment curves are widely used to assess the performance of ranking algorithms with regard to their ability to identify early enrichment. Unfortunately, researchers almost never consider the uncertainty associated with estimating such curves before declaring differences between performance of competing algorithms. Uncertainty is often large because the testing fractions of interest to researchers are small. Appropriate inference is complicated by two sources of correlation that are often overlooked: correlation across different testing fractions within a single algorithm, and correlation between competing algorithms. Additionally, researchers are often interested in making comparisons along the entire curve, not only at a few testing fractions. We develop inferential procedures to address both the needs of those interested in a few testing fractions, as well as those interested in the entire curve. For the former, four hypothesis testing and (pointwise) confidence intervals are investigated, and a newly developed EmProc approach is found to be most effective. For inference along entire curves, EmProc-based confidence bands are recommended for simultaneous coverage and minimal width. While we focus on the hit enrichment curve, this work is also appropriate for lift curves that are used throughout the machine learning community. Our inferential procedures trivially extend to enrichment factors, as well.

Keywords: Virtual Screening; Enrichment Factor; Lift Curve; Early Enrichment; Ranking Algorithm; Empirical Process

1 Introduction

Ranking algorithms order items according to the belief that they possess some desired feature. When the presence/absence of the desired feature is known, ranking algorithms are often evaluated by “testing” items according to the relative rank or testing order. The focus of this paper will be on a performance curve used to assess whether early tests reveal the desired feature – the hit enrichment curve. The statistics and machine learning communities use a number of variations of this performance curve to evaluate ranking algorithms as well, including the enrichment factor or lift curve. Since the variations of these curves involve scaling by a constant on either the x or y axis, all of the inferential procedures discussed in this paper are equivalent. Popular software such as the R package `caret` [1], SAS Enterprise Miner [2], and JMP [3] can be used to construct these curves. These curves are used extensively in the evaluation of virtual screens of chemical compounds for drug discovery [4]. Due to the interpretability of these metrics when classes are extremely imbalanced and testing fractions are small, we believe these are important metrics

1 to report when evaluating virtual screens. The curves are also used in a number of
2 other applications such as the evaluation of marketing campaigns [5].

3 In the context of virtual screening, the desired feature is often a biological activ-
4 ity. Typically the desired activity is binding to a protein target, so we will refer to
5 the chemical compounds as ligands. Ligands are scored, where the scores are pro-
6 vided by one of the many ranking algorithms available, such as molecular docking
7 algorithms, pharmacophore models, or quantitative structure-activity relationship
8 (QSAR) models.

9 Empereur-Mot et al. [6] recently provided a web application for constructing per-
10 formance curves to evaluate virtual screening methods. One of the performance
11 curves they provide is the hit enrichment curve. The software provides many nice
12 interactive features for exploring performance metrics at points along a curve. It also
13 provides utilities for combining the scores from multiple methods into a consensus
14 score that may improve ranking performance.

15 Of tremendous importance in a world where virtual screening data is tightly
16 guarded, Empereur-Mot et al. [6] make some of their data freely available. We use
17 one of the case studies as a demonstration dataset. The target was the protein
18 regulating gene peroxisome proliferator-activated receptor gamma (PPAR γ), which
19 has been linked to several diseases such as obesity, diabetes, atherosclerosis, and
20 cancer [7]. While this is a small dataset for a retrospective screening evaluation by
21 present-day standards, the rarity of actives ($\hat{\pi}_+ = 0.0265$ is the observed fraction
22 of active ligands) and the small testing fractions are very consistent with larger
23 screening studies [8]. The primary goal in such a study is to discover scoring methods
24 that are able to correctly identify active ligands, and to do so very early in the testing
25 phase [4, 9, 10].

26 The hit enrichment curve is commonly used to summarize effectiveness of a screen-
27 ing campaign. It plots the proportion of active ligands identified (i.e., the recall)
28 as a function of the fraction of ligands tested, where testing order is based on the
29 score produced by a virtual screening method. Larger recall values are preferred
30 and, going a step further, these are more relevant when they occur at small testing
31 fractions. In other words, one hopes to demonstrate improvement in *early enrich-*
32 *ment*.

33 While the hit enrichment curve is designed to show results for all testing fractions,
34 it is common to focus on fractions below 0.1 and even below 0.001 [8]. The size of our
35 demonstration dataset limits us to consider 0.001 as the smallest testing fraction
36 (resulting in just three ligands tested), but present-day screening campaigns can
37 easily involve millions of ligands and hence testing fractions below 0.001 are reason-
38 ably considered. For a given testing fraction, one may want to compare observed
39 recall values for competing scoring methods.

40 For the PPAR γ study, Empereur-Mot et al. [6] use three popular docking meth-
41 ods to score ligands: Surflex-dock, ICM, and Vina. We scale all docking scores so
42 that a larger value is consistent with active ligands; as a result, ICM and Vina
43 scores have been negated. Empereur-Mot et al. [6] also evaluate several methods
44 for constructing consensus scores, and we limit investigations to two of their best
45 performers: maximum of the z-scores from Surflex-dock and ICM, and the minimum
46 of the ranks from Surflex-dock and ICM.

Figure 1 shows estimated hit enrichment curves for the five scoring methods over the full range of testing fractions, along with the ideal hit enrichment curve that would be expected if the 85 active ligands were identified in the first 85 tests, and the hit enrichment curve consistent with random identification of active ligands. This figure differs slightly from that produced by the web application of Empereur-Mot et al. [6] due to a difference in how ties are handled; we use inverse distribution functions (strategy discussed below) to avoid arbitrary indications of better performance due simply to random ordering.

Comparing consensus methods to individual docking methods, is the observed improvement in recall at testing fraction 0.1 significant? With recall being the proportion of active ligands identified at this testing fraction of 321 tests, it may be tempting to apply a procedure based on comparing independent binomial proportions. There are, however, two complications. The first is that determination of testing order requires scores from all ligands, and hence this introduces correlation between the 321 tests that are applied for a single scoring method. The second complication is that competing scoring methods are very likely positively correlated and so the uncertainty associated with differences between hit enrichment curves may possibly be reduced, thus improving the power to uncover differences. This paper develops appropriate techniques for comparing hit enrichment curves that result from competing scoring methods. Our inferential procedures trivially extend to enrichment factors, as well. The need for proper inferential methods for these early enrichment metrics has been emphasized by many recent papers [11, 12]. In Hawkins et al. [13], the authors argue against the enrichment factor by stating “It is difficult to calculate analytically errors in enrichment, and there is no available literature for such a calculation.” We hope to address this concern in this paper. Perhaps most importantly, this paper establishes a statistical foundation that is necessary for carefully thinking about the uncertainty in early enrichment measures when classes are extremely imbalanced and testing fractions are small.

The paper is organized as follows. First, we introduce hit enrichment curves constructed from ranking algorithms. Next, we present four approaches for hypothesis

1 testing and confidence intervals to compare two hit enrichment curves, along with
 2 simulation studies to compare effectiveness of the approaches; EmProc is newly pro-
 3 posed here, while three other approaches are applied in new ways. Then, we present
 4 confidence band procedures and simulation results for an entire hit enrichment
 5 curve from a single algorithm. Next, we consider bands for the difference between
 6 two hit enrichment curves for competing algorithms. Then, we revisit the PPARg
 7 application, using our inferential methods. Finally, we discuss general findings from
 8 this study, and also make connections to the broader task of assessment of virtual
 9 screening campaigns.

10 Methods

11 Evaluation of Ranking Algorithms

12 Let S denote the score from a ranking algorithm, where larger values of S suggest
 13 stronger belief that a ligand is active. S is reasonably regarded as a random vari-
 14 able. Activity of a ligand may also be regarded as a random variable: $X = I(\text{active})$,
 15 where $I(\cdot)$ is the indicator function. That is, $X = 1$ when a ligand belongs to the
 16 active class (+) and $X = 0$ when a ligand belongs to the inactive class (-). Let
 17 $P(X = 1) = \pi_+$. Given that a ligand is active, S has cumulative distribution
 18 function $F_+(s)$, and given that a ligand is not active, S has cumulative distribu-
 19 tion function $F_-(s)$. Combining the scores from both classes results in the mixture
 20 distribution $F_S(s) = \pi_+ F_+(s) + (1 - \pi_+) F_-(s)$.

21 Once ligands have been ranked according to their score, S , a threshold on the
 22 score, t , will prioritize a top fraction of the data set for testing. Let this top fraction
 23 be $r = P(S > t)$, which is the x-axis of the population hit enrichment curve. The
 24 hit enrichment curve (also known as the enrichment curve, accumulation curve,
 25 or percent captured response curve) is often used when an entire curve is used
 26 to evaluate a virtual screening campaign. Population hit enrichment curves plot
 27 $P(S > t|+)$ on the y-axis, where $P(S > t|+)$ is known as *recall* at threshold t .

28 So far, we have described the population level distributions of random variables
 29 X and S , that is, the distribution of ligand activity for the population of drug
 30 candidates from which a data set is sampled and the distribution of scores that a
 31 ranking algorithm would assign to them. We consider a data set under examination
 32 to be a random sample of activity and score pairs $\{(X_i, S_i); i = 1, \dots, n\}$ from
 33 these population level distributions. Let $\{S_i^+; i = 1, \dots, n^+\}$ be the scores that were
 34 sampled from the + class mixture component, $F_+(s)$, and $\{S_i^-; i = 1, \dots, n^-\}$ be
 35 the scores that were sampled from the - class mixture component, $F_-(s)$.

36 The empirical hit enrichment curve plots the cumulative fraction of actives on
 37 the y-axis, identified as a function of the top r fraction of ranked ligands. This
 38 means all compounds with scores beyond the percentile $100(1 - r)$ are “tested”
 39 and the cumulative fraction of actives determined. Another way of determining this
 40 percentile would be to choose a threshold $\hat{t}_r$ such that the fraction of items with
 41 $S > \hat{t}_r$ is r . We use $\hat{t}$ instead of t to denote that this threshold defines a fraction of
 42 the sample data and not the population.

Specifically, we define, $\hat{F}(\cdot)$ and $\hat{F}_+(\cdot)$ to be the empirical cumulative distribution functions (cdfs) for all and + scores:

$$\hat{F}(s) = \frac{1}{n} \sum_{i=1}^n I(S_i \leq s), \quad \hat{F}_+(s) = \frac{1}{n^+} \sum_{i=1}^{n^+} I(S_i^+ \leq s).$$

1 Given a testing fraction r , we ideally choose a threshold $\hat{t}_r$ to be the score percentile
 2 selected for testing such that $r = 1 - \hat{F}(\hat{t}_r)$. To accommodate the possible existence
 3 of ties in the observed data on scores, we define $\hat{t}_r = \min\{t : \hat{F}(t) \geq 1 - r\}$.
 4 Estimated recall at testing fraction r is the fraction of the active compounds that
 5 are correctly predicted to be active (i.e. $S_i^+ > \hat{t}_r$) and is thus obtained as $\hat{\theta}_r =$
 6 $1 - \hat{F}_+(\hat{t}_r) = 1 - \hat{F}_+(\hat{F}^{-1}(1 - r))$. Thus, the empirical hit enrichment curve plots
 7 the pairs $\{r, \hat{\theta}_r\}$.

8 Jiang and Zhao [14] have shown that if $\pi_+ \in (0, 1)$, then the empirical hit enrich-
 9 ment curve is an unbiased estimator of the population hit enrichment curve.

10 Compare Hit Enrichment Curves from Competing Algorithms

11 We wish to determine whether one ranking algorithm has significantly better
 12 performance than another at a given testing fraction r . For $r = 1 - F(t_r)$, let
 13 $\theta_r = P(S > t_r | +)$ denote the true population-level recall for a ranking algorithm
 14 at testing fraction r . Let $\{(X_i, S_{1i}, S_{2i}); i = 1, \dots, n\}$ be a random sample from the
 15 ligand activity distribution and the score distributions of ranking algorithm 1 and
 16 2. We assume that the triplets (X_i, S_{1i}, S_{2i}) are independent across i . However, S_{1i}
 17 and S_{2i} are likely correlated. The amount of correlation will depend on the extent
 18 to which the scores are measures of the same characteristics of the ligands. For
 19 example, competing docking scoring functions are often parameterized in similar
 20 ways (e.g., Glide SP and Glide XP; see [15]), and competing QSAR models often
 21 utilize highly correlated sets of descriptors. Using the random sample, we estimate
 22 the difference in performance between two algorithms at a given r , $\hat{\theta}_{1r} - \hat{\theta}_{2r}$, and
 23 perform a hypothesis test to determine if the difference is significant.

First considering a single algorithm, let $Q_r = \sum_{i=1}^n X_i I(S_i > \hat{t}_r)$ represent the number of active ligands that are examined at testing fraction r . We estimate recall at testing fraction r using $\hat{\theta}_r = Q_r / \sum_{i=1}^n X_i$, noting that both the numerator and denominator are random variables. The activity rate π_+ is reasonably estimated as $\hat{\pi}_+ = \sum_{i=1}^n X_i / n$, so an alternative expression for estimated recall is $\hat{\theta}_r = Q_r / (n\hat{\pi}_+)$. Because $\hat{t}_r$ is estimated using the entire dataset through the empirical cdf $\hat{F}(\cdot)$, Q_r is not binomially distributed. To properly account for this, Jiang and Zhao [14] take an empirical process approach to derive asymptotic normality of $\hat{\theta}_r$. Their result is that $\sqrt{n}(\hat{\theta}_r - \theta_r) \xrightarrow{d} N(0, \tau_{\theta_r}^2)$ for $r \in (0, 1)$ as $n \rightarrow \infty$, with corresponding asymptotic variance expression

$$\tau_{\theta_r}^2/n := \text{Var}_{JZ}(\hat{\theta}_r) = \text{Var}_B(\hat{\theta}_r) \left[1 - 2\Lambda_r + \frac{\Lambda_r^2(1-r)r}{\pi_+\theta_r(1-\theta_r)} \right], \quad (1)$$

where

$$\text{Var}_B(\hat{\theta}_r) = (n\pi_+)^{-1}\theta_r(1-\theta_r) \quad (2)$$

is the simple binomial variance, and $\Lambda_r = P(+|S = t_r)$ is a threshold-specific activity rate. This result assumes that $\pi_+ > 0$, and that the conditional densities $f_+(s)$ and $f_-(s)$ are positive and continuously differentiable in a neighborhood of $S = t_r$; henceforth called Conditions 1 and 2.

When it comes to comparing estimated recall across two competing algorithms, there are two sources of correlation that should be addressed. One source is correlation induced by needing to estimate t_r using $\hat{t}_r$, and that is addressed by using the result of Jiang and Zhao [14]. The other source of correlation arises because competing algorithm scores are derived using some common data, and this source of correlation has not been previously addressed in the literature. By accounting for both types of correlation, we expect to improve the power to detect real differences in algorithmic performance.

The following subsections present four methods of testing for significant differences in recall for two competing algorithms. First, we extend the approach of Jiang and Zhao [14] to use an empirical process approach that accounts for the correlation *between* algorithms, in addition to the correlation induced *within* a particular algorithm by estimation of t_r ; this method is called EmProc. Second, we present details on how a McNemar procedure for correlated proportions may be applied to the application; as far as we know, this has not been previously done. Third, we apply the Jiang and Zhao [14] result for hypothesis testing; this method is called IndJZ and is only optimized to address correlation within each algorithm but not correlation between algorithms. And fourth, we treat Q_{1r} and Q_{2r} as correlated binomial random variables, thus ignoring correlations induced by estimating t_r but accounting for correlations between algorithms; this method is called CorrBinom. The four methods are compared using a simulation study in section *Simulation Results*.

All four methods are based on asymptotic normality of a test statistic of the form

$$Z_r = \frac{\hat{\theta}_{1r} - \hat{\theta}_{2r}}{SE(\hat{\theta}_{1r} - \hat{\theta}_{2r})}.$$

We reject $H_{0r} : \theta_{1r} = \theta_{2r}$ if $|Z_r| > z_{\alpha/2}$, where $z_{\alpha/2} = 1.96$ for an $\alpha = .05$ level test. Pointwise confidence intervals are obtained as

$$(\hat{\theta}_{1r} - \hat{\theta}_{2r}) \pm z_{\alpha/2} SE(\hat{\theta}_{1r} - \hat{\theta}_{2r}). \quad (3)$$

The methods differ in their approach to estimating $Var(\hat{\theta}_{1r} - \hat{\theta}_{2r})$.

EmProc: Adjust for correlation between algorithms and correlation within each algorithm

Taking an empirical process approach, the functional delta method employed by Jiang and Zhao [14] was extended to derive the following asymptotic normality result concerning $\hat{\theta}_{1r} - \hat{\theta}_{2r}$.

Theorem 1: Given that Conditions 1 and 2 are satisfied for both θ_{1r} and θ_{2r} , then

$$\sqrt{n} \left\{ (\hat{\theta}_{1r} - \hat{\theta}_{2r}) - (\theta_{1r} - \theta_{2r}) \right\} \xrightarrow{d} N(0, \tau_{\theta_{1r}, \theta_{2r}}^2)$$

Table 1 Frequency distribution for the tested/not-tested status of active ligands according to both ranking algorithms, at a fixed testing fraction r .

	Algorithm 2 tested	Algorithm 2 not-tested	Row total
Algorithm 1 tested	$Q_{12 \cdot r}$	$Q_{1r} - Q_{12 \cdot r}$	Q_{1r}
Algorithm 1 not-tested	$Q_{2r} - Q_{12 \cdot r}$	$n\hat{\pi}_+ - (Q_{1r} + Q_{2r} - Q_{12 \cdot r})$	$n\hat{\pi}_+ - Q_{1r}$
Column total	Q_{2r}	$n\hat{\pi}_+ - Q_{2r}$	$n\hat{\pi}_+$

as $n \rightarrow \infty$. Furthermore, the asymptotic variance expression is

$$\tau_{\theta_{1r}, \theta_{2r}}^2/n := Var_{\text{EmProc}}(\hat{\theta}_{1r} - \hat{\theta}_{2r}) = Var_{JZ}(\hat{\theta}_{1r}) + Var_{JZ}(\hat{\theta}_{2r}) - 2Cov_{\text{EmProc}}(\hat{\theta}_{1r}, \hat{\theta}_{2r}), \quad (4)$$

where: $Var_{JZ}(\cdot)$ is as given in equation (1) and applied for each algorithm;

$$Cov_{\text{EmProc}}(\hat{\theta}_{1r}, \hat{\theta}_{2r}) = Cov_B(\hat{\theta}_{1r}, \hat{\theta}_{2r}) \left\{ (1 - \Lambda_{1r} - \Lambda_{2r}) + \frac{(\gamma_{12 \cdot r} - r^2)\Lambda_{1r}\Lambda_{2r}}{\pi_+(\theta_{12 \cdot r} - \theta_{1r}\theta_{2r})} \right\};$$

$$Cov_B(\hat{\theta}_{1r}, \hat{\theta}_{2r}) = (n\pi_+)^{-1}(\theta_{12 \cdot r} - \theta_{1r}\theta_{2r}) \quad (5)$$

1 is the covariance between binomial counts; $r = P(S_j > t_{jr})$ and determines the
 2 threshold for algorithm j , for $j = 1, 2$; $\theta_{12 \cdot r} = P(S_1 > t_{1r}, S_2 > t_{2r} | +)$ is the
 3 conditional probability that both algorithms result in testing an active ligand be-
 4 cause it is highly ranked by both algorithms; $\gamma_{12 \cdot r} = P(S_1 > t_{1r}, S_2 > t_{2r})$ is the
 5 unconditional probability that a ligand is highly ranked by both algorithms; and
 6 $\Lambda_{jr} = P(+ | S_j = t_{jr})$ for $j = 1, 2$. Details and derivations are in the Appendix.

7 To estimate $Var_{\text{EmProc}}(\hat{\theta}_{1r} - \hat{\theta}_{2r})$, we replace the population parameters with
 8 consistent estimators. The frequency distribution for the tested/not-tested status
 9 according to both ranking algorithms for the active ligands is shown in Table 1,
 10 where $Q_{jr} = \sum_{i=1}^n X_i I\{S_{ji} > \hat{t}_{jr}\}$ counts the number of active ligands tested by
 11 algorithm j for $j = 1, 2$, and $Q_{12 \cdot r} = \sum_{i=1}^n X_i I\{S_{1i} > \hat{t}_{1r}, S_{2i} > \hat{t}_{2r}\}$ counts the
 12 number of active ligands tested by both algorithms. As previously discussed, we
 13 estimate θ_{jr} with $\hat{\theta}_{jr} = Q_{jr}/(n\hat{\pi}_+)$. Additional estimates are obtained as $\hat{\gamma}_{12 \cdot r} =$
 14 $\sum_{i=1}^n I\{S_{1i} > \hat{t}_{1r}, S_{2i} > \hat{t}_{2r}\}/n$, $\hat{\theta}_{12 \cdot r} = Q_{12 \cdot r}/(n\hat{\pi}_+)$, and $\hat{\Lambda}_{jr}$ is obtained using
 15 Nadaraya-Watson kernel regression with the “rule-of-thumb” bandwidth selector
 16 [16].

17 Under the null hypothesis $H_{0r} : \theta_{1r} = \theta_{2r}$, we could alternatively use a pooled
 18 estimator of θ_{jr} for $j \in \{1, 2\}$, namely $\hat{\theta}_r = \frac{1}{2}(\hat{\theta}_{1r} + \hat{\theta}_{2r})$ to replace both $\hat{\theta}_{1r}$ and $\hat{\theta}_{2r}$
 19 in variance expression (4). We consider variance estimates using both the unpooled
 20 and pooled approaches.

21 McNemar’s test for difference in recall

When estimating recall, the same set of active ligands simultaneously serves as the
 set of “trials” for both ranking algorithms, with the decision being whether or not
 each algorithm selects the active ligand for testing. The consequence is that the
 data for both algorithms may be viewed as fully paired. The standard test used
 for paired proportions is McNemar’s test [17, 18]. Table 1 shows how the number
 of active ligands tested by either ranking algorithm can be written as a 2×2
 contingency table. The estimated recall values present themselves as the marginal

probabilities of testing an active ligand for each ranking algorithm. Consequently, the asymptotic McNemar test is based on test statistic

$$Z_r = \frac{((Q_{1r} - Q_{12\cdot r}) - (Q_{2r} - Q_{12\cdot r}))}{\sqrt{(Q_{1r} - Q_{12\cdot r}) + (Q_{2r} - Q_{12\cdot r})}} = \frac{(Q_{1r} - Q_{2r})}{\sqrt{Q_{1r} + Q_{2r} - 2Q_{12\cdot r}}} = \frac{(\hat{\theta}_{1r} - \hat{\theta}_{2r})}{\sqrt{Q_{1r} + Q_{2r} - 2Q_{12\cdot r}/(n\hat{\pi}_+)}}$$

and it assumes that discordant counts $(Q_{2r} - Q_{12\cdot r})$ and $(Q_{1r} - Q_{12\cdot r})$ are large. In a simulation study comparing the asymptotic McNemar test to a number of other tests for paired nominal data, the asymptotic McNemar test was found to be the most powerful across simulation scenarios, though slightly liberal in terms of type I error [18].

While the asymptotic McNemar test enforces the null condition $\theta_{1r} = \theta_{2r}$ to replace $Q_{1r} - Q_{2r}$ with zero in the variance expression, an alternative approach is needed to obtain pointwise confidence intervals. Pointwise Wald confidence intervals use the following standard error expression in equation (3):

$$SE(\hat{\theta}_{1r} - \hat{\theta}_{2r}) = \sqrt{Q_{1r} + Q_{2r} - 2Q_{12\cdot r} - (Q_{1r} - Q_{2r})^2/(n\hat{\pi}_+)/(n\hat{\pi}_+)}.$$

Unfortunately, several studies [18–20] demonstrate inadequate coverage properties of the Wald interval. The Bonett-Price [21] adjusted interval is a simple modification of the Wald interval, and it has been shown to have good coverage properties [18, 20]. We refer to the Bonett-Price adjustment as a “plus” adjustment because it adds one unit to each of the discordant counts shown in Table 1, then applies the Wald formula. More precisely, discordant count $Q_{1r} - Q_{12\cdot r}$ becomes $Q_{1r} - Q_{12\cdot r} + 1$ and discordant count $Q_{2r} - Q_{12\cdot r}$ becomes $Q_{2r} - Q_{12\cdot r} + 1$, thus adding one to each of the marginal counts and two to the overall total. As a result, the Bonett-Price plus interval is

$$\frac{Q_{1r} - Q_{2r}}{n\hat{\pi}_+ + 2} \pm z_{\alpha/2} \sqrt{\frac{1}{(n\hat{\pi}_+ + 2)^2} \left[(Q_{1r} + Q_{2r} - 2Q_{12\cdot r} + 2) - \frac{(Q_{1r} - Q_{2r})^2}{(n\hat{\pi}_+ + 2)} \right]},$$

noting that both the center point and the standard error have been adjusted.

IndJZ: Adjust for correlation within each algorithm but not correlation between algorithms

If we assume that $\hat{\theta}_{1r}$ and $\hat{\theta}_{2r}$ are independent, then $Var_{JZ}(\hat{\theta}_{1r} - \hat{\theta}_{2r}) = Var_{JZ}(\hat{\theta}_{1r}) + Var_{JZ}(\hat{\theta}_{2r})$, where $Var_{JZ}(\hat{\theta}_{jr})$ is obtained as in equation (1) for $j = 1, 2$. For testing equality of recall for the algorithms, either a pooled or unpooled estimator of the variance could be used, as previously discussed.

When the competing algorithms have scores that are highly positively correlated, it is expected that the IndJZ approach will lead to standard errors that are unnecessarily large, resulting in an underpowered test.

CorrBinom: Adjust for correlation between algorithms but not correlation within each algorithm

In this approach, we treat Q_{jr} as if it follows a simple binomial distribution, even though it does not. As a result, the relevant variance expression is $Var_B(\hat{\theta}_{1r} - \hat{\theta}_{2r}) =$

1 $Var_B(\hat{\theta}_{1r}) + Var_B(\hat{\theta}_{2r}) - 2Cov_B(\hat{\theta}_{1r}, \hat{\theta}_{2r})$, where $Var_B(\cdot)$ is defined in equation (2)
2 and $Cov_B(\cdot, \cdot)$ is defined in equation (5).

3 For testing equality of recall for the algorithms, either a pooled or unpooled esti-
4 mator of the variance could be used, as previously discussed.

5 *Simulation Results*

6 *Simulation of Benchmarking Data Sets* Geppert et al. [4] and Xia et al. [22] have
7 recently reviewed the standard data sets used to benchmark virtual screening tools.
8 The goal in designing a benchmark data set is to mimic real world chemical collec-
9 tions – this means that the score and activity distributions of the compounds in the
10 benchmark should resemble these populations. However there are frequently biases
11 in benchmark data sets.

12 One type of bias is *analogue bias* [23]. Data sets with known activities toward a
13 target often have very limited chemical scaffolds (or *chemotypes*) because they were
14 assembled by medicinal chemists for structure-activity relationship (SAR) studies.
15 In an actual prospective virtual screen the chemotypes are more diverse. Another
16 type of bias is *artificial enrichment* [24]. This occurs when the inactives do not
17 resemble the actives in terms of low dimensional physicochemical properties. This
18 results in an over simplified classification problem that does not reflect the true
19 complexity of the SARs present in a prospective virtual screen. The third type of
20 bias is *false negative bias* [25]. Molecules included in the benchmarking data sets
21 as inactives (also known as negatives or *decoys*) are often chemically similar to
22 the active ligands and have not been tested experimentally. In the past, this was
23 necessary because it was uncommon for a large data set of inactives to be available
24 in the commonly used chemical databases like ChEMBL [26].

25 Prospective virtual screens are typically conducted on large databases such as
26 ZINC [27]. These databases can be considered random samples from “drug like”
27 chemical space. The performance curves estimated by retrospective virtual screens
28 will likely mis-estimate performance in a prospective virtual screen if the benchmark
29 data set is not also a random sample. To this date, directory of useful decoys (DUD)
30 is the most widely used collection of benchmarking data sets used in the evaluation
31 of retrospective virtual screens, however each of the biases mentioned have been
32 observed in these data sets [22, 23, 28]. The directory of useful decoys, enhanced
33 (DUD-E) data sets [29] were developed to address some biases in these data sets,
34 though the data sets still lack the experimental testing of decoys (i.e., there is still
35 false negative bias) and there is an unrealistic frequency of actives included in each
36 of the data sets. The MUV benchmarking data sets [30] have also been developed
37 with the intention of minimizing these biases. A clear advantage over DUD-E is
38 that the decoys in MUV have been tested experimentally. The authors of MUV
39 collected 18 primary high-throughput screen assays from PCBioAssay [31]. Actives
40 were further confirmed with low throughput assays to minimize the number of false
41 positives, and additional checks for false negatives were performed. We modeled our
42 simulations on the MUV benchmarking data sets, because we believe these data sets
43 to be the most representative of the population of drug candidates.

44 Basing our simulations on MUV, we simulated data sets with $n = 150,000$, $\pi_+ =$
45 $.002$, and $skew = 499$ (where $skew$ is ratio of inactives to actives, $(1 - \pi_+)/\pi_+$).

In general, this is representative of a typical virtual screening data set with large sample sizes and extreme class imbalance.

Study Design A study was conducted to investigate the power of EmProc, McNemar, IndJZ, and CorrBinom to detect differences between competing algorithms. For each of EmProc, IndJZ, and CorrBinom, a total of four modifications were considered according to whether pooling was applied or whether the Bonett-Price plus adjustment was applied. Although the Bonett-Price plus adjustment was proposed for the Wald confidence interval to improve coverage probability, we wondered if it might also be used for hypothesis testing. McNemar was not modified. In fact, McNemar is identical to CorrBinom with pooling and without plus adjustment.

A study was also conducted to investigate coverage probabilities of the confidence intervals associated with EmProc, IndJZ, CorrBinom, and the Bonett-Price plus-adjusted Wald interval (the latter being referred to as the McNemar interval, for brevity). There is no justification for pooling when considering confidence intervals, so no pooling is applied. We continue to investigate the impact of the Bonett-Price plus adjustment. The McNemar interval is identical to the CorrBinom interval with plus adjustment.

Four data-generating mechanisms were considered: binormal or bibeta; and correlation of 0.9 or 0.1. In the binormal model, both algorithms have F_- as the standard normal distribution function, while Algorithm 1 has F_+ as the normal with mean $0.8\sqrt{2}$ and variance one, and Algorithm 2 has F_+ as the normal with mean $0.6\sqrt{2}$ and variance one. Algorithm 1 has a relatively large separation (Cohen's $D = 0.8$) between the score distributions of the + and - classes; Algorithm 2 has a slightly diminished performance (Cohen's $D = 0.6$). To capture the fact that both algorithms are scoring the same compounds, we simulated the positive scores from a bivariate normal with marginals as described above and correlation parameter $\rho = 0.1$ or $\rho = 0.9$. The negative scores were simulated from a separate bivariate normal with the described marginals and the same correlation parameter. In the bibeta model, both algorithms have F_- as the beta distribution with $\alpha = 2$ and $\beta = 5$, while Algorithm 1 has F_+ as the beta(5,2) and Algorithm 2 has F_+ as the beta(4,2). Sampling was done using the bivariate beta distributions to incorporate correlations similar to the binormal model. Sampling was conducted using the `copula` R package [32].

There is much greater separation between F_- and F_+ in the bibeta model, so the true hit enrichment curves are higher than in the binormal model, and we expect greater correlation of scores across early testing fractions. But for both the binormal and bibeta models, parameters were chosen to result in very similar hit enrichment curves for the two competing algorithms, thus creating a challenging task for hypothesis testing of differences between hit enrichment curves.

Studies were conducted using 10,000 Monte Carlo replicates. We estimated the type I error rate for the hypothesis test methods assuming that both ranking algorithms had either the score distributions of Algorithm 1 or Algorithm 2.

Results Pooling has no impact on either the power or protection from type I errors for EmProc, IndJZ, or CorrBinom (results not shown for brevity), so we limit further discussion to the versions of these tests constructed without pooling. The impact

of the Bonett-Price plus adjustment on power is mixed (results not shown). The plus adjustment had no noticeable impact on power under the bibeta model. But under the binormal model, the plus adjustment caused a noticeable decrease in power for a medium range of tests performed (approximately 30 to 500). For this reason we limit further discussion to the versions of these tests constructed without plus adjustment. As a reminder, McNemar is based on pooling and no plus adjustment.

Figure 2 shows estimated power curves for correlation of 0.1 (A) and 0.9 (B) under the bibeta model. We focus on results for number of tests (nr) ranging from two tests to testing ten percent of the total sample size; in practice, screening campaigns will interrogate only a tiny fraction of the available virtual screening library [8]. The CorrBinom and McNemar approaches are noticeably suboptimal. In the presence of weak correlation, EmProc and IndJZ have comparable performance. EmProc dominates in the presence of strong correlation between scores. Indeed, EmProc is the only approach that is designed to address both the correlation between scores of competing algorithms, and the correlation that is induced within a particular algorithm as a result of having to estimate thresholds. The binormal model (results not shown) yielded similar findings. Type I error rates (results not shown) are well controlled to their nominal values of 0.05.

For hypothesis testing, we recommend EmProc without pooling and without plus adjustment, because EmProc has the greatest power compared to IndJZ, CorrBinom, and McNemar, while maintaining control of type I error rates. If one chooses to use either IndJZ or CorrBinom, the unpooled and non-plus-adjusted versions should be used.

Figure 2 shows estimated coverage probabilities (C) and average widths (D) for confidence intervals EmProc, IndJZ, and CorrBinom, based on the binormal model with correlation of 0.9. The McNemar interval is equivalent to the CorrBinom interval with plus adjustment, so while the figure does not explicitly include the label of McNemar, it is included.

The most obvious finding is that the Bonett-Price plus adjustment dramatically improves coverage probabilities when the number of tests is small; this is because the plus adjustment results in wider intervals when the number of tests is small. As the number of tests increase, the plus and no-plus versions converge, and they approach nominal coverage. By not accounting for correlation across competing algorithms, IndJZ standard errors are unnecessarily large, resulting in wide intervals that provide conservative coverage. The plus-adjusted versions of EmProc and CorrBinom (and hence also McNemar) provide conservative coverage when the number of tests is small, but coverage approaches the nominal level as number of tests increase.

For confidence intervals, we recommend the plus-adjusted version of EmProc, because it is best able to balance achieving nominal coverage rates while minimizing the width of confidence intervals. And if other procedures are used, the plus adjustment should be used as well.

Confidence Bands

Bands for a Single Algorithm

Methods An ideal scenario would be to accompany hit enrichment curves, such as those shown in Figure 1, with confidence regions. Non-overlapping regions would provide an alternative justification for claiming significant differences between competing algorithms. Let θ denote the vector of recall values from a single algorithm at the vector $\mathbf{r} = (r_1, r_2, \dots, r_k)$ of k ordered testing fractions $r_1 < r_2 < \dots < r_k$. We seek a $100(1 - \alpha)$ percent confidence region for θ . While the pointwise confidence interval approach of equation (3) could be modified using a Bonferroni adjustment, such corrections are known to be conservative when k is large, leading to unnecessarily wide intervals.

In their technical report, Jiang and Zhao [33] suggested an alternate confidence band estimation procedure, and gave brief comments on simulation results, but some details were omitted. We complete these details to state the following result. Under the previously mentioned Conditions 1 and 2, as $n \rightarrow \infty$, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$, where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_{r_1}, \dots, \hat{\theta}_{r_k})$ is the vector of recall estimators as previously defined, and $\mathbf{V} = \{V_{ij}\}_{i,j=1,\dots,k}$. Moreover, $V_{ii}/n = \text{Var}_{JZ}(\hat{\theta}_{r_i})$, and, for $r_i < r_j$,

$$V_{ij}/n = \text{Cov}_{\text{EmProc}}(\hat{\theta}_{r_i}, \hat{\theta}_{r_j}) = \frac{\theta_{r_i}(1 - \theta_{r_j})}{n\pi_+} \left\{ (1 - \Lambda_{r_i} - \Lambda_{r_j}) + \frac{r_i(1 - r_j)\Lambda_{r_i}\Lambda_{r_j}}{\pi_+\theta_{r_i}(1 - \theta_{r_j})} \right\}. \quad (6)$$

Derivation details are omitted because they are similar to the steps in the appendix; see [34] for further details. To provide a working distribution for $\hat{\boldsymbol{\theta}}$, an estimator of $\boldsymbol{\theta}$ is obtained by replacing population parameters with consistent estimators. This working distribution is the basis of our approximate confidence regions.

Our most straight-forward approach is to use a Wald $100(1 - \alpha)$ percent confidence ellipsoid, defined as $\{\boldsymbol{\theta} : n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \hat{\mathbf{V}}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq \chi_{k,1-\alpha}^2\}$, where $\chi_{k,1-\alpha}^2$ is the $1 - \alpha$ percentile of the chi-squared distribution with k degrees of freedom. But the Wald confidence ellipsoid does not produce regions that are of the rectanguloid form $\{\boldsymbol{\theta} : \hat{\theta}_i \pm q \cdot SE(\hat{\theta}_i) \forall i \in \{1, \dots, k\}\}$. We have chosen to use confidence regions with a rectangular structure (i.e., a confidence band) and not ellipsoids because this allows confidence regions of high dimensions to be easily visualized.

Clearly, Bonferroni regions are rectanguloid, with $q = \sqrt{\chi_{1,1-\alpha/k}^2}$. We mention two additional rectanguloid regions, following the naming conventions in Montiel Olea and Plagborg-Møller [35]: the θ -projection and sup-t bands. θ -projection bands are obtained by identifying the smallest rectanguloid that contains the Wald ellipsoid, and results in $q = \sqrt{\chi_{k,1-\alpha}^2}$. Upon further inspection, it becomes clear that θ -projection bands are always at least as wide as Bonferroni bands, so they are not considered further. On the other hand, sup-t bands are the smallest rectanguloid that maintains the simultaneous coverage probability of $1 - \alpha$, and are expected to have good performance. Their critical value q must be obtained using Monte Carlo sampling. Briefly,

$$1 - \alpha \leq \Pr\left(|\hat{\theta}_i - \theta_i| \leq q \cdot SE(\hat{\theta}_i) \forall i \in \{1, \dots, k\}\right) = \Pr\left(\sup_{i=1,\dots,k} \frac{|\hat{\theta}_i - \theta_i|}{SE(\hat{\theta}_i)} \leq q\right).$$

Monte Carlo sampling is used to estimate q as the $(1 - \alpha)100$ percentile for the distribution of $\sup_{i=1,\dots,k} |\hat{\theta}_i - \theta_i|/SE(\hat{\theta}_i)$.

Simulation Results A study compared coverage probabilities achieved by confidence bands constructed using sup-t and Bonferroni approaches. Results are shown for both the standard and plus-adjusted versions of sup-t and Bonferroni bands. The familiar trick of “add two successes and add two failures” [36] before estimating proportions is what we refer to as the plus adjustment for bands corresponding

1 to a single hit enrichment curve, not the Bonett-Price plus adjustment for compar-
2 ing two algorithms that was used earlier. We computed bands for θ using a grid of
3 number of compounds tested between two and 15,000. We considered 25 points on
4 the grid, defined as: 2^k for $k = 1, \dots, 13$; 3^k for $k = 1, \dots, 8$; 105, 300, 1500, 15000.

5 Figure 3 shows estimated coverage probabilities (A) and average widths (B) of
6 confidence bands created based on five distributional cases. The distributional cases
7 represent varying degrees of separation between + and - classes, and are chosen to
8 mimic real-world scenarios. Case 1 is a binormal model, with equal unit variance and
9 means zero and 1.4. Case 2 is another binormal model, with equal unit variance and
10 means zero and 0.5; Case 2 offers much less separation than Case 1, so Case 2 results
11 in lower values of recall. Case 3 is a bibeta model, with Beta(2,5) and Beta(5,2)
12 distributions. Case 4 is another, more separated, bibeta model, with Beta(1,20)
13 and Beta(20,1) distributions. Case 5 is made up of distributions of limited extent,
14 namely uniform on (0,0.75) and uniform on (0.25,1).

15 The plus-adjusted Bonferroni bands have the highest coverage, but they are also
16 the widest. The plus-adjusted sup-t bands are not as wide, yet have excellent cov-
17 erage. As such, for confidence bands applied to a single hit enrichment curve, we
18 recommend the plus-adjusted sup-t bands.

1 Bands for the Difference Between Competing Algorithms

2 *Methods* While the pointwise confidence intervals offer effective comparisons of
 3 competing algorithms at a few selected testing fractions, it may be more desirable
 4 to perform comparisons across a large range of testing fractions. This may be ac-
 5 complished by converting the pointwise confidence intervals into confidence bands,
 6 in much the same way that confidence bands were obtained.

Letting θ_ℓ denote the vector of recall values from Algorithm ℓ ($\ell = 1, 2$), the method is based on the asymptotic result $\sqrt{n} \left((\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2) \right) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$, where $n \rightarrow \infty$, and for $r_i \leq r_j$,

$$V_{ij}/n = Cov_{EmProc}(\hat{\theta}_{1r_i}, \hat{\theta}_{1r_j}) + Cov_{EmProc}(\hat{\theta}_{2r_i}, \hat{\theta}_{2r_j}) - Cov_{EmProc}(\hat{\theta}_{1r_i}, \hat{\theta}_{2r_j}) - Cov_{EmProc}(\hat{\theta}_{1r_j}, \hat{\theta}_{2r_i}). \quad (7)$$

The first two components of equation (7) are obtained using equation (6), and the latter two components are obtained using

$$Cov_{EmProc}(\hat{\theta}_{1r_i}, \hat{\theta}_{2r_j}) = \frac{(\theta_{12 \cdot r_i r_j} - \theta_{1r_i} \theta_{2r_j})}{n\pi_+} \left\{ (1 - \Lambda_{1r_i} - \Lambda_{2r_j}) + \frac{(\gamma_{12 \cdot r_i r_j} - r_i r_j) \Lambda_{1r_i} \Lambda_{2r_j}}{\pi_+ (\theta_{12 \cdot r_i r_j} - \theta_{1r_i} \theta_{2r_j})} \right\}, \quad (8)$$

where $\theta_{12 \cdot r_i r_j} = P(S_1 > t_{1r_i}, S_2 > t_{2r_j} | +)$ and $\gamma_{12 \cdot r_i r_j} = P(S_1 > t_{1r_i}, S_2 > t_{2r_j})$ are the conditional and unconditional probabilities that both algorithms test a ligand because it is highly ranked by both algorithms, albeit at different testing fractions r_i and r_j . Equation (8) does not impose any restrictions between testing fractions r_i and r_j .

As described in Section *Bands for a Single Algorithm*, matrix $\mathbf{V}$ is estimated and used to construct sup-t and Bonferroni bands.

Simulation Results A study was conducted to compare coverage probabilities and average widths of confidence bands constructed using sup-t and Bonferroni approaches under four settings of two competing algorithms: binormal or bibeta, and $\rho = 0.1$ or 0.9 . Bands were computed for $\boldsymbol{\theta}$ using a grid of size 25, with number of tested compounds being: 2^k for $k = 1, \dots, 13$; 3^k for $k = 1, \dots, 8$; 105, 300, 1500, 15000.

Figure 3 shows estimated coverage probabilities (C) and average widths (D). Results are very similar to those observed for confidence bands for a single algorithm, namely that the plus-adjusted sup-t bands provide the best balance between coverage probabilities and average width.

For confidence bands applied to the difference between two hit enrichment curves, we recommend the plus-adjusted sup-t bands for achieving nominal coverage rates and minimizing width. The covariance used in constructing these bands arise from the EmProc approach.

Results and Discussion

We revisit the PPARg application. For testing fractions 0.001, 0.01, and 0.1, Table 2 provides details for all pairwise comparisons between scoring methods Surflex-dock (the best docking method), ICM (the worst docking method), and the maximum z-score (the best consensus method). For each of the three pairs, we provide the estimated difference between hit enrichment curves. Standard errors and resulting p -values are provided for the EmProc approach to conducting inference, and also for the remaining approaches McNemar, IndJZ, and CorrBinom.

Acknowledging the multiple-testing scenario required to compare two scoring methods, Table 2 also provides multiplicity-adjusted p -values. Given choice of a particular approach to inference (for example, EmProc), there are nine tests based on three pairs of scoring methods and three testing fractions. A simple Benjamini-Hochberg [37] step-up procedure is used to control the false discovery rate. At testing fraction 0.1 (321 tests), the difference between Surflex-dock and the consensus method is not significant, but ICM is significantly worse than both Surflex-dock and

1 the consensus method. These conclusions are clearly supported by all inferential ap-
2 proaches. At testing fraction 0.01 (32 tests), no significant differences are observed,
3 but the EmProc procedure is seen to produce the smallest standard errors. With
4 our relatively small dataset, testing fraction 0.001 results in only three tests, and
5 there is too much uncertainty to make a reliable conclusion.

6 Ignoring the need for multiplicity adjustments, Figure 4 provides information sim-
7 ilar to Table 2, except for many more testing fractions. The upper grid shows (unad-
8 justed) p -values from testing for equality of hit enrichment curves, while the lower
9 grid shows (unadjusted) pointwise 95 percent confidence intervals of differences be-
10 tween hit enrichment curves. EmProc is the only method that is consistently among
11 the best performers. The Pearson correlation is largest (but still only moderate at
12 0.624) between scores from Surflex-dock and the consensus method, so the poor
13 performance of IndJZ for this comparison is not surprising. On the other hand, the
14 performance of IndJZ understandably improves when comparing Surflex-dock and
15 ICM, with Pearson correlation of only 0.156 between scores.

16 The diagonal grid of Figure 4 shows estimated score densities within activity
17 classes for each scoring method. The consensus method has the biggest separation
18 of activity-class densities, with Kullback-Leibler divergence [38] of 2.50 compared to
19 the substantially smaller values of 0.00162 for surflex-dock and 2.21×10^{-5} for ICM,
20 and this is consistent with the consensus scoring method seeming to outperform the
21 others.

22 Taking an alternative approach, Figure 5 shows plus adjusted sup-t confidence
23 bands for the three scoring methods. These bands account for correlation between
24 recall values at distinct testing fractions for a single curve, and offer simultane-
25 ous coverage across the curve. While they do not account for correlations between
26 curves, they are helpful visualizations of uncertainty that go beyond simply graph-
27 ing the curves alone.

28 A more direct approach to pairwise comparisons between curves, while adjusting
29 for the many comparisons that occur along the curve, is shown in Figure 6. These
30 bands for the differences between hit enrichment curves offer simultaneous coverage
31 across the curve. They account for correlation between recall values at distinct
32 testing fractions for a single curve, and for correlation between estimated recall
33 from different scoring methods.

Table 2 Pairwise comparison of scoring methods surflex-dock (surf), ICM, and their consensus (maxz), at three testing fractions decided *a priori*. For each pair of scoring methods, differences and standard errors of estimated hit enrichment curves are shown, along with raw and Benjamini-Hochberg adjusted *p*-values.

	N tested = 3 ($r = 0.001$)				N tested = 32 ($r = 0.01$)				N tested = 321 ($r = 0.1$)			
	Diff	Std Err	Raw p	Adj p	Diff	Std Err	Raw p	Adj p	Diff	Std Err	Raw p	Adj p
EmProc												
maxz - surf	0.0000	0.0005	1.000	1.000	-0.0118	0.0237	0.6200	0.6970	0.0588	0.0254	2.07e-02	6.21e-02
maxz - icm	0.0118	0.0143	0.410	0.527	0.0824	0.0402	0.0407	0.0733	0.3060	0.0541	1.60e-08	1.44e-07
surf - icm	0.0118	0.0142	0.409	0.527	0.0941	0.0429	0.0281	0.0632	0.2470	0.0626	7.91e-05	3.56e-04
McNemar												
maxz - surf	0.0000	0.0000	1.000	1.000	-0.0118	0.0311	0.705	0.794	0.0588	0.0255	2.53e-02	7.60e-02
maxz - icm	0.0118	0.0203	0.564	0.725	0.0824	0.0557	0.144	0.260	0.3060	0.0552	2.07e-06	1.86e-05
surf - icm	0.0118	0.0203	0.564	0.725	0.0941	0.0614	0.131	0.260	0.2470	0.0642	3.86e-04	1.74e-03
IndJZ												
maxz - surf	0.0000	0.0138	1.000	1.000	-0.0118	0.0497	0.8130	0.915	0.0588	0.0609	3.34e-01	5.28e-01
maxz - icm	0.0118	0.0143	0.411	0.528	0.0824	0.0482	0.0874	0.197	0.3060	0.0668	4.74e-06	4.26e-05
surf - icm	0.0118	0.0143	0.409	0.528	0.0941	0.0471	0.0458	0.137	0.2470	0.0693	3.63e-04	1.64e-03
CorrBinom												
maxz - surf	0.0000	0.0000	1.000	1.000	-0.0118	0.0311	0.705	0.793	0.0588	0.0255	2.12e-02	6.35e-02
maxz - icm	0.0118	0.0203	0.563	0.724	0.0824	0.0557	0.139	0.251	0.3060	0.0552	3.07e-08	2.76e-07
surf - icm	0.0118	0.0203	0.563	0.724	0.0941	0.0614	0.125	0.251	0.2470	0.0642	1.20e-04	5.40e-04

1 Early enrichment is broadly recognized as a primary goal of virtual screening [9].
 2 There is, however, considerable debate about how to assess achievement of this goal
 3 [12]. The receiver operating characteristic (ROC) curve, and its summary metric
 4 of area under the curve, lack sensitivity to both the early recognition goal and the
 5 rarity of active compounds that often exists in screening studies [9, 39]. Partial area
 6 under the ROC curve makes a step towards addressing the early recognition goal
 7 but not the rarity of active compounds.

8 The robust initial enhancement (RIE, [40]) and its normalized version the Boltz-
 9 mann enhanced discrimination of ROC (BEDROC, [9]) directly address early recog-
 10 nition by incorporating an exponential weight that decreases for active compounds
 11 that are discovered later in testing. The versions of RIE and BEDROC that are pro-
 12 posed by Truchon and Bayly [9] are computationally efficient. They are, however,
 13 global measures that are not tied to a specific testing fraction, and they depend
 14 on a parameter α that must be specified by the user. A larger α provides greater
 15 weight to active ligands found early, but weights are applied to all active ligands,
 16 even those found very late. Moreover, statistical inference is not straightforward.

17 For the PPAR γ study using the default value of $\alpha = 20$, the Empereur-Mot et al.
 18 [6] web application reports BEDROC values as follows, where larger is better: 0.743
 19 for the consensus; 0.687 for Surflex-dock; and 0.447 for ICM. Without uncertainty

measures associated with these scores, it is difficult to conclude separation of the scoring methods. Furthermore, the results presented in Table 2 suggest conclusions that are more nuanced, with no significant differences at testing fraction 0.01 but some big differences at testing fraction 0.1. The improvement in BEDROC for the consensus method is mostly driven by the improvement near testing fraction 0.1, but this is a much larger fraction than what is typically used in screening evaluations. This illustrates how BEDROC can provide a misleading sense of improvement in early enrichment. It averages performance over all testing fractions and weights early fractions more heavily, but it is often difficult to determine from the tuning parameter alone to what extent the early fractions are weighted.

The hit enrichment curve offers multiple benefits for assessing early enrichment. First, it directly addresses the measure of interest, namely enrichment. For a given dataset, the closer the estimated hit enrichment curve is to the ideal hit enrichment curve, the more desirable is the associated algorithm. Second, the user is able to directly enforce their definition of “early” by specifying testing fractions of interest, without needing to rely on the indirect methods offered through measures such as BEDROC or partial area under the ROC curve. And third, localized assessment is possible, rather than whole-curve assessment. These are the reasons we have chosen to focus this work on hit enrichment curves.

Enrichment factors, and the associated enrichment factor curve, are also very popular for assessing early enrichment. As it turns out, all results in this paper can be trivially modified to obtain inference for the enrichment factor curve; see Ash [34] for further details. While the hit enrichment curve plots $\{r, \hat{\theta}_r\}$, the enrichment factor curve plots $\{r, \hat{\theta}_r/r\}$, so standard error expressions are easily modified by division by the testing fraction r . Enrichment factors (sometimes denoted EF_r) simply focus on the enrichment factor curve at a specific testing fraction.

In this article, we provide a template for rigorously comparing competing algorithms while accounting for uncertainty, two types of correlation, and the multiple

1 testing issue. If interest is restricted to comparing hit enrichment curves for com-
2 peting algorithms at a few pre-selected testing fractions, the hypothesis testing and
3 confidence interval procedures offer effective strategies. On the other hand, while
4 it is best practice to decide testing fractions *a priori*, we acknowledge this is not
5 always done or even possible, so we also provide confidence bands to compare entire
6 hit enrichment curves. Additionally, bands allow augmented graphical presentation
7 of the entire hit enrichment curves for competing algorithms.

8 The EmProc procedure is newly proposed here and is expected to perform as well
9 or better than the other procedures considered (CorrBinom, McNemar, and IndJZ)
10 in the presence of correlation between competing algorithms and/or correlation
11 across different testing fractions within a single algorithm. The other procedures
12 considered address only a single type of correlation and hence are not recommended
13 for general-purpose use.

14 Additional Files

15 Additional file 1

16 appendix.pdf: Proof of Theorem 1

17 Author's contributions

18 JRA and JHO conceived, developed and implemented the method, performed the calculations, and wrote the
19 manuscript.

20 Funding

21 This work was supported by NIH training grant T32ES007329; the Triangle Center for Evolutionary Medicine; and a
22 graduate research assistantship from SAS Institute.

23 Availability of data and materials

24 The inferential procedures and confidence band plotting functions are conveniently made available in R package
25 chemmodlab which can be installed from the following github repository: <https://github.com/jrash/chemmodlab>

1 The full data and code required to recreate the results of this paper are available in the following github repository:
2 <https://github.com/jrash/Enrichment-Inference-Supplemental-Materials>

3 Competing interests

4 The authors declare that they have no competing interests.

5 Acknowledgements

6 Not applicable

7 Author details

8 ¹Department of Statistics, Bioinformatics Research Center, North Carolina State University, Raleigh, United States.

9 ²JMP Division, SAS Institute, Cary, United States.

10 References

- 11 1. Kuhn, M.: Building predictive models in R using the caret package. *Journal of Statistical Software* **28**(5), 1–26
12 (2008). doi:[10.18637/jss.v028.i05](https://doi.org/10.18637/jss.v028.i05)
- 13 2. SAS Institute Inc.: SAS Enterprise Miner 15.1, Cary, NC (2020)
- 14 3. SAS Institute Inc.: JMP 16.0, Cary, NC (2021)
- 15 4. Geppert, H., Vogt, M., Bajorath, J.: Current trends in ligand-based virtual screening: Molecular representations,
16 data mining methods, new application areas, and performance evaluation. *Journal of Chemical Information and*
17 *Modeling* **50**(2), 205–216 (2010). doi:[10.1021/ci900419k](https://doi.org/10.1021/ci900419k)
- 18 5. Rosset, S., Neumann, E., Eick, U., Vatnik, N., Idan, I.: Evaluation of Prediction Models for Marketing
19 Campaigns, pp. 456–461. ACM Press, New York, NY (2001). doi:[10.1145/502512.502581](https://doi.org/10.1145/502512.502581)
- 20 6. Empereur-Mot, C., Zagury, J.-F., Montes, M.: Screening Explorer—An Interactive Tool for the Analysis of
21 Screening Results. *Journal of Chemical Information and Modeling* **56**(12), 2281–2286 (2016).
22 doi:[10.1021/acs.jcim.6b00283](https://doi.org/10.1021/acs.jcim.6b00283). (Web application at <http://stats.drugdesign.fr>)
- 23 7. NCBI <https://www.ncbi.nlm.nih.gov/gene/5468> (2021)
- 24 8. Zhu, T., Cao, S., Su, P.-C., Patel, R., Shah, D., Chokshi, H.B., Szukala, R., Johnson, M.E., Hevener, K.E.: Hit
25 identification and optimization in virtual screening: Practical recommendations based on a critical literature
26 analysis. *Journal of Medicinal Chemistry* **56**(17), 6560–6572 (2013). doi:[10.1021/jm301916b](https://doi.org/10.1021/jm301916b)
- 27 9. Truchon, J.-F., Bayly, C.I.: Evaluating virtual screening methods: Good and bad metrics for the “early
28 recognition” problem. *Journal of Chemical Information and Modeling* **47**(2), 488–508 (2007).
29 doi:[10.1021/ci600426e](https://doi.org/10.1021/ci600426e)
- 30 10. Jain, A.N., Nicholls, A.: Recommendations for evaluation of computational methods. *Journal of*
31 *Computer-Aided Molecular Design* **22**(3–4), 133–139 (2008). doi:[10.1007/s10822-008-9196-5](https://doi.org/10.1007/s10822-008-9196-5)
- 32 11. Nicholls, A.: Confidence limits, error bars and method comparison in molecular modeling. part 1: the calculation
33 of confidence intervals. *Journal of computer-aided molecular design* **28**(9), 887–918 (2014)
- 34 12. Robinson, M.C., Glen, R.C., Lee, A.A.: Validating the validation: reanalyzing a large-scale comparison of deep
35 learning and machine learning models for bioactivity prediction. *Journal of computer-aided molecular design*,
36 1–14 (2020). doi:[10.1007/s10822-019-00274-0](https://doi.org/10.1007/s10822-019-00274-0)
- 37 13. Hawkins, P.C.D., Warren, G.L., Skillman, A.G., Nicholls, A.: How to do an evaluation: pitfalls and traps.
38 *Journal of Computer-Aided Molecular Design* **22**(3–4), 179–190 (2008). doi:[10.1007/s10822-007-9166-3](https://doi.org/10.1007/s10822-007-9166-3)
- 39 14. Jiang, W., Zhao, Y.: On asymptotic distributions and confidence intervals for lift measures in data mining.
40 *Journal of the American Statistical Association* **110**(512), 1717–1725 (2015).
41 doi:[10.1080/01621459.2014.993080](https://doi.org/10.1080/01621459.2014.993080)
- 42 15. Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H.,
43 Shelley, M., Perry, J.K., Shaw, D.E., Francis, P., Shenkin, P.S.: Glide: A new approach for rapid, accurate
44 docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry* **47**(7),
45 1739–1749 (2004). doi:[10.1021/jm0306430](https://doi.org/10.1021/jm0306430)
- 46 16. Li, Q., Racine, J.S.: *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, New York,
47 NY (2007)
- 48 17. Agresti, A.: *An Introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics. John Wiley
49 & Sons, Inc., Hoboken, NJ (2007)
- 50 18. Fagerland, M.W., Lydersen, S., Laake, P.: The McNemar test for binary matched-pairs data: mid-p and
51 asymptotic are better than exact conditional. *BMC Medical Research Methodology* **13**(1), 91 (2013).
52 doi:[10.1186/1471-2288-13-91](https://doi.org/10.1186/1471-2288-13-91)
- 53 19. Newcombe, R.G.: Improved confidence intervals for the difference between binomial proportions based on
54 paired data. *Statistics in Medicine* **17**, 2635–2650 (1998).
55 doi:[10.1002/\(SICI\)1097-0258\(19981130\)17:22<2635::AID-SIM954>3.0.CO;2-C](https://doi.org/10.1002/(SICI)1097-0258(19981130)17:22<2635::AID-SIM954>3.0.CO;2-C)
- 56 20. Rodriguez de Gil, P., Pham, J.R.T., Nguyen, D., Kromrey, J.D., Kim, E.S.: SAS Macros CORR-P and TANGO:
57 Interval Estimation for the Difference Between Correlated Proportions in Dependent Samples. In: *Proceedings*
58 *of the SouthEast SAS Users Group 2013* (2013)
- 59 21. Bonett, D.G., Price, R.M.: Adjusted Wald confidence interval for a difference of binomial proportions based on
60 paired data. *Journal of Educational and Behavioral Statistics* **37**(4), 479–488 (2012).
61 doi:[10.3102/1076998611411915](https://doi.org/10.3102/1076998611411915)
- 62 22. Xia, J., Tilahun, E.L., Reid, T.-E., Zhang, L., Wang, X.S.: Benchmarking methods and data sets for ligand
63 enrichment assessment in virtual screening. *Methods* **71**, 146–157 (2015). doi:[10.1016/j.ymeth.2014.11.015](https://doi.org/10.1016/j.ymeth.2014.11.015)
- 64 23. Good, A.C., Oprea, T.I.: Optimization of CAMD techniques 3. virtual screening enrichment studies: a help or
65 hindrance in tool selection? *Journal of Computer-Aided Molecular Design* **22**(3–4), 169–178 (2008).
66 doi:[10.1007/s10822-007-9167-2](https://doi.org/10.1007/s10822-007-9167-2)
- 67 24. Stumpfe, D., Bajorath, J.: *Applied Virtual Screening: Strategies, Recommendations, and Caveats*, pp. 291–318
68 (2011). doi:[10.1002/9783527633326.ch11](https://doi.org/10.1002/9783527633326.ch11)

25. Bauer, M.R., Ibrahim, T.M., Vogel, S.M., Boeckler, F.M.: Evaluation and optimization of virtual screening workflows with dekois 2.0 – a public library of challenging docking benchmark sets. *Journal of Chemical Information and Modeling* **53**(6), 1447–1462 (2013). doi:[10.1021/ci400115b](https://doi.org/10.1021/ci400115b)
26. Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Motow, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M.P., Overington, J.P., Papadatos, G., Smit, I., Leach, A.R.: The ChEMBL Database in 2017. *Nucleic Acids Research* **45**(D1), 945–954 (2017). doi:[10.1093/nar/gkw1074](https://doi.org/10.1093/nar/gkw1074)
27. Sterling, T., Irwin, J.J.: Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling* **55**(11), 2324–2337 (2015). doi:[10.1021/acs.jcim.5b00559](https://doi.org/10.1021/acs.jcim.5b00559)
28. Irwin, J.J.: Community benchmarks for virtual screening. *Journal of Computer-Aided Molecular Design* **22**(3–4), 193–199 (2008). doi:[10.1007/s10822-008-9189-4](https://doi.org/10.1007/s10822-008-9189-4)
29. Mysinger, M.M., Carchia, M., Irwin, J.J., Shoichet, B.K.: Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry* **55**(14), 6582–6594 (2012). doi:[10.1021/jm300687e](https://doi.org/10.1021/jm300687e)
30. Rohrer, S.G., Baumann, K.: Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. *Journal of Chemical Information and Modeling* **49**(2), 169–184 (2009). doi:[10.1021/ci8002649](https://doi.org/10.1021/ci8002649)
31. NCBI Resource Coordinators.: Database resources of the national center for biotechnology information. *Nucleic Acids Research* **44**(D1), 7–19 (2016). doi:[10.1093/nar/gkv1290](https://doi.org/10.1093/nar/gkv1290)
32. Hofert, M., Kojadinovic, I., Maechler, M., Yan, J.: Copula: Multivariate Dependence with Copulas. (2020). R package version 1.0-1. <https://CRAN.R-project.org/package=copula>
33. Jiang, W., Zhao, Y.: Some technical details on confidence intervals for lift measures in data mining. Technical report (2014)
34. Ash, J.R.: Methods Development for Quantitative Structure-Activity Relationships. North Carolina State University, Raleigh, NC (2020)
35. Montiel Olea, J.L., Plagborg-Møller, M.: Simultaneous confidence bands: Theory, implementation, and an application to svars. *Journal of Applied Econometrics* **34**(1), 1–17 (2019). doi:[10.1002/jae.2656](https://doi.org/10.1002/jae.2656)
36. Agresti, A., Coull, B.A.: Approximate is better than ‘exact’ for interval estimation of binomial proportions. *The American Statistician* **52**(2), 119–126 (1998). doi:[10.2307/2685469](https://doi.org/10.2307/2685469)
37. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995). doi:[10.2307/2346101](https://doi.org/10.2307/2346101)
38. Kullback, S., Leibler, R.A.: On information and sufficiency. *The Annals of Mathematical Statistics* **22**(1), 79–86 (1951). doi:[10.1214/aoms/1177729694](https://doi.org/10.1214/aoms/1177729694)
39. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10**(3), 0118432 (2015). doi:[10.1371/journal.pone.0118432](https://doi.org/10.1371/journal.pone.0118432)
40. Sheridan, R.P., Singh, S.B., Fluder, E.M., Kearsley, S.K.: Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *Journal of Chemical Information and Computer Sciences* **41**(5), 1395–1406 (2001). doi:[10.1021/ci0100144](https://doi.org/10.1021/ci0100144)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementalmaterials.pdf](#)