Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem

Jean-François Truchon* and Christopher I. Bayly

Department of Medicinal Chemistry, Merck Frosst Centre for Therapeutic Research, 16711 TransCanada Highway, Kirkland, Québec, Canada H9H 3L1

Received October 2, 2006

Many metrics are currently used to evaluate the performance of ranking methods in virtual screening (VS), for instance, the area under the receiver operating characteristic curve (ROC), the area under the accumulation curve (AUAC), the average rank of actives, the enrichment factor (EF), and the robust initial enhancement (RIE) proposed by Sheridan et al. In this work, we show that the ROC, the AUAC, and the average rank metrics have the same inappropriate behaviors that make them poor metrics for comparing VS methods whose purpose is to rank actives early in an ordered list (the "early recognition problem"). In doing so, we derive mathematical formulas that relate those metrics together. Moreover, we show that the EF metric is not sensitive to ranking performance before and after the cutoff. Instead, we formally generalize the ROC metric to the early recognition problem which leads us to propose a novel metric called the Boltzmann-enhanced discrimination of receiver operating characteristic that turns out to contain the discrimination power of the RIE metric but incorporates the statistical significance from ROC and its well-behaved boundaries. Finally, two major sources of errors, namely, the statistical error and the "saturation effects", are examined. This leads to practical recommendations for the number of actives, the number of inactives, and the "early recognition" importance parameter that one should use when comparing ranking methods. Although this work is applied specifically to VS, it is general and can be used to analyze any method that needs to segregate actives toward the front of a rank-ordered list.

1. INTRODUCTION

Virtual screening (VS) is the term applied to a type of problem which is conceptually the computational equivalent of (experimental) high-throughput screening, wherein a large number of samples are quickly assayed to discriminate active samples from inactive samples. Evaluating the performance of VS methods is a necessary practice that both the method developers and the end-users perform. For the developers, it is done to parametrize and validate the methods, for the end users, it is a way to select which method performs best in a given situation. Recently, many pharmaceutical companies have conducted large evaluations of commercial software against a diverse set of targets;[1−7] an interesting book chapter that reviews the major software evaluations is being printed.[8] A great deal of attention has been focused in finding the right biological targets and the right database of compounds, since the character of both strongly impacts the results. Ultimately, the results are reported as a single number that estimates the ability of a VS method to retrieve active compounds out of a mixed set of active compounds and decoys (compounds presumably inactive against the examined target). It has been recently stated[9] that "medicinal and computational chemists suffer from a lack of standard methods to evaluate the accuracy of a newly designed *in silico* assay". Indeed, there are many metrics that are currently in use, often borrowed from other fields. However, it is surprising that almost none of these metrics address the "early recognition" problem specific to VS. The key requirement for success in VS is that it must rank actives very early in the larger set of compounds, since only a very small proportion of the compounds ranked will actually be tested experimentally. For instance, sample repositories of pharmaceutical companies commonly contain more than 1 million compounds. Identifying a set of a few thousand compounds to screen experimentally requires that the VS method perform well within the first 0.1% of the scored compounds! Even if the VS method is excellent at retrieving all actives for any target within the first half of the data set, it is useless if the early performance is bad. It is also surprising that, in other fields where the early recognition problem applies, such as information retrieval, the metrics used are not particularly good at early recognition.[10]

The area under the receiver operating characteristic (ROC) curve is used by influential groups[3,9,11−13] to measure VS performance in part because it does possess desirable statistical behaviors, and it has been widely used in other fields. For example, it is claimed to be independent of the ratio of actives and decoys; it has a value of 1/2 if the ranking method does not do better than random picking; it can be interpreted as the probability that an active will be ranked before an inactive; it has a value between 0 (worst performance attainable) and 1 (best performance). However, recent works have shown that some of these advertised properties are not fully realized.[14,15] Along those lines, in this work, we show that the area under the ROC curve, noted *ROC* (in this paper, we use italics to note the scalar metric), does depend on the ratio of actives to inactives. More importantly,

* Corresponding author tel.: (514) 428-3144; fax: (514) 428-4930; e-mail: jeanfrancois_truchon@merck.com.

10.1021/ci600426e CCC: $37.00
Published on Web 02/09/2007

the *ROC* metric is clearly a bad metric to discriminate among VS methods because it is not sensitive to early recognition. Consider three basic cases: (1) half of the actives are retrieved at the very beginning of the rank-ordered list and the other half at the end; (2) the actives are randomly distributed all across the ranks; (3) all of the actives are retrieved in the middle of the list. In all three cases, the *ROC* metric is 1/2 when, in terms of the "early recognition", case 1 is clearly better than case 2, which is also significantly better than case 3. In this paper, we give a mathematical proof that shows that the *ROC* metric corresponds to a linearly scaled average of the positions of the actives without preference for those that are found early in the rank-ordered list. There are several other metrics available: the area under the accumulation curve (*AUAC*), the average position of the actives,[2] analysis of variance (*ANOVA*),[16] the Z-score,[17] the enrichment factor (*EF*),[18] the robust initial enhancement (*RIE*),[7,19] and so forth. In the case of *EF*, it has been criticized for its lack of discrimination before the retrieval threshold and the absence of consideration after the retrieval threshold.[2,7,9,19] Also, the maximum value for *EF* is strongly dependent on the number of actives and inactives.[9] We will see that *AUAC* suffers from the same liabilities as *ROC*, the other "area under the curve" metric. The *ANOVA* and Z-score metrics are familiar from the field of statistics. They are in part based on the averaged position of actives and the spread of their distributions in the rank-ordered list. Of all of these, *RIE* from Sheridan et al.[19] is an "early recognition" metric which addresses these shortcomings in *EF*, but it lacks the advantages of *ROC*: For *ROC*, the limiting values (0−1) are independent of the number of actives, and the probalistic interpretation is well-defined. To address these deficiencies, we develop a novel metric, called the Boltzmann-enhanced discrimination of receiver operating characteristic (*BEDROC*), which is derived as a generalization of *ROC*, but which addresses the "early recognition" problem as well as *RIE*.

It is important to state our position regarding the precedent usage of metrics in VS method evaluation. Metrics such as *ROC* are reassuring because they directly address the question: Is this ranking method bringing added value compared to the random picking of compounds? This is a relevant question, but for VS, it is more important to evaluate if the ranking method can help us find leads in real pharmaceutical programs. Coming back to our previous example, if no lead shows up until 20% of the compound list is screened, even if then the ranking method finds all of the actives and gives a superb *ROC* of 0.8, this is not going to really add value to our work if experimentally we can only screen less than 20% of the list. This is an oversimplified example, but the argument holds for performance evaluation in general. Therefore, an appropriate metric needs to favor ranking methods that have a higher tendency to rank actives early in the list. This is the motivation to generalize *ROC* to the "early recognition" case.

In what follows, we start by describing the sampling method that was used to study the behavior of the seven metrics encompassed in the current study: *ROC*, *AUAC*, the average rank of actives, *EF*, weighted *AUAC* (*wAUAC*), *RIE*, and *BEDROC*. We review each metric, and more importantly, we propose a probability theory framework to establish mathematical relationships in between all of the metrics, unifying their understanding. Next, we examine the statistical variance of the metrics and derive useful analytical formulas. Errors due to the "saturation effect" are then characterized, and finally we outline heuristics for the use of *BEDROC*, which is the best metric adapted to the "early recognition" problem.

2. METHODS

Probability Distribution Function in Performance Evaluation Metrics. The concept of a probability distribution function (PDF), which we write as $f(x)$, is a basic and important notion in probability theory. In the case of VS, x represents the normalized rank in the ordered list that is a real number going from 0 to 1 obtained by dividing the rank of a compound by the total number of compounds. In a continuous world, the PDF is defined such that the probability that a random variable (normalized rank of an active in VS) takes a value between x and $x + \mathrm{d}x$ (normalized rank range) is given by $f(x)\,\mathrm{d}x$. In a real virtual screen, one can build an empirical PDF by binning the fractional occurrence of actives. A cumulative distribution function (CDF), which we note as $F(x)$, is the probability that a random variable takes a value between 0 and x. In the context of VS, the resulting empirical CDF is often referred to as the accumulation curve or the enrichment curve. Simply put, this is the cumulative number of actives found when going through the rank-ordered list, from best to worst, up to a normalized rank x. A CDF is simply the integral from 0 to x of the PDF, and the PDF is the derivative of the CDF at x, and both mathematical constructs contain all the information about a single random variable. These definitions directly link a VS method to probability theory. A curious reader can find examples of CDF and PDF elsewhere,[16] although Seifert uses the score itself and not the rank. In VS, either the CDF or the PDF of the actives contains all the information about the performance of the virtual screening method. In VS performance evaluations, a specific CDF for each biological target is obtained, which depends on the VS algorithm parameters, actives, decoys, and so forth. With that in mind, an evaluation constitutes a sample from an immense potential set, and this leads to an intrinsic variability in the metric results. This is in part due to the specific choice of actives and decoys. In practice, the CDF we obtain can also be tainted by other kinds of bias not addressed in this work, such as molecular structure similarity in the choice of the actives. The role of a VS evaluation is to first get the best CDF estimate possible and to summarize this function as a single number most relevant for the question being asked. The "early recognition" aspect of the problem requires particular characteristics of a metric as discussed above.

Simulation of Ranking Experiments. In the present article, we derive formulas and relationships between metrics and we examine the effects of varying different parameters on the quality of the metric used to judge the quality of a ranking method. Our analyses require that real ranking situations be simulated through Monte Carlo sampling of a hypothetical idealized CDF. Given the usual form of CDF in ranking evaluation, literature precedence,[20,21] and mathematical simplicity, we decided to use an exponential PDF given by eq 1 with its corresponding CDF given by eq 2.

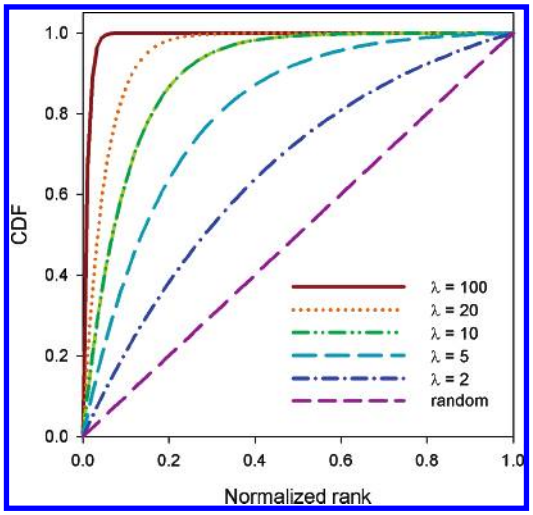

Figure 1. Cumulative density function (CDF) curves calculated from eq 2. A large value of λ corresponds to good performance and a small value of λ to poor performance; random performance is obtained with $F(x) = x$.

One advantage of the exponential distribution is that it can mimic both extremes: actives appearing at the very front of the list to actives appearing randomly in the list. The mean and the standard deviation of the PDF distribution are both given by $1/\lambda$. Hence, a larger λ corresponds to a better ranking method. The CDFs of exponential distributions with various λ parameters are shown in Figure 1.

$$f(x) = \frac{\lambda\, e^{-\lambda x}}{(1 - e^{-\lambda})} \tag{1}$$

$$F(x) = \frac{1 - e^{-\lambda x}}{1 - e^{-\lambda}} \tag{2}$$

In order to simulate a ranking experiment in which n actives are ranked among N compounds in total ($N - n$ inactives) with a PDF given by an exponential distribution of parameter λ, we used the inverse transformation method,[22] which relates a uniform (0,1) random variable U to a random variable X with a CDF given by F; the relationship is given by eq 3.

$$F(x) = \frac{1 - e^{-\lambda x}}{1 - e^{-\lambda}}$$

$$\Rightarrow F^{-1}(F) = \frac{-1}{\lambda} \ln[1 - F(1 - e^{-\lambda})]$$

$$X = \frac{-1}{\lambda} \ln[1 - U(1 - e^{-\lambda})] \tag{3}$$

In practice, we produce pseudo-random numbers U with values between 0 and 1. The real value X is obtained with eq 3, and X has the desired exponential distribution. The value X corresponds to a relative position of an active and is therefore transformed into a rank integer between 1 and N. This becomes the position of the first active in the set. The process is repeated until n new positions are obtained. One might be interested, instead, in the uniform distribution (0,1), and this is simply obtained directly from the pseudo-random number generator. A single rank can be occupied by only one active, and whenever a clash occurs, the proposed position is ignored and a new random number is generated.

Typical results from this kind of simulation are shown in Figure 2 where an exponential PDF with $\lambda = 20$ (burgundy) and a uniform PDF (orange) are sampled. We can assess how well we can approximate the ideal CDFs (dashed lines) with a finite sample of actives and decoys. For instance, we see some variability around the ideal CDFs (dashed lines) when we simulate only 10 actives among 2000 compounds (Figure 2a) and much less variability for 50 actives among 10 000 (Figure 2b). This variability is a feature that will allow us, later in this paper, to quite precisely calculate the variance of the different metrics.

3. METRICS

In this section, after defining *ROC* and *AUAC*, we examine their formal inter-relationships to improve our understanding of their strengths and weaknesses. Also, we express the *ROC*

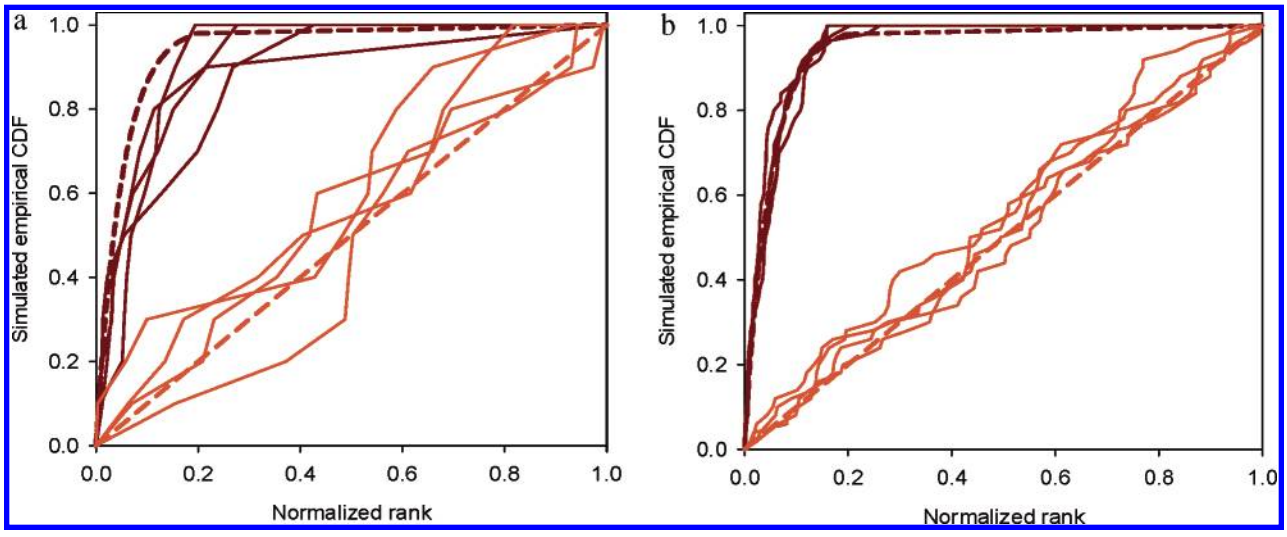

Figure 2. Accumulation curves obtained with the inverse transformation method for an exponential distribution with parameter $\lambda = 20.0$ (burgundy) and a uniform distribution (orange). The ideal CDFs are shown in dashed lines, and four simulated active distributions are illustrated with solid lines for two situations: (a) 10 actives among 2000 compounds and (b) 50 actives among 10 000 compounds.

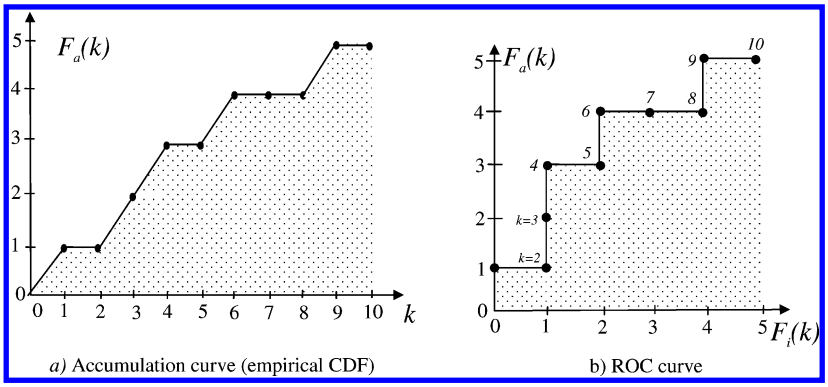

Figure 3. Accumulation curve (empirical CDF) (a) and receiver operating characteristic curve (b) for the same situation in which five actives among 10 compounds are found at rank $k = 1$, 3, 4, 6, and 9. The *AUAC* (a) is 0.59, and the *ROC* metric (b) is 0.68. We can calculate the *ROC* in b from the *AUAC* in a with $ROC = AUAC/R_i - R_a/(2R_i) = 0.59 \times 2 - 1/2 = 0.68$ (eq 11).

and the *AUAC* metrics as a function of the average rank of actives, recently suggested to be superior to *EF* by Kairys et al.[2] We then introduce the *BEDROC* metric as a logical consequence of the analysis, incorporating the notion of early recognition into the *ROC* metric formalism.

Area under the Accumulation Curve. Accumulation curves (corresponding to an empirical CDF as defined above) are widely used to display ranking performances. However, the corresponding area under the curve, the *AUAC*, is not as often used partly because it is believed to be largely dependent on the ratio of actives in the set. It is based on the empirical CDF where on the abscissa is the relative rank, *x*, and on the ordinate is the cumulative fractional count of actives retrieved up to *x* when the compounds are examined from best to worst according to a scoring or ranking method. We note this empirical CDF, $F_a(x)$, which is the probability that an active will be found before the relative rank *x*. With the empirical CDF in hand, the *AUAC* is obtained by calculating the area under the curve as shown in eq 4.

$$AUAC = \int_0^1 F_a(x)\,dx \tag{4}$$

Here, $F_a(x)$ is normalized and goes from the origin to the couple (1,1). The best value for the continuous *AUAC* is 1; the worst value is 0, and if the actives are uniformly distributed in the ranked list, the expected *AUAC* becomes 0.5. On the basis of eq 4, the *AUAC* can be interpreted as the probability that an active, selected from the empirical CDF defined by the rank-ordered list, will be ranked before a compound randomly selected from a uniform distribution. This comes from the fact that a uniform PDF is given by $f(x) = 1$, the multiplicative function of $F_a(x)$ in eq 4, whose interpretation is already common practice in the probability analysis of the *ROC* formalism (see eq 10 and related references). For the discrete case, the *AUAC* metric can be calculated with eq 5, using the trapezoid rule, where *k* is used throughout this paper to indicate the actual rank of the active. In eq 5 as opposed to eq 4, $F_a(k)$ and *k* are not normalized, hence the division by *nN*. We adopt this convention in all of our equations: the continuous formulas use the relative rank and scaled CDF or PDF; the discrete formulas use nonscaled quantities. The maximum value for the discrete *AUAC* is related to the number of actives and inactives since, if all of the actives are aligned at the beginning of the list, the maximum value for the *AUAC* is $1 - n/(2N)$, which tends to 1 as n/N gets smaller; the minimum value is $n/(2N)$ if all of the actives are ranked at the end of the list. An example of an empirical CDF is shown in Figure 3a for a case where five actives are ranked at positions 1, 3, 4, 6, and 9 among a total of 10 compounds; the *AUAC* calculates to $29.5/50 = 0.59$ according to eq 5.

$$AUAC = \frac{1}{2nN}\sum_{k=0}^{N-1}[F_a(k) + F_a(k+1)] \tag{5}$$

In appendix A.1, we derive an alternative equation for the *AUAC* as a function of the average relative rank of the actives, noted $\langle x \rangle$, shown by eqs 6 and 7 where r_i is the rank of the *i*th active in the list and x_i is the relative rank of the same active given by r_i/N.

$$AUAC = 1 - \langle x \rangle \tag{6}$$

$$AUAC = 1 - \frac{1}{nN}\sum_{i=1}^{n} r_i \tag{7}$$

$$= 1 - \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8}$$

If we use eq 7 to calculate the *AUAC* of the example from Figure 3a, we obtain 0.54 instead of 0.59. This is due to the fact that the continuous formulation is like a Riemann integral instead of the trapezoid rule, and the difference is $1/(2N)$, which goes very rapidly to zero. Equation 6 shows that the average relative rank of actives used in the literature[2] has the same meaning as the *AUAC* except that a small value of the former means good ranking performances as opposed to the *AUAC* where a large value means good ranking performances. Equation 6 also clearly shows that an active found early in the list or at the end of the list will influence the metric equally. The *AUAC* cannot discriminate any of the three cases outlined in the Introduction. We call this kind of metric an unweighted metric, and it is clearly a bad metric to discriminate among the ranking methods when early recognition is important. Finally, eq 7 is the starting equation to calculate the variance of the *AUAC* and the *ROC* metrics discussed in the "Effect of Metric Variances" section.

Area under the Receiver Operating Characteristic Curve. The *ROC* metric is widely used across many disciplines. It has its roots in signal detection analysis and was widely applied in the medical community to evaluate the discriminatory power of tests for normal or abnormal

patients or radiography images for example. The interested reader is referred to very interesting reviews.[21,23−26] The popularity of *ROC* over other metrics lies in the fact that it is nonparametric: no assumption is made about the shape of the distribution. Also, its graphical representation gives a good feeling about the performance of the ranking. The discrete formula for a set of ranked elements is given below:

$$ROC = \frac{1}{(nN)} \sum_{k=2}^{N} F_a(k)\,[F_i(k) - F_i(k-1)] \qquad (9)$$

where $F_i(k)$ is the cumulative count of inactives at rank position k (inactive empirical CDF). The continuous definition of *ROC* is given by eq 10 and can be interpreted as the probability that an active compound will be ranked earlier than an inactive one[27] within a rank-ordered list. This is highly related to the Wilcoxon rank sum test, which is also well-known by statisticians.

$$ROC = \int_0^1 F_a(x)\, f_i(x)\, dx \qquad (10)$$

Here, $f_i(x)$ is the PDF of the inactives from the ordered list. The ROC curve can be confused with the accumulation curve (or enrichment curve), while it is really a distinct mathematical entity as shown in Figure 3 where the accumulation curve (a) is compared to the ROC curve (b) for the same ranking result. The ROC curve is drawn by moving across the ranked list of compounds, and for each rank, the cumulative counts of actives and inactives are calculated. In principle, more than two class models can be examined as well. Many papers also call $F_a(x)$ the sensitivity and $F_i(x)$ is called (1 − specificity), or the true positive rate and the false positive rate, respectively. The *ROC* metric calculated with the discrete formula can formally be 1, as opposed to the *AUAC*. We will discuss further in the "Saturation Effect" section that, when the *AUAC* has a maximum value less than 1, this biases the *ROC* as well. A last point to note is that the real discrete *ROC* takes the average value of $^1/_2 + 1/[2(N - n)]$ if the actives are randomly distributed (see the Appendix).

Coming back to the three cases outlined in the Introduction, we see that the failure of *ROC* to discriminate them can be generalized. In fact, this situation can occur anytime that two ROC curves cross one another. If two ROC curves do not cross, one is said to dominate the other,[23] and it means that the dominating ranking method will always be better than the other. This should be reflected by any metric, including the one that we are proposing.

Link between *AUAC* and the *ROC* Metrics. In the Appendix, we derive a formal relationship between the *ROC* and *AUAC* metrics given by eq 11 where R_a is the ratio of actives in the list (n/N) and R_i is the ratio of inactives in the same list. The basis for the proof lies in the fact that, if we know the PDF of the actives in a ranking experiment, we then know the PDF of the inactives; hence, the two distributions are dependent.

$$ROC = \frac{AUAC}{R_i} - \frac{R_a}{2R_i} = \frac{1 - \langle x \rangle}{R_i} - \frac{R_a}{2R_i} \qquad (11)$$

If $n \ll N$, then $R_i \rightarrow 1$ and $R_a \rightarrow 0$

$$ROC \approx AUAC \qquad (12)$$

Importantly, eq 11 shows that any *AUAC* score (e.g., from the literature) can be converted into a *ROC* score, and vice versa, as long as n and N are reported. *ROC* does show an advantage over *AUAC* in that the minimum and maximum values are not dependent on the number of actives and inactives used in the evaluation. However, eq 12 shows that, as the ratio of inactives to actives becomes large, which is the desirable and general case, this advantage disappears and *ROC* becomes equal to *AUAC*. The ROC curve has the advantage of having a vertical line for an ideal curve (meaning all actives are found at the beginning), easing visual comparisons of ranking methods between evaluations. Again, this benefit quickly disappears as the ratio of inactives to actives becomes large.

In terms of statistical information, *ROC* and *AUAC* are both based on the average position of the actives: position 1 and N have the same contribution. This clearly shows that they do not discriminate the early part of the rank-ordered list from the last part and are therefore not appropriate for application to an "early recognition" problem. This is not to say that the *ROC* score is a bad metric in general: in situations where the discrimination power of a classifier (actives/inactives classes) is all that matters, we believe that the ROC curve and the *ROC* score can be useful, but in the case where the discrimination needs to occur at the very beginning of the rank-ordered list, they need to be modified.

Equations 11 and 12 allow us to re-examine several statements and claims made about the advantage of *ROC* over *AUAC*. The statement that the accumulation curves contain half of the aspect of a ranking experiment[9] should be softened since the false positive rates used in ROC curves are linearly dependent on the true positive rate (see the Appendix). Also, *ROC* was recently touted as having the advantage that it is independent of the proportion of positives to negatives (actives/decoys in our case). However, more recently, this idea was refuted,[14,15] and eq 11 gives strong evidence to support it. Rather, it is correct to say that only the end points (0 and 1) are independent of the ratio of actives. This gives rise to a second and conceptually easier way to derive eq 11 which involves a linear scaling of the *AUAC* metric to make the maximum and minimum values 0 and 1 (see the Appendix); this second route is relevant since it is the one that will be used later on when deriving *BEDROC*.

Enrichment Factor. The *EF* metric is simply the measure of how many more actives we find within a defined "early recognition" fraction of the ordered list relative to a random distribution. The discrete formula is given by eq 13 and the continuous formulas by eqs 14 and 15.

$$EF = \frac{\sum_{i=1}^{n} \delta_i}{\chi n} \quad \text{where } \delta_i = \begin{cases} 1, & r_i \le \chi N \\ 0, & r_i > \chi N \end{cases} \qquad (13)$$

$$EF = \frac{\int_0^1 f_a(x)\, w(x)\, dx}{\int_0^1 w(x)\, dx} \quad \text{where } w(x) = \begin{cases} 1, & x \le \chi \\ 0, & x > \chi \end{cases} \tag{14}$$

$$= \frac{\int_0^\chi f_a(x)\, dx}{\chi} \tag{15}$$

In theses equations, χ is the fraction of the ordered list that is considered and goes from 0 to 1. The maximum value that *EF* can take is $1/\chi$ if $\chi \geq n/N$ and N/n if $\chi < n/N$, and the minimum value is 0. In the situation of a uniform distribution of the actives in the ordered list, *EF* takes the average value of $\lfloor \chi N \rfloor / (\chi N)$. Here, the lower brackets mean "the largest integer smaller than". This metric has the advantage of answering the question: how enriched in actives will the set of 300 compounds that I select for screening be compared to the case where I would just pick the 300 compounds randomly? This is relevant only if the database is of considerable size ($N > 300$), hence, the importance of the χ parameter. A second advantage is that it does not weight equally all compounds. Coming back to our three cases, *EF* correctly ranks the hypothetical methods as opposed to *ROC*. However, a disadvantage of *EF* is that it equally weights the actives within the cutoff such that it cannot distinguish the better ranking algorithm where all the actives are ranked at the very beginning of the ordered list from a worse algorithm where all the actives are ranked just before the cutoff (e.g., 10%). A second disadvantage concerns the fact that the value obtained is highly dependent on χ, and the maximum value depends on χ, *n*, and *N*. A third disadvantage is due to the lack of discrimination after the χ cutoff. The *EF* metric has also been criticized for the variability of the score when a small amount of actives is used and likely to be close to the hard cutoff at χN. This point has not yet been proven either by simulations or by analytical work, and we will examine this in the "Effect of Metric Variance" section. Finally, the *EF* metric has been criticized for the arbitrariness of the "early recognition" parameter χ that necessitates the user to make a choice that is clearly avoided in the *ROC* metric (and *AUAC*).[2] Many choices of χ are possible, but avoiding the choice means the "early recognition" problem cannot be addressed, and the practical usefulness of a ranking method where only a very small portion of the database can be screened cannot be assessed. In the case of the *ROC* metric, this choice is avoided, which results in a metric insensitive to early performance.

Robust Initial Enhancement. *RIE*, developed by Sheridan et al.,[19] is a metric using a continuously decreasing exponential weight as a function of rank. The rationale advanced by Sheridan et al. for the proposal of this new metric is mainly that it is less susceptible than the *EF* metric to have large variations when a small number of actives are used. When applied to the three hypothetical ranking scenarios from the Introduction, *RIE* correctly identifies their performance. The original discrete formula is given by eqs 16 and 17 where r_i is the rank of the *i*th active in the ordered list and x_i is the relative rank when scaled.

$$RIE = \frac{\sum_{i=1}^{n} e^{-\alpha r_i/N}}{\left\langle \sum_{i=1}^{n} e^{-\alpha r_i/N} \right\rangle_r} \tag{16}$$

$$= \frac{\sum_{i=1}^{n} e^{-\alpha x_i}}{\left\langle \sum_{i=1}^{n} e^{-\alpha x_i} \right\rangle_r} \tag{17}$$

The denominator corresponds to the average sum of the exponential when *n* actives are uniformly distributed in the ordered list containing *N* compounds. In this paper, we use the angle brackets to denote averaging, and the subscript *r* means that it is over a uniform distribution. Originally, the *RIE* metric was calculated through a Monte Carlo simulation using 1000 trials. However, realizing that there are $N!/(N-n)!n!$ possible combinations, for example, 17×10^{12} different combinations when 10 actives and 90 inactives are used, it becomes clear that when $N = 20\,000$, the 1000 trials might not be enough to obtain an accurate value. In practice, very accurate numbers were obtained only with 100 000 samples or more. To avoid this issue, we have analytically calculated this random average and obtained the exact formula for the *RIE* metric given in eqs 18 and 19 (see the Appendix).

$$RIE = \frac{\sum_{i=1}^{n} e^{-\alpha x_i}}{\frac{n}{N}\left(\frac{1 - e^{-\alpha}}{e^{\alpha/N} - 1}\right)} \tag{18}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} e^{-\alpha x_i}}{\frac{1}{N}\left(\frac{1 - e^{-\alpha}}{e^{\alpha/N} - 1}\right)} \tag{19}$$

Originally, the motivation for obtaining eq 18 was simply to avoid the unnecessary Monte Carlo simulation. However, this formulation allows us to recast eq 18 into eq 19, wherein the relationship of the numerator and denominator with the active PDF becomes obvious. In fact, the numerator of eq 19 is simply the average of the exponential according to the ranking method active PDF. The denominator is the exponential average according to a uniform PDF. Therefore, one can rewrite eq 19 and obtain eq 20 in the continuous representation: a Boltzmann distribution. The numerator is the averaged exponential of the active position, and the denominator is the averaged exponential over a uniform distribution.

$$RIE = \frac{\int_0^1 f_a(x)\, e^{-\alpha x}\, dx}{1/\alpha(1 - e^{-\alpha})} \tag{20}$$

To verify the consistency of our reasoning, one can calculate the denominator of eq 19 when *N* becomes very

large (condition for continuity), and the result obtained is the integral of the exponential from 0 to 1 corresponding to the denominator of eq 20. The mathematical proof is straightforward, and the result is shown in eq 21.

$$\lim_{N\to\infty} \frac{e^{-\alpha/N}}{N}\left(\frac{1-e^{-\alpha}}{1-e^{-\alpha/N}}\right) = \frac{1-e^{-\alpha}}{\alpha} \tag{21}$$

The *RIE* metric has a meaning similar to *EF* since it also shows how many times better than random is the exponential average of the distribution generated by the ranking method. The meaning of $1/\alpha$ is very close to the meaning of χ in *EF* since, in an exponential distribution, the standard deviation is given by the inverse of the exponent factor and it more or less corresponds to the width of the distribution. If we compare the denominators of eqs 15 and 20, it is obvious that, if α is sufficiently large, $1/\alpha$ corresponds to χ. Therefore, $1/\alpha$ can be understood as the fraction of the list where the weight is important. However, as opposed to *EF*, the *RIE* metric has the advantage of including the contributions of all actives into the final score. Another advantage of the *RIE* metric over *EF* is that it distinguishes the situation where all the actives are ranked at the beginning of the ordered list from the situation where the actives would be ranked close to the limit at $1/\alpha$ (*RIE*) or χ (*EF*). The claimed robustness stability regarding a slight change in the active position will be discussed in the section "Effect of Metric Variances", but if α is large, one can imagine that this problem could arise with *RIE*.

A significant disadvantage of the *RIE* metric is that its minimum value (when all the actives are ranked at the tail of the list) and its maximum value (when all the actives are ranked at the beginning of the list) are dependent on *n*, *N*, and α as shown by eqs 22 and 23. This is also true in other non-weighted metrics such as the *AUAC* or the average rank.

$$RIE_{\min} = \frac{1-e^{\alpha R_a}}{R_a(1-e^{\alpha})} \approx \frac{\alpha}{e^{\alpha}-1} \quad \text{when } \alpha R_a \ll 1 \tag{22}$$

$$RIE_{\max} = \frac{1-e^{-\alpha R_a}}{R_a(1-e^{-\alpha})} \approx \frac{\alpha}{1-e^{-\alpha}} \quad \text{when } \alpha R_a \ll 1 \tag{23}$$

In other words, the scale significantly changes depending on α and R_a, which makes the comparisons between two *RIE* values dangerous if the condition $\alpha R_a \ll 1$ is not met. This condition is particularly demanding on the number of decoys required when α is high. The RIE_{max} from eq 23 is plotted in Figure 4 as a function of R_a for $\alpha = 100$ as used in previous work.[7,19] The curve indicates the maximum value of the *RIE* metric if a perfect ordering is obtained from a ranking method. From this figure, it is clear that a *RIE* of 20 is the best performance that can be achieved when $R_a = 0.05$ but is not as optimal when $R_a = 0.01$. This clearly shows that using a weighted metric must be done with care. We will provide more insight on this issue when the "saturation effect" is discussed.

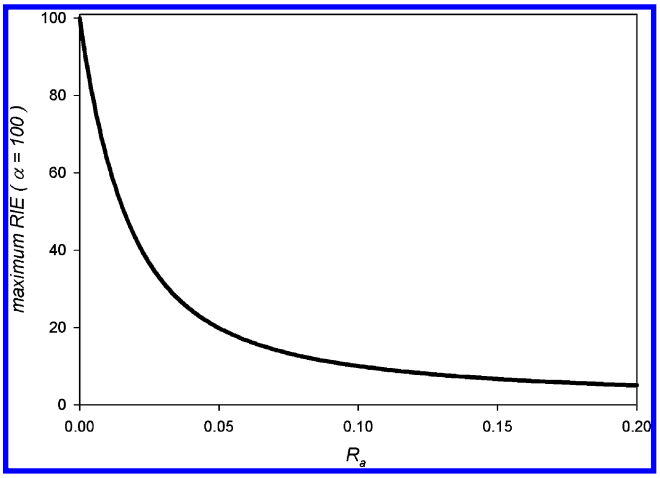

Figure 4. *RIE* maximum as a function of the ratio of actives R_a with $\alpha = 100$. This figure shows that the maximum RIE has a very high dependency over R_a.

The *RIE* uses an exponential weight formula, but eqs 19 and 20 can be generalized for any kind of smooth weighting function as shown by eqs 24 and 25 within the continuous and the discrete formulation, respectively. Other possibilities, such as an inverse logistic "S" shape function or a Gaussian function, would have the drawback that the logistic function would tend to behave like *EF* in not discriminating the compounds close the threshold and would be more difficult to track mathematically, and the Gaussian function goes to zero too rapidly.

$$wRIE = \frac{\int_0^1 f_a(x)\, w(x)\, dx}{\int_0^1 w(x)\, dx} \tag{24}$$

$$wRIE = \frac{\frac{1}{n}\sum_{i=1}^{n} w(x_i)}{\frac{1}{N}\sum_{k=1}^{N} w(k/N)} \tag{25}$$

Interestingly, we were able to relate the *RIE* metric to *ROC* (proof in the Appendix) as shown by eqs 26 and 27 below:

$$ROC = \frac{1}{R_i}\left(\frac{\partial RIE}{\partial \alpha}\right)_{\alpha=0} + \frac{1}{2} \tag{26}$$

$$\approx \frac{RIE(\alpha)-1}{\alpha R_i} + \frac{1}{2};\ \alpha \text{ is small} \tag{27}$$

This tells us that the *ROC* metric is linearly related to the *RIE* variation rate when α is small (uniform weight function is applied). When the finite difference derivative is used, the *ROC* metric is linearly dependent on the *RIE* metric only if the exponential prefactor used in the calculation of *RIE* is very small.

The Boltzmann-Enhanced Discrimination of ROC (*BEDROC*) Metric. The *RIE* metric has the desirable property that it correctly assesses "early recognition" in a ranking method. We previously showed that *ROC* can simply be obtained by forcing *AUAC* to be bounded by 0 and 1 by

applying a linear transformation. So, now we want to define a weighted CDF metric as below:

$$wAUAC = \frac{\int_0^1 F_a(x)\, w(x)\, dx}{\int_0^1 w(x)\, dx} \tag{28}$$

The weighting function could in principle weight appropriately any region of the relative rank x, but in the present context, we upweight the early region of the ordered list. The weighting function should have a finite area between 0 and 1, and then it would behave like a PDF leading to the equations below (note the similarity with eq 10).

$$\tilde{f}(x) = \frac{w(x)}{\int_0^1 w(y)\, dy} \tag{29}$$

$$wAUAC = \int_0^1 F_a(x)\, \tilde{f}(x)\, dx \tag{30}$$

Here, *wAUAC* can be interpreted as the probability that an active in the ordered list be ranked before a compound ranked by an algorithm with an underlying PDF given by $\tilde{f}(x)$. Equation 30 is not bounded by 1 because the CDF, $F_a(x)$, must have an initial slope in practice. Therefore, it is very tempting to repeat the same linear scaling that transformed *AUAC* into a *ROC* score as below (see the Appendix). We name this general metric *swAUAC* for scaled weighted *AUAC*.

$$swAUAC = \frac{wAUAC - wAUAC_{min}}{wAUAC_{max} - wAUAC_{min}} \tag{31}$$

$$= \frac{wAUAC}{wAUAC_{max} - wAUAC_{min}} - \frac{wAUAC_{min}}{wAUAC_{max} - wAUAC_{min}} \tag{32}$$

It is clear that any *swAUAC* metric will take a value formally between 0, when all the actives are ranked at the end of the ordered list, and 1, when all the actives are ranked at the beginning. If one chooses $w(x) = \exp(-\alpha x)$, the corresponding *wAUAC* is a function of *RIE* (see the Appendix):

$$wAUAC = \frac{RIE}{\alpha} + \frac{1}{1 - e^{\alpha}} \tag{33}$$

$$\approx \frac{RIE}{\alpha};\ \text{if } e^{\alpha} \gg 1 \tag{34}$$

We believe that the decreasing exponential has great advantages over other weighting functions: (1) The extent of the weight importance is controllable via the single "early recognition" parameter α. (2) When interpreted as a PDF against which a ranking result is compared, the maximum weight is 1.0 and can decrease at a tunable rate. (3) It is mathematically well-behaved and easily tractable. This well-behaved weighting function leads to

$$BEDROC = \frac{wAUAC - wAUAC_{min}}{wAUAC_{max} - wAUAC_{min}} = \frac{RIE - RIE_{min}}{RIE_{max} - RIE_{min}} \tag{35}$$

$$= \frac{\sum_{i=1}^{n} e^{-\alpha r_i/N}}{\frac{n}{N}\left(\frac{1 - e^{-\alpha}}{e^{\alpha/N} - 1}\right)} \frac{R_a\, e^{\alpha R_a}(e^{\alpha} - 1)}{(e^{\alpha} - e^{\alpha R_a})(e^{\alpha R_a} - 1)} + \frac{1}{1 - e^{\alpha(1-R_a)}}$$

$$= \frac{\sum_{i=1}^{n} e^{-\alpha r_i/N}}{\frac{n}{N}\left(\frac{1 - e^{-\alpha}}{e^{\alpha/N} - 1}\right)} \times \frac{R_a \sinh(\alpha/2)}{\cosh(\alpha/2) - \cosh(\alpha/2 - \alpha R_a)} + \frac{1}{1 - e^{\alpha(1-R_a)}} \tag{36}$$

$$\approx \frac{RIE}{\alpha} + \frac{1}{1 - e^{\alpha}},\ \text{if } \alpha R_a \ll 1 \text{ and } \alpha \neq 0 \tag{37}$$

$$= wAUAC$$

The *BEDROC* metric is bounded by 0 and 1. The *RIE* metric in the above conditions is defined in eq 18 and is kept in the definition to ease the derivation of the *BEDROC* metric and to avoid calculating a numerical integral, which is a little less convenient than summing exponentials. Under the conditions $\alpha R_a \ll 1$ and $\alpha \neq 0$, the *BEDROC* metric is the probability that an active ranked by the evaluated method will be found before a compound that would come from a hypothetical exponential PDF with parameter α (see eq 1), or $\tilde{f}(x)$ in the more general case. In this case, the point of comparison is no longer a uniform PDF (like *ROC* when R_a is small) but a PDF with a higher weight for small values of x. In the case of the *ROC* metric, a value of $\sim$1/2 is obtained when the actives are uniformly distributed; in parallel, *BEDROC* takes a value of 1/2 if the observed empirical CDF has the shape of the CDF produced by $\tilde{f}(x)$, an exponential of parameter α in the case of *BEDROC* (cf. eqs 1 and 2 in the well-behaved case where $\alpha R_a \ll 1$). Thus, the comparison PDF is changed in the case of *BEDROC* in order to obtain a metric that identifies the usefulness of a ranking method, in line with the concept of "early recognition". A consequence is that the intuition the user develops for the area under the ROC curve is fully applicable for *BEDROC* except that the *BEDROC* metric is adapted to early recognition. In fact, *BEDROC* should be understood as a "virtual screening usefulness scale" as opposed to an "improvement over random scale" (*ROC*). Even the probability meaning is similar, the only difference being that *ROC* relates to a uniform distribution and *BEDROC* to an exponential distribution. The equality between *wAUAC* and *BEDROC* is subject to the condition $\alpha R_a \ll 1$ and is more restrictive than $R_a \ll 1$, which comes from the equivalence between *AUAC*

and *ROC*. This is simply saying that, in focusing on the beginning of the ranking axis, there are fewer actives that count in the metric evaluation. The effect of deviating from $\alpha R_a \ll 1$ can be analyzed by comparing the *RIE* multiplicative factor in eq 36 to $1/\alpha$. In most cases, the second term of eq 36 can be neglected since it goes to zero very rapidly as long as R_a is reasonably small and α is 5 or higher. Another point to note, if the actives are uniformly distributed, then *BEDROC* becomes

$$BEDROC_r = \frac{e^{\alpha R_a} - R_i}{e^{\alpha R_a} - 1} - \frac{R_i}{1 - e^{-\alpha R_i}} \approx \frac{1}{\alpha} + \frac{1}{1 - e^{\alpha}} \text{ if } \alpha R_a \ll 1 \quad (38)$$

To conclude, we have generalized the *ROC* metric to incorporate a weighting function that adapts it for use in "early recognition" problems. We propose to use an exponential weighting function incorporated into eq 36 which has a probabilistic meaning when $\alpha R_a \ll 1$. Now, when the three cases outlined in the introduction are reconsidered, the *BEDROC* metric succeeds in discriminating them in the correct order. Finally, it is important to realize that comparing *BEDROC* scores obtained with different α values is not correct. That would be like comparing EFs where the thresholds χ are different. The critical task of choosing the right α is equivalent to asking: what is the baseline enrichment from which you would consider a ranking method useful in the context of a VS?

4. EFFECT OF METRIC VARIANCES

Comparing the Effects of Metric Variance. In using any of these metrics to compare ranking methods, we must be cognizant of the statistical error on the values of the metrics obtained. However, very few researchers in the computational chemistry community report an error on *ROC* values in spite of the literature precedence in other fields.[24,27] We show in this section that the statistical source of error should be characterized not only when reporting the final results but also when planning an evaluation in order to reduce this potentially dominant source of error. For instance, it would be very important to know if the differences found between ranking methods in published performance evaluations[1−7,20] are actually statistically significant.

To help understand why a ranking evaluation has a statistical error, one needs to realize that any given accumulation curve can be considered a finite and specific sampling of all possible actives and decoys. Therefore, accumulation curves are not perfect static representations but can change with a slightly different set of actives and decoys. Intuitively, accumulation curves are better represented with more data points (actives) as shown in Figure 2. The difference between a perfect representation and the actual evaluation case is characterized by a statistical variance that we want to approximate.

We can start by answering the question: what parameters are important among the performance of the ranking method (λ), the number of actives (n), the total number of compounds (N), and the "early recognition" parameter (α)? Unfortunately, we were unable to answer this question analytically for the general case, so instead we approached this question by simulating the many possibilities, according to the protocol outlined in the "Methods" section. In the results shown in Figure 5, the standard deviations in the metric (STDs) were calculated by building 150 000 samples and two types of ranking methods: poor ($\lambda = 1$) and good ($\lambda = 10$). Figure 5a shows that keeping a fixed number of 100 actives, when the number of decoys or N increases, causes the STD to converge to a constant value. This is also observed with the other metrics (results not shown). In the case of *ROC*, a smaller number of compounds (N) increases the STD as opposed to *AUAC*, which has a lower STD. However, smaller values of N are not necessarily better, because this lower STD is due to the bias induced by a larger ratio of actives. From the same graph, we can observe that the poor accumulation curves ($\lambda = 1$) yield a higher variance, which is simply reflecting the fact that the actives are more spread throughout the ranks. This also sets the requirement for a high number of decoys to be used for a controlled statistical behavior. For the next parameters, the total number of compounds N was set to a large number in order to decouple the effect of N. The STD of the *ROC*, *AUAC*, and *BEDROC* metrics were calculated when the number of actives was varied, keeping $N = 25\,000$. As shown in Figure 5b, the STD goes down as the number of actives increases, which is a healthy and expected behavior. The STD of *ROC* and *AUAC* are equal, which is also expected because of the small $R_a = 0.004$. The same dependency of STD over n is observed in Figure 5c for the *EF* and the *RIE* metrics where the *RIE* STD is always slightly smaller than the *EF* STD as predicted, although this is true for all n, not just when n is small. Finally, the effect of the "early recognition" parameter α on the STD for *EF*, *RIE*, and *BEDROC* (Figure 5d) is demonstrated. In the case of *EF* and *RIE*, the STD increases with α, the *EF* STD being consistently higher.

How can these results be applied to real evaluations? The first and easiest situation is when a uniform PDF is obtained (actives uniformly distributed across the ranks). This is where the relative error [STD/mean or sqrt(Var)/mean] reaches a maximum, this could be our needed upper bound estimation of the statistical error. In this special case, analytical variances are provided in Table 1 for the different metrics examined in this work; the mathematical proofs are given in the Appendix. It is important to realize that the STDs calculated from the variances of Table 1 are exact only when the actives are uniformly distributed among the ranked compounds. We have found in the literature an upper bound of the *ROC* variance which reaches its maximum with the uniform PDF;[27] this corroborates our formulas. Furthermore, in the statistics literature, the *ROC* metric variance estimates have been derived for a variety of situations.[24,27] In most of these derivations, probabilities based on the ranking results are required; thus, it is especially useful to calculate error bars on the obtained results but useless when planning the performance evaluation. For *ROC*, we strongly encourage the calculation of statistical error on the basis of the formulas previously reported[24,27] as a standard practice. From Table 1, a few more observations are worth noting. First, the *ROC* average is not exactly 1/2, and this can be understood simply

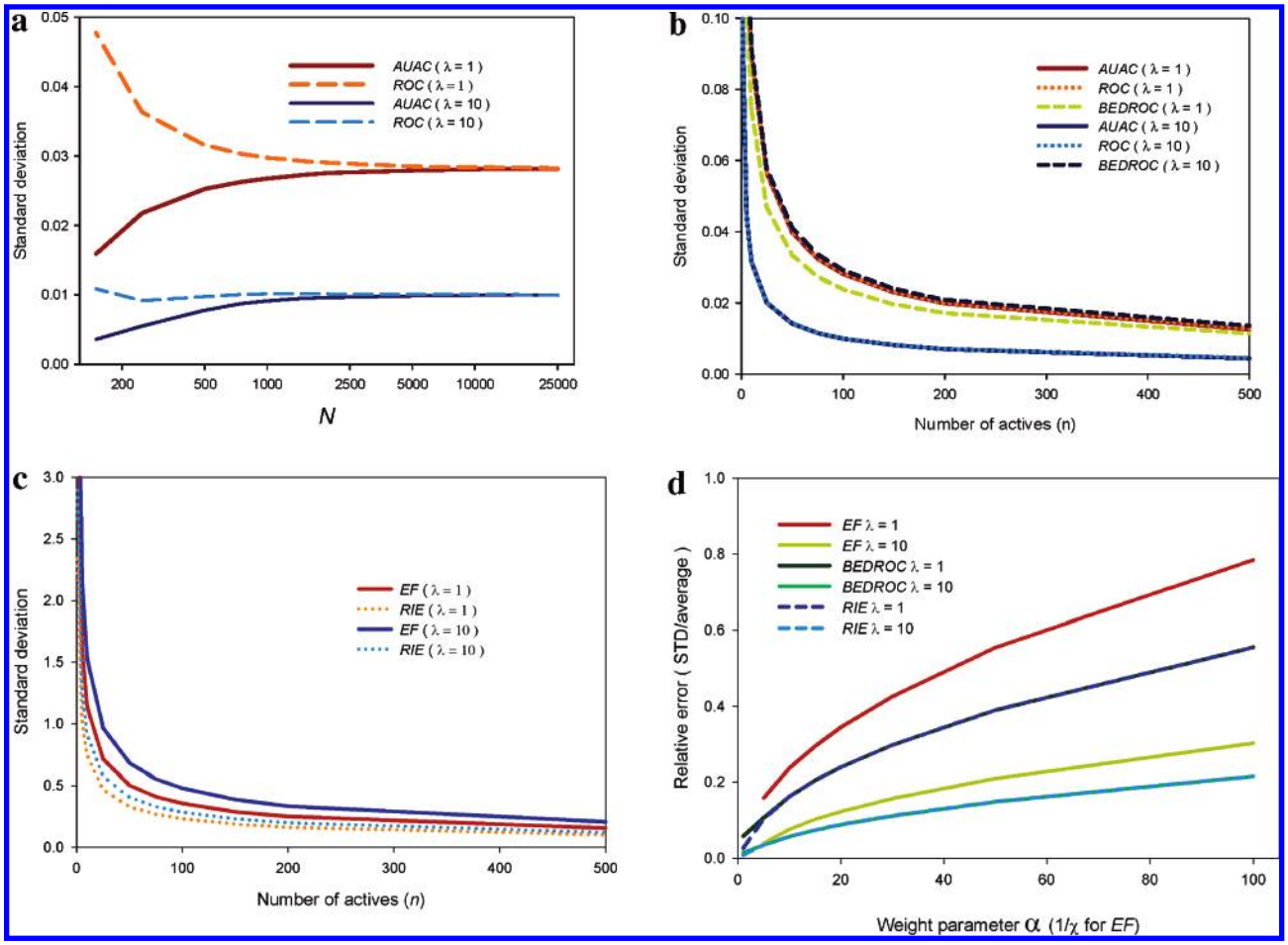

Figure 5. Calculated STDs for *ROC*, *EF*, *RIE*, and *BEDROC* metrics as a function of *N* using a semilog scale (a), *n* (b,c), and the exponential factor α (d). The standard deviation converges as *N* gets large and decreases with the number of actives. The relative error increases with α for *EF*, *RIE*, and *BEDROC*. In d, the *RIE* and *BEDROC* curves are almost superimposed. These results are obtained with simulations with λ set to 1 and 10; 150 000 samples are used in every case. When not varied, *n* is set to 100, *N* is set to 25 000, and α is set to 10.

Table 1. Analytical Variance Formula for a Uniform PDF for Each Metric Examined in This Work[a]

metric	analytical variance for uniform PDF	average for a uniform PDF	
AUAC	$\frac{(N-n)(N+1)}{12nN^2}$	$\frac{N+1}{2N}$	eq. 39
ROC	$\frac{N+1}{12n(N-n)}$	$\frac{1}{2}+\frac{1}{2(N-n)}$	eq. 40
EF	$\frac{W}{nN\chi^2}\left[1+\frac{n-1}{N-1}(W-1)\right]-\frac{W^2}{\chi^2N^2}$	$\frac{W}{\chi N}$	eq. 41
RIE	$\frac{\left[\frac{1-e^{-2\alpha}}{e^{2\alpha/N}-1}\right]+\frac{2(n-1)}{(N-1)}\frac{e^{-2\alpha}(e^{\alpha/N}-e^{\alpha})(1-e^{\alpha})}{(e^{\alpha/N}-1)^2(1+e^{\alpha/N})}}{\frac{n}{N}\left(\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}\right)^2}-1$	1	eq. 42
wAUAC	$\frac{\mathrm{Var}[RIE]_r}{\alpha^2}$	$\frac{1}{\alpha}+\frac{1}{1-e^{\alpha}}$	eq. 43
BEDROC	$\frac{\mathrm{Var}[RIE]_r \cdot R_a^2 \sinh^2(\alpha/2)}{[\cosh(\alpha/2)-\cosh(\alpha/2-\alpha R_a)]^2}$	$\frac{e^{\alpha R_a}-R_i}{e^{\alpha R_a}-1}-\frac{R_i}{1-e^{-\alpha R_i}}$	eq. 44

[a] $W=\lfloor \chi N \rfloor$, $R_i=(N-n)/N$, $R_a=n/N$, $\alpha > 0$.

by the fact that the actual random line is bumpy, which is equivalent to adding half an area unit for each active: $[1/n \times 1/(N-n) \times 1/2]n = 1/2/(N-n)$. Second, the *BEDROC* metric behaves like *ROC* when α goes to zero (no weighting of the rank), and the *BEDROC* mean of a uniform PDF goes to 0.5 in this situation. When $\alpha \gg 1$, the *BEDROC* mean goes to R_a and to the *wAUAC* mean when $\alpha R_a \ll 1$ (eq 38).

Table 1 could also be used to assess the relative error on a metric. For instance, if we want the relative error of *BEDROC* to be ≤5%, we could then find the number of

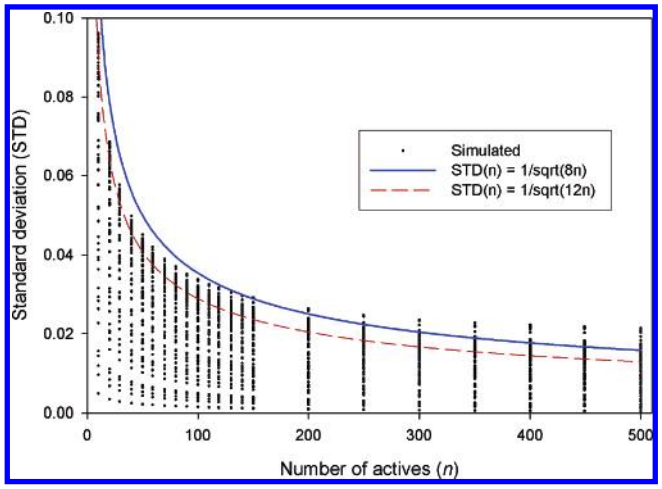

Figure 6. Standard deviation of *BEDROC* obtained by simulation for the combination of α and λ set to 1, 5, 10, 15, 20, 30, 40, 50, and 100 where 150 000 samples were generated; a total of 25 000 compounds (N) are used. Two enveloping lines are drawn, one corresponding to the STD of a uniform distribution (dashed) and a second as an empirical fit to the maximum standard deviation observed (solid).

required pairs of actives and inactives to meet this criteria. However, this turns out to lead to too many actives for the general case, and especially in the cases of *RIE*, *EF*, and *BEDROC* metrics, it is impractical.

To be more practical, we have performed another set of simulations in order to get a better upper bound for the STD but, this time, specifically designed for the *BEDROC* metric. With the simulation method described in the Methods section, we have calculated the standard deviation coming from the ranking of n actives when a total of 25 000 compounds is used for many values of α and λ (1, 5, 10, 15, 20, 30, 40, 50, and 100). Although the exponential PDF used to generate samples is only mimicking true distributions, we believe that it gives a good approximation that can be generalized since it is sampling a diverse range of accumulation curves. Figure 6 shows the results for the STDs calculated as a function of the number of actives, and each STD was obtained with 150 000 samples. Although we have not found an analytical formula to relate the STD to n, α, and λ, we can calculate the equation of the envelope that forms the upper bound of all STDs. The maximum STDs were observed in the simulations with $\alpha \approx \lambda$ independently of n; when α is different than λ, the STD is lower. An obvious example of this situation is with a uniform PDF and $\alpha = 0$. This can be analytically calculated from the *BEDROC* variance in Table 1 by taking the limit when N is large and α goes to zero, and the result is 1/sqrt(12n), the standard deviation for a general uniform distribution. This function is drawn as a red dashed line Figure 6, and it is slightly lower than the empirically fit maximum STD given by 1/sqrt(8n) (solid blue line).

As a final point in this section, we would like to mention that one of the best ways to assess the error on a given ranking is to proceed with a bootstrapping statistical analysis. We have been able to approximate STDs of true ranking experiments by randomly removing a certain percentage of entries in the rank-ordered list and calculating the obtained metric for many such samples. We found that, with an exponentially weighted metric, the STD could be reasonably well approximated by removing about 20% of the compounds and performing the operation on 10 000 different samples. The obtained variance can be used as an error estimate on the metric.

5. SATURATION EFFECTS

An advantage of using the theoretical PDF defined in eq 1 and the simulation is that we can assess what the effect of increasing R_a is on the value of a metric for the same PDF, whether it is a PDF corresponding to high (e.g., $\lambda = 20$) or low (e.g., $\lambda = 1$) enrichment. All metrics rely on the correct assessment of the PDF that is mapped into a simple scalar according to the formulas outlined in previous sections. For various values of λ of the exponential distribution defined in eq 1, for seven values of R_a, and in the case of weighted metrics for α set to 20, we calculate analytically the expected values of *AUAC* (eq 4), *EF* (eq 15), *RIE* (eq 20), and *BEDROC* (eq 30). The *ROC* metric is a special case since there is no analytical expected value, although we know that it should be equal to *AUAC* when R_a is small. Here, we want to highlight an important curse accompanying the "early recognition" problem, namely, the "saturation effect". In fact, once actives "saturate" the early part of the list, the performance metric cannot get any higher. Obviously, when one defines "early" as a smaller effective number of positions (i.e., when α is high), the saturation issue becomes more acute.

In Figure 7, the results are shown not only for the analytical expected exact value but also for numerical simulations that mimic practical situations. Each point was calculated using $N = 25\,000$ (to make the ranking scale almost continuous), and 150 000 samples were generated using the inverse transformation method described earlier. The exact values are calculated using the continuous formulas and the analytical formula of the sampled PDF (eq 1). The standard deviations obtained during these simulations are also shown as error bars on the data points. These graphs allow us to observe the deviation of the metric values from ideal when R_a is increased. In the case of *ROC*, the deviation is larger in the medium regime of λ (medium enrichment) than that with small or large values (bad and good ranking) and spans 0.07 at its widest (Figure 7a). On the other hand, the *AUAC* metric reaches the maximum deviation at large values of λ, because of the initial slope in the accumulation curve (Figure 7b); the maximum span seen at $R_a = 0.3$ is 0.11, similarly to the *ROC* deviation. This reinforces, as a second argument, our assertion that the *ROC* metric is dependent on the ratio of actives/inactives as opposed to what others have claimed;[9] the first argument being eq 11. In fact, the ROC curve is based on true positive and false positive rates which are affected by the prior probability (different ratios of actives). The only difference from the *AUAC* metric is the region where the maximum deviation is reached. In the case of the *ROC*, the maximum deviation is reached directly in the region of interest (where most of the performances are observed), whereas *AUAC* reaches its maximum in the region where an excellent ranking is obtained; this is the manifestation of the "saturation effect". The *ROC* analysis is not a panacea to solve the "saturation effect"; rather, a careful planning of the evaluation is required to ensure that either this effect is constant or it is absent because of a

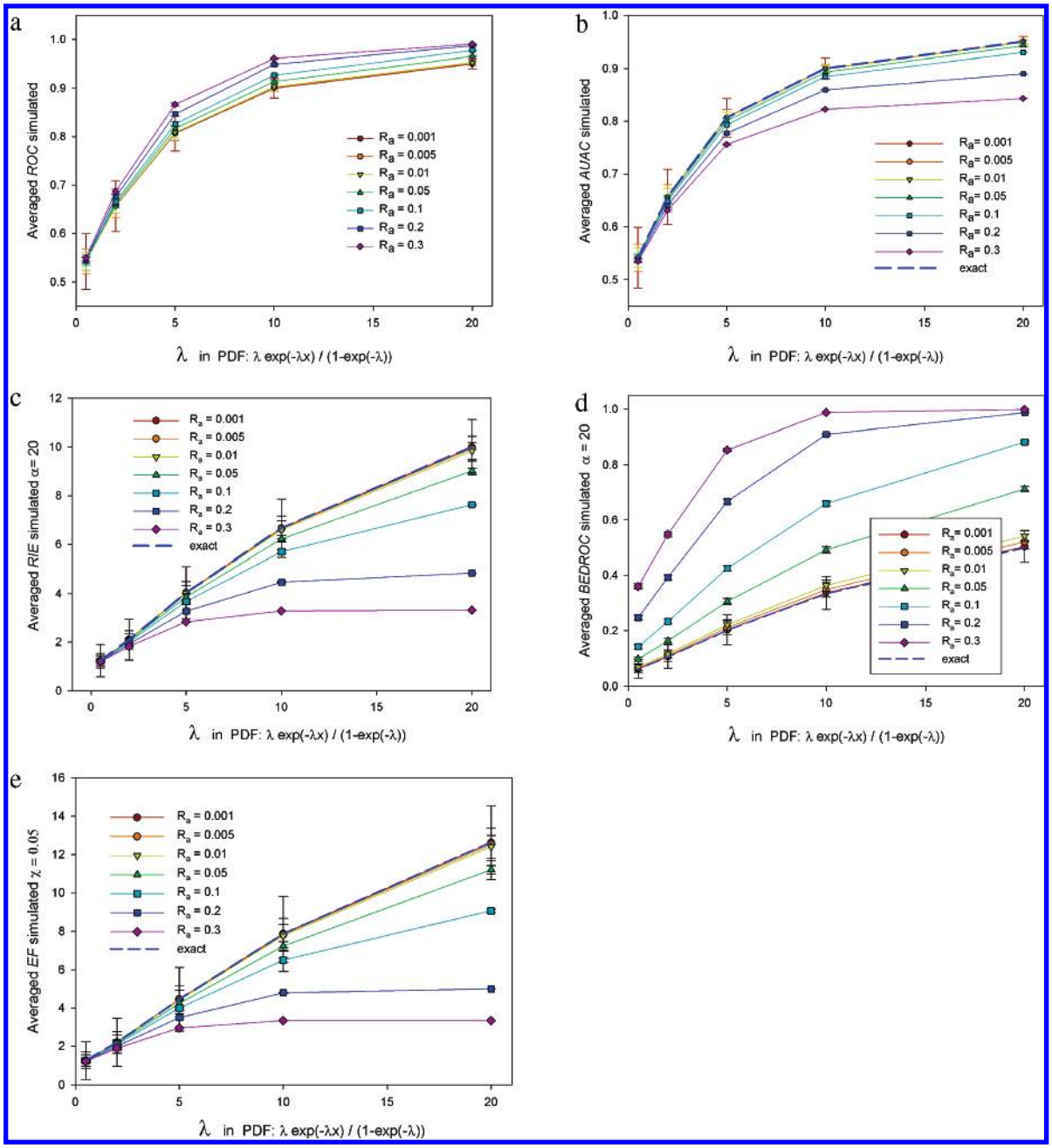

Figure 7. Graphs showing averaged scores obtained by (a) *ROC*, (b) *AUAC*, (c) *RIE* ($\alpha = 20$), (d) *BEDROC* ($\alpha = 20$), and (e) *EF* ($\chi = 0.05$) metrics for different values of λ according to eq 2. The ratio of actives (R_a) is varied from 0.001 to 0.3 to show how the metrics change their value as a function of the ratio of actives for a constant true performance. The curves are the results of averaging 150 000 samples, simulated using the inverse transformation method with 25 000 ranked compounds.

reasonably small ratio of actives. The comparison within the groups *ROC*−*AUAC* and *BEDROC*−*RIE*−*EF* is clear: only the effects are amplified in the case of the exponentially weighted metrics. Indeed, the *RIE* metric deviates more when λ increases (the PDF is steeper) unless $\alpha R_a \ll 1$ is respected. The deviation can be as large as 75% of the scale for large values of λ. In the case of *BEDROC*, the deviation can be important too, converging nevertheless to a common value of 1 for larger λ values. The problem with these deviations is that, when studies with different R_a values are performed, one cannot compare or average the obtained scores because the deviation is not directly related to the performance of the ranking method and can significantly shift the score. This behavior is an unfortunate but unavoidable consequence of the "early recognition" aspect of the problem. On the basis of these results, we believe that ranking algorithm evaluations should operate in the regime where $R_a \ll 1$ (*ROC* and *AUAC*) and $\alpha R_a \ll 1$ for *BEDROC*. This allows healthy comparisons between results from different studies. In the next section, we provide guidelines on how to fulfill this condition in a practical way.

6. PLANNING AN EVALUATION WHERE EARLY RECOGNITION IS IMPORTANT

The Choice of the Exponent Prefactor α. In metrics such as the weighted *RIE*, *wAUAC*, and *BEDROC*, and in the *EF*, the user-defined parameter α (or χ for *EF*, but we will refer only to α for the subsequent discussion) needs to be chosen. This parameter embodies the "early recognition" element, adopting higher values to move the region of importance

further to the front of the list. A challenge for evaluations is to choose a value which could be a useful middle ground for comparisons between different studies, and we want here to propose a sensible way to do it. In *ROC*, such a choice is not necessary because the comparator is a random distribution (when R_a is small), which is equivalent to $\alpha \rightarrow 0$. In what follows, we will focus on the *wAUAC* metric that is equal to the *BEDROC* metric when the condition $\alpha R_a \ll 1$ is met; we believe *BEDROC* is the relevant metric to study. We know that, if α is high, the early part of the accumulation curve will count relatively more, eventually to the point that the rest of the curve will not matter. Thus, choosing the appropriate α is important to the extent that it determines what part of the curve is "early." We outline below a scheme to find an α value based on the contribution of the weight. First, we rewrite eq 28 with $w(x) = \exp(-\alpha x)$ in the explicit form shown in eq 45 using the approximation that the ranking is continuous (N is sufficiently high for the integral to be accurate).

$$wAUAC = \frac{\int_0^1 F_a(x)\, e^{-\alpha x}\, dx}{(1 - e^{-\alpha})/\alpha} \tag{45}$$

We then calculate the maximum contribution to the *wAUAC* that z% of the rank can generate by setting $F_a(x) = 1$ (perfect method). The *wAUAC* obtained is noted as θ, and the value of θ is calculated with eq 46. This equation is easily integrated to lead to eq 47.

$$\theta = \frac{\int_0^z e^{-\alpha x}\, dx}{(1 - e^{-\alpha})/\alpha} = \frac{1 - e^{-\alpha z}}{1 - e^{-\alpha}} \tag{46}$$

$$\Rightarrow 0 = \theta(1 - e^{-\alpha}) + 1 - e^{-\alpha z} - 1 \tag{47}$$

We can read eq 46 as the following question: what is the value of α that will contribute to θ% of the total score at z% of the rank? In other words, if one has a sample collection of 1 million compounds and can test only 3000 compounds in an assay, then it is quite important that a reasonable amount of actives be found in the top scoring 0.3% of the collection. For example, in this case, we could ask that 50% of the overall score come from 1% of the relative rank. Hence, eq 47 would become "$0 = 0.5[1 - \exp(-\alpha)] + \exp(-\alpha 0.01) - 1$", and solving for α leads to $\alpha = 69.3$. Alternatively, requiring that 80% of the score come from the top 5% of the beginning of the ordered compounds results in $\alpha = 32.2$. Many cases are summarized in Figure 8 where θ is plotted versus z for a few values of α. This graph can be used to approximate the α needed. For instance, if we want the first 10% of the rank ($z = 0.1$) to contribute for 80% of the score ($\theta = 0.8$), the figure gives $\alpha \approx 15$ since the desired coordinate (0.1, 0.8) is between the curve where $\alpha = 20$ and the curve obtained from $\alpha = 10$. Of course, a numerical solution can easily be obtained. In terms of "early recognition" discrimination, all values of $\alpha > 0$ would rank correctly the three cases given in the Introduction. The authors would also like to stress that, whatever an investigator picks as a specific α value, or otherwise specifies as a metric

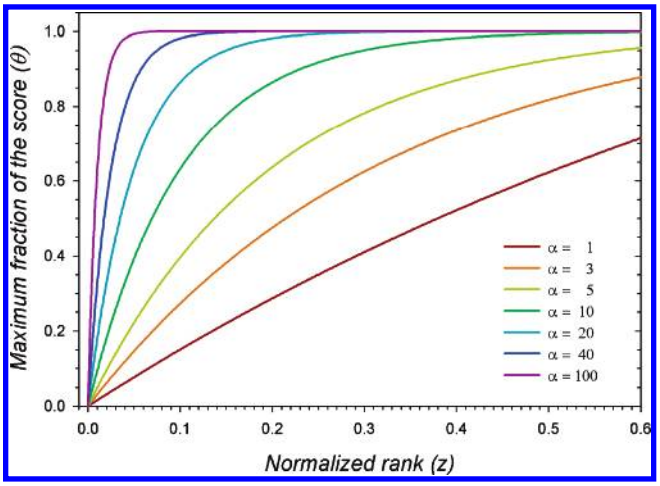

Figure 8. Maximum fraction of the *BEDROC* score (θ) as a function of the contributing normalized rank (z) for various values of α in the context of eq 47.

Table 2. Correspondence between *EF* and *BEDROC* Parameters Calculated by Solving for α in eq 47 when $\theta = 80\%$ and $z = \chi$

EF χ	*BEDROC* α
1.0%	160.9
1.6%	100.0
3.0%	53.6
3.2%	50.0
5.0%	32.2
8.0%	20.0
10.0%	16.1
16.1%	10.0
20.0%	8.0

in a publication, it is important that the actual rank data be added as Supporting Information so others may apply their own metrics for comparison purposes. For comparison with precedent literature, we present in Table 2 the corresponding values between the *EF* thresholds χ and the *BEDROC* parameter α calculated from eq 47 where 80% of the score (Θ) comes from the fraction χ of the ordered list. Of course, the advantages of *BEDROC* over the *EF* metric still hold, and they mainly concern the performance before and after the threshold cutoff χ in the case of *BEDROC*. It is to be noted that α should not be chosen in such a way that it represents the best performance expected by a ranking method, but rather it should be considered as a useful standard to discriminate better or worse performance in a real problem to which the ranking method will be applied.

In the special case when $\theta = 80\%$ and $z = 8\%$, which leads to $\alpha = 20.0$—a reasonable choice for a VS evaluation—it is interesting to compare *BEDROC* values with *ROC* values for a few ranking scenarios obtained with explicit ranks when $n = 50$ and $N = 25\,000$, as illustrated in Figure 9. Two comparator curves are drawn in blue: one is the CDF (accumulation curve) resulting from the random distribution of actives (dotted blue line), and the other is the CDF enriched in actives corresponding to an exponential distribution with the λ parameter set to 20 (solid blue line). In the case of *ROC*, a value of approximately 0.5 is normally obtained for the random distribution of actives, and in the case of *BEDROC*, the enriched CDF following the exponential CDF is close to 0.5. In all the empirical CDFs but one, *ROC* and the *BEDROC* correctly identify the dominance of the ranking methods. The *BEDROC* score shows a higher

sensitivity than the *ROC* one in the good ranking end. The *ROC* metric fails in comparing the magenta CDF that is very good at retrieving the first 60% of the actives. In terms of "early recognition", this CDF should be better than most of the other CDFs, but *ROC* gives a score of only 0.82, smaller than most of the other CDFs. In contrast, the *BEDROC* metric weights more the early part of the curve and therefore identifies it as being one of the best CDFs.

Best Selection of the Number of Actives and Decoys in Weighted Metrics. This subsection is about practical heuristics to control the two sources of error described in the previous sections, the metric variance and the saturation effect. At this stage, we suppose that α is chosen and we examine the optimal choice of the number of active and decoy compounds to use in an evaluation. Two conditions need to be taken into account: we want to minimize intrinsic variance, which requires a large number of actives, but we also want to avoid saturation, which requires a large number of decoys. Obviously, having a large number of both would be optimal, but we want to minimize both for practical reasons: for a given target, only a limited number of actives may be available, and limiting the total number of compounds is necessary to make an evaluation tractable in terms of CPU costs. Although the analysis can be applied on most of the metrics discussed so far, we will focus our attention on *BEDROC*, which we consider to be superior to *ROC*, *EF*, and *RIE*.

First of all, the maximum statistical STD is given by eq 48, and this equation should be used to decide the number of actives to use. For example, if $n = 10$, 50, 100, or 200, the corresponding maximum STDs for *BEDROC* are 0.11, 0.05, 0.035, and 0.025, respectively, on a scale that goes from 0 to 1. The best benefit from variance reduction occurs between 50 and 100 actives as found earlier.

$$\mathrm{STD}_{\max} = \frac{1}{\sqrt{8n}} \tag{48}$$

The second source of error is related to the saturation effect that can be evaluated, in the case of the *BEDROC* metric, by comparing eqs 36 and 37 term by term. As discussed in the "Saturation Effects" section, fulfilling the requirement that $\alpha R_a \ll 1$ allows reliable cross-evaluation comparisons. The difference in the second terms of eqs 36 and 37 is small for $\alpha \geq 5$; for example, if $\alpha = 5$ and $R_a = 20\%$, the difference is 0.01. The first term however commands more attention. We can calculate the relative deviation Δ as follows

$$\Delta = \frac{\alpha R_a \sinh(\alpha/2)}{\cosh(\alpha/2) - \cosh(\alpha/2 - \alpha R_a)} - 1 \tag{49}$$

which comes from the difference in the first terms of eqs 36 and 37 divided by $1/\alpha$, the expected value. The calculated Δ corresponds to the relative difference between *wAUAC* and *BEDROC*, which vanishes when the saturation effects are negligible. Therefore, Δ is a percent deviation that has a maximum value when the *BEDROC* value equals 1. According to eq 35, Δ also corresponds to a deviation of the $RIE_{\max} - RIE_{\min}$ range from the expected saturationless value of α. The values of n and α being known, we want to find the minimum number of compounds ($N_{\min}$) necessary to make sure that the relative deviation from *wAUAC* is smaller than $\Delta_{\max}$. Thus, we need to minimize the function E

$$E(N;n,\alpha) = [\Delta_{\max} - \Delta(n, N, \alpha)]^2 \tag{50}$$

which is easily accomplished with the Nelder−Mead simplex algorithm by setting $N = n + 10$ as the starting point. The results for $\Delta_{\max} = 5\%$ and 1% are reported in Table 3 for a few values of n and α. These results are also applicable to the *RIE* based on the denominator of eq 35.

Repeating the optimization with a few $\Delta_{\max}$ values in the range 1−20% unveils the following empirical relationships:

$$N_{\min} = \frac{\alpha n}{2\Delta_{\max}} \tag{51}$$

$$\Rightarrow R_a = \frac{2\Delta_{\max}}{\alpha} \tag{52}$$

Therefore, when planning an evaluation where the early recognition is important, the source of statistical variability can be controlled with eq 48, and the departure from a correct asymptotic behavior that allows interstudy comparisons can be ensured up to a certain percent by using $N_{\min}$, given by eq 51. The choice of α needs to be based on the purpose of the evaluation (eq 47), and it affects the total number of compounds (N) needed via eq 51.

When many actives are available, the throughput of the VS method might not allow the screening of large amounts of compounds as suggested by Table 3 to avoid saturation effects. Of course, saturation will only occur with a sufficiently good ranking method. This difficulty can be avoided by scoring only a fraction of the actives at a time in a bootstrap average. Bootstrapping is also a good method to evaluate the actual STD obtained from a ranking process, as mentioned earlier. These are simple ways to approximate error bars on reported performance scores.

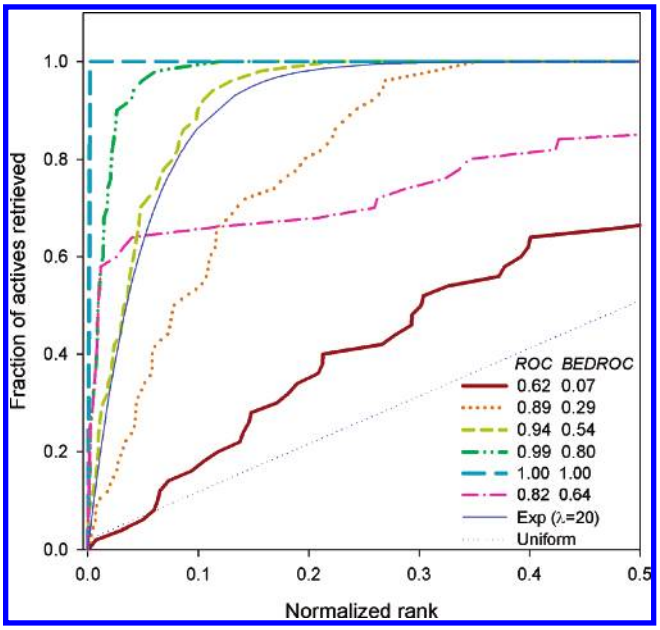

Figure 9. Different accumulation curves from sampling ($n = 50$, $N = 25000$) shown together with the corresponding *ROC* and *BEDROC* values where $\alpha = 20.0$. An exact CDF with $\lambda = 20$ is also shown to highlight the fact that the *BEDROC* metric returns a value of 1/2 for a curve close to this CDF.

Table 3. Minimum Number of Total Compounds N Solving eq 50 for $\Delta_{max} = 5\%$ and 1%

	N_{min} for $\Delta_{max} = 5\%$					N_{min} for $\Delta_{max} = 1\%$				
n	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$	$\alpha = 30$	$\alpha = 100$	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$	$\alpha = 30$	$\alpha = 100$
20	1031	2033	4066	6098	20 328	5086	10 034	20 066	30 100	100 332
40	2062	4066	8131	12 197	40 656	10 171	20 068	40 133	60 199	200 664
60	3093	6099	12 197	18 295	60 984	15 257	30 102	60 199	90 299	300 997
80	4124	8132	16 262	24 393	81 312	20 342	40 137	80 266	120 399	401 329
100	5156	10165	20 328	30 492	101 640	25 428	50 171	100 332	150 498	501 661
120	6187	12 198	24 393	36 590	121 967	30 513	60 205	120 399	180 598	601 993
140	7218	14 231	28 459	42 689	142 295	35 599	70 239	140 465	210 698	702 326
160	8249	16 264	32 525	48 787	162 623	40 685	80 273	160 532	240 797	802 658
180	9280	18 297	36 590	54 885	182 951	45 770	90 307	180 598	270 897	902 990
200	10 311	20 330	40 656	60 984	203 279	50 856	100 341	200 664	300 997	1 003 322

7. CONCLUSION

Despite the increasing numbers of performance evaluations of ranking methods in the context of VS, there is still no consensus on the metrics used to analyze the results. Area under the curve metrics such as ROC are not suited to the "early recognition" problem. In this study, we have used a probability framework to show that some metrics that seem to be different are in fact intimately related: the area under the accumulation curve, the average rank, and the area under the ROC curve. Parametric methods such as *Z*-score and ANOVA, also based on the average rank, are likewise not suited to the "early recognition" problem. These metrics dramatically fail to discriminate among three trivial cases outlined in the Introduction that must be correctly ranked by any metric intended to be usefully applied to VS. While some publications[9] assert that the ROC curve is not affected by the ratio of actives contained in the list of ranked compounds, we have shown with analytical work and through proper statistical simulation methods that this is not the case. Rather, we have found that the ratio of actives should be kept relatively small to ensure cross-evaluation comparisons of the scalar metrics.

The enrichment factor *EF* is another widely used metric intended to account for the "early recognition" problem. However, we criticize it, in line with precedent papers, for not discriminating the ranking "goodness" before the fractional threshold. Furthermore, if two ranked lists have similar initial enhancements, but differ significantly just after the threshold, they would not be differentiated by *EF*. Finally, *EF* has high variability, and its maximum value directly depends on the ratio of actives in the list. A better metric discussed previously in the literature is the robust initial enhancement *RIE* for which we derived an analytical formula of the denominator, avoiding a Monte Carlo simulation. *RIE* addresses the "early recognition" problem but suffers from several liabilities. First, its maximum varies significantly with the ratio of actives, making the comparison across evaluations difficult; the *RIE* upper bound is α, but we showed that it can be reached only if $\alpha R_a \ll 1$. It is also missing a probability-related interpretation that *ROC* has.

In this paper, we generalize the *ROC* metric to a new class of metrics that are adapted to the "early recognition" problem. In particular, we instantiate the *BEDROC* that uses an exponential of parameter α, which embodies the degree of "early recognition" required, to weight the contribution of the rank axis to the final score. *BEDROC*, like *ROC*, is bounded by the interval [0,1] and can be interpreted as the probability that a ranked active randomly selected will be positioned before a randomly selected compound distributed following an exponential of parameter α. This is true only when $\alpha R_a \ll 1$, that is, when the metric score is exempt from the saturation effect that would hinder cross-evaluation comparisons.

Any reported metric value is associated with a STD that we characterized both for the uniform distribution of actives and for a more general performance. The former STD is analytically calculated, and the latter is examined through extensive simulations. Trends were outlined: the STD is independent of N (the total number of scored compounds) when N is large enough, and the STD is reduced when the number of actives increases. The maximum STD for the *BEDROC* metric was obtained when the simulated CDF (corresponding to the accumulation or enrichment curve) matched the *BEDROC* weighting function determined by the choice of α. The many STDs calculated for a wide range of simulated CDFs and α values suggested an upper-bound $STD_{max} = 1/\mathrm{sqrt}(8n)$ that can be used in planning evaluations.

The "saturation effect" has been characterized as the variation of a metric with the ratio of actives while the nature of the ranking method remains unchanged (same PDF). It arises from having so many actives, which can potentially saturate the front of the list, that it is no longer possible to distinguish between good and excellent ranking methods. Saturating the scored compounds with actives modifies the shape of the accumulation and ROC curves, hence, all of the metric values. Deviations of 10% or more are observed in *AUAC* and *ROC* and even larger ones for weighted metrics like *BEDROC*. This must be taken into account when planning evaluations because the validity of any interevaluation comparisons (target, data set, etc.) depends on that. This is only important when good performances are obtained in the case of *RIE*, *EF*, and *AUAC* and is important for midrange performances in the case of *ROC* and *BEDROC*.

The parameter α of the *BEDROC* metric should be selected in light of the importance given to the early part of an ordered list. We give a basis for the choice of α, and by comparison to the enrichment factor, we propose that α be set to 20.0, which means that 80% of the maximum contribution to the *BEDROC* comes from the first 8% of the list. The corresponding model accumulation curve drawn from an exponential distribution of parameter $\alpha = 20$ takes a *BEDROC* value of 0.5. The comparison point is thus this model accumulation curve; the *BEDROC* metric becomes an early recognition "usefulness" metric as opposed to the *ROC* that is only a "departure from random" metric.

Practical guidelines are given to plan a healthy evaluation when the "early recognition" matters. More precisely, at least 50 actives are necessary to have an acceptable maximum STD of 0.05 for the *BEDROC* metric and $N_{min} = \alpha n/(2\Delta_{max})$ where Δ_{max} is the maximum relative deviation due to saturation effect allowed. We also suggest that if saturation is an issue with too many actives for a given set of inactives, the desired metric can be calculated by a bootstrap average with active replacement where the number of actives would match the condition for N_{min}. Also, the STD of a ranking result can be approximated using a bootstrap without replacement where a fraction of actives and inactives is randomly removed creating many samples of a reduced list size that keeps, on average, the same performance. The calculated variance through sampling of these reduced lists approximates the true STD on the metric value.

ACKNOWLEDGMENT

The authors would like to thank Robert Sheridan for his invaluable comments on the manuscript. We are also grateful to Georgia McGaughey for a careful proofread of this manuscript.

APPENDIX

The authors would like to acknowledge at this stage the usefulness of the Mathematica software in deriving and verifying the mathematical formulas presented in this work.

A.1. Other Formulations for the Area under the Accumulation Curve. We can define a metric *wAUAC* (weighted area under the accumulation curve) calculated as follows:

$$wAUAC = \frac{\int_0^1 F_a(x)\, w(x)\, dx}{\int_0^1 w(x)\, dx} \tag{A.1}$$

where $w(x)$ is any wel-behaved weighting function of the relative rank x. We now take the numerator and perform the integration by parts

$$\int_0^1 F_a(x)\, w(x)\, dx = F_a(1)\, F_w(1) - F_a(0)\, F_w(0) - \int_0^1 F_w(x) \frac{dF_a(x)}{dx} dx$$

$$= F_w(1) - \int_0^1 F_w(x)\, f_a(x)\, dx \tag{A.2}$$

where we have used the fact that the derivative of a CDF is the corresponding PDF. $F_w(x)$ is simply the integrated weighting function $F_w(x) = \int_0^x w(z)\, dz$. This leads to this recast of the *wAUAC*

$$wAUAC = \frac{F_w(1) - \int_0^1 F_w(x)\, f_a(x)\, dx}{F_w(1)} \tag{A.3}$$

and if we apply a uniform weighting function which, according to eq A.1 and eq 4, is just the *AUAC*, then we obtain eq A.4.

$$w(x) = 1$$
$$\Rightarrow F_w(x) = x$$
$$AUAC = 1 - \int_0^1 x\, f_a(x)\, dx$$
$$= 1 - \langle x \rangle \tag{A.4}$$

A.2. The *ROC* Metric as a Function of the *AUAC* Metric. Starting with a continuous definition of the *ROC* given below

$$ROC = \int_0^1 F_a(x)\, f_i(x)\, dx$$

we realize that, in a ranking experiment, if one knows the active PDF, the inactive PDF must also be known. In fact, if we focus our attention in a window of relative ranks dx that is not necessary infinitesimal, we know that the sum of actives and inactives within this window is proportional to the width of the window. This is simply due to the fact that the rank score is uniformly distributed. This leads to eq A.5. Like in the main text, $R_a = n/N$ is the ratio of actives and $R_i = (N - n)/N$

$$f_a(x)\, dx\, n + f_i(x)\, dx\, (N - n) = N\, dx \tag{A.5}$$
$$f_i(x) = \frac{1 - R_a\, f_a(x)}{R_i}$$

This leads to eq A.6, where the second term equals 1/2 as shown using integration by parts.

$$ROC = \int_0^1 F_a(x) \left(\frac{1 - R_a\, f_a(x)}{R_i} \right) dx$$

$$= \frac{1}{R_i} \int_0^1 F_a(x)\, dx - \frac{R_a}{R_i} \int_0^1 F_a(x)\, f_a(x)\, dx \tag{A.6}$$

$$\int_0^1 F_a(x)\, f_a(x)\, dx = [F_a(x)\, F_a(x)]_0^1 - \int_0^1 F_a(x)\, f_a(x)\, dx$$
$$= \frac{1}{2}$$

We finally obtain the relationship by assembling these results into eq A.7.

$$ROC = \frac{\int_0^1 F_a(x)\, dx}{R_i} - \frac{R_a}{2R_i}$$
$$= \frac{AUAC}{R_i} - \frac{R_a}{2R_i} \tag{A.7}$$

A second and conceptually easier path to obtain eq A.7 uses the maximum and minimum values of *AUAC*: $AUAC_{max} = 1 - n/(2N)$ and $AUAC_{min} = n/(2N)$. In fact, on the basis of eq A.7,

$$ROC = \frac{AUAC - AUAC_{min}}{AUAC_{max} - AUAC_{min}}$$
$$= \frac{AUAC}{R_i} - \frac{R_a}{2R_i} \tag{A.8}$$

This unveils the simplicity of the area under the ROC curve idea.

A.3. The Analytical Formula for the Average Uniformly Distributed Exponential Sum. The calculation of the *RIE*, from eq 16, necessitates evaluating the average of the sum of exponentials of all distributions of n actives in N possible ranks if each position is equiprobable. This was originally calculated via a Monte Carlo simulation that was slow to converge, but we propose here an analytical formula. The first step is to explicitly write the average over all combinations of position of n actives in the form of eq A.9, where $C_n^N = N!/(N-n)!n!$ is the number of ways that n elements can be arranged in N compartments and C_i is a particular realization and consists in a list of n ranks r_j taking values between 1 and N without repetition. Equation A.9 can be transformed by counting the number of times that a specific position k is counted through all the combinations, and the count is simply the number of ways one can arrange the $n-1$ remaining actives (not at position k) in the $N-1$ remaining positions; this is expressed in eq A.10. Finally, the sum of N exponentials is a geometric series for which the sum is known and can be found with the telescoping series method. The final answer is given by eq A.13.

$$\left\langle \sum_{i=1}^{n} e^{-\alpha x_i} \right\rangle_r = \frac{1}{C_n^N} \sum_{i=1}^{C_n^N} \sum_{j \in C_i}^{n} e^{-\alpha r_j/N} \tag{A.9}$$

$$= \frac{1}{C_n^N} \sum_{k=1}^{N} C_{n-1}^{N-1} e^{-\alpha k/N} \tag{A.10}$$

$$= \frac{n}{N} \sum_{k=1}^{N} (e^{-\alpha/N})^k \tag{A.11}$$

$$= \frac{n}{N} e^{-\alpha/N} \left(\frac{1 - e^{-\alpha}}{1 - e^{-\alpha/N}} \right) \tag{A.12}$$

$$= \frac{n}{N} \left(\frac{1 - e^{-\alpha}}{e^{\alpha/N} - 1} \right) \tag{A.13}$$

A.4. The *wAUAC* as a Function of the *RIE* when the Weight is a Decreasing Exponential. The proof is straightforward from eq A.3

$$F_w(x) = \int_0^x e^{-\alpha z} dz$$

$$= \frac{1 - e^{-\alpha x}}{\alpha}$$

$$wAUAC = 1 - \frac{\int_0^1 \left(\frac{1 - e^{-\alpha x}}{\alpha} \right) f_a(x)\, dx}{(1 - e^{-\alpha})/\alpha}$$

$$= 1 - \frac{{}^1/_\alpha \int_0^1 f_a(x)\, dx}{(1 - e^{-\alpha})/\alpha} + \frac{{}^1/_\alpha \int_0^1 e^{-\alpha x} f_a(x)\, dx}{(1 - e^{-\alpha})/\alpha}$$

$$= \frac{RIE}{\alpha} + \frac{1}{1 - e^{\alpha}} \tag{A.14}$$

$$wAUAC \approx \frac{RIE}{\alpha} \text{ if } \alpha \text{ is sufficiently large}$$

Of course, only the *wAUAC* with an exponentially decreasing weighting function can be written as a function of the *RIE* in eq A.14.

A.5. The *ROC* and the *AUAC* Metrics as a Function of the *RIE*. From eq 20, we write the *RIE* in the continuous formulation, modify this equation with basic algebra, and take the derivatives on both sides evaluated at $\alpha = 0$ to obtain eq A.15.

$$RIE = \frac{\int_0^1 f_a(x)\, e^{-\alpha x}\, dx}{{}^1/_\alpha (1 - e^{-\alpha})}$$

$$\frac{RIE(1 - e^{-\alpha})}{\alpha} = \int_0^1 f_a(x)\, e^{-\alpha x}\, dx$$

$$\left[\frac{\partial}{\partial \alpha} \left(\frac{RIE(1 - e^{-\alpha})}{\alpha} \right) \right]_{\alpha=0} = \left[\frac{\partial}{\partial \alpha} \left(\int_0^1 f_a(x)\, e^{-\alpha x}\, dx \right) \right]_{\alpha=0} \tag{A.15}$$

Then the right-hand side (rhs) and the left-hand side (lhs) are simultaneously treated. From the rhs, the derivatives can be taken inside the integral if $f_a(x)$ is well-behaved as expected, and using a Taylor series expansion of the exponential function and setting α to zero, we obtain eq A.16. This follows from the moment-generating functions property.[22]

$$\left(\frac{\partial}{\partial \alpha} \int_0^1 f_a(x)\, e^{-\alpha x}\, dx \right)_{\alpha=0} = -\langle x \rangle \tag{A.16}$$

Next, the lhs is calculated, and we need to be careful here because α cannot be zero because of the denominator. We can avoid this difficulty by taking the limit when α goes to zero, and this leads to eq A.17.

$$\lim_{a \to 0} \left[\frac{\partial}{\partial \alpha} \left(\frac{RIE(1 - e^{-\alpha})}{\alpha} \right) \right] = \left(\frac{\partial RIE}{\partial \alpha} \right)_{\alpha=0} - \frac{1}{2} \tag{A.17}$$

Combining both the lhs and rhs leads to the following equation which can be rewritten in terms of *AUAC* and *ROC* in eqs A.18 and A.19 using eqs A.4 and A.7.

$$\frac{1}{2} - \langle x \rangle = \left(\frac{\partial RIE}{\partial \alpha} \right)_{\alpha=0}$$

$$AUAC = \left(\frac{\partial RIE}{\partial \alpha} \right)_{\alpha=0} + \frac{1}{2} \tag{A.18}$$

$$ROC = \frac{1}{R_i} \left(\frac{\partial RIE}{\partial \alpha} \right)_{\alpha=0} - \frac{R_a}{2R_i} + \frac{1}{2R_i}$$

$$= \frac{1}{R_i} \left(\frac{\partial RIE}{\partial \alpha} \right)_{\alpha=0} + \frac{1}{2} \tag{A.19}$$

Approximating the *RIE* derivative with a finite difference leads to the following equations:

$$\left(\frac{\partial RIE}{\partial \alpha} \right)_{\alpha=0} \approx \frac{RIE(\Delta\alpha) - RIE(0)}{\Delta\alpha}$$

$$\approx \frac{RIE(\alpha) - 1}{\alpha}$$

$$ROC \approx \frac{RIE(\alpha) - 1}{\alpha R_i} + \frac{1}{2};\ \alpha \text{ is small}$$

$$AUAC \approx \frac{RIE(\alpha) - 1}{\alpha} + \frac{1}{2};\ \alpha \text{ is small}$$

If α is small enough, both the *ROC* and the *AUAC* can be calculated from the *RIE*.

A.6. Variance of *AUAC* Metric. We have shown the following identities for the *AUAC* in section A.1:

$$\begin{aligned} AUAC &= 1 - \langle x \rangle \\ &= 1 - \frac{1}{nN}\sum_{i=1}^{n} r_i \end{aligned} \tag{A.20}$$

We can use eq A.20 to calculate the variance of the *AUAC* metric as shown in the next equation where the *r* in subscript means taken over a random (or uniform) distribution.

$$\mathrm{Var}[AUAC]_r = \frac{\mathrm{Var}[\sum_{i=1}^{n} r_i]_r}{n^2N^2}$$

In order to calculate the variance of the summation of rank, we first need to calculate the average value of the rank summation over random. This is done below in a similar way as the average *RIE* value was calculated previously (section A.3), so we will skip detailed explanations as the math is self-explanatory. The resulting average is given by eq A.21.

$$\begin{aligned} \langle \sum_{i=1}^{n} r_i \rangle_r &= \frac{1}{C_n^N}\sum_{i=1}^{C_n^N}\sum_{j\in C_i}^{n} j \\ &= \frac{C_{n-1}^{N-1}}{C_n^N}\sum_{i=1}^{N} i \\ &= \frac{n}{N}\frac{N(N+1)}{2} \\ &= \frac{n(N+1)}{2} \end{aligned} \tag{A.21}$$

Since we are there, it is straightforward from eq A.21 to calculate the average *AUAC* metric over a uniform distribution of actives:

$$\begin{aligned} \langle AUAC \rangle_r &= \frac{\langle \sum_{i=1}^{n} r_i \rangle_r}{nN} \\ &= \frac{N+1}{2N} \end{aligned} \tag{A.22}$$

To calculate the variance of the sum, we use the identity A.21 and make explicit the average over all possible combinations of *n* actives among *N* possible ranks, which is simply replacing the averaging brackets over random in the first equation below. The two summations inside the averaging summation are developed into two grouped summations. The first one pertains to the same active ranks (diagonal of the matrix of pairs). The second summation term corresponds to the upper triangle of the matrix of pair of ranks (this should also count the lower triangle of the matrix, hence, the factor 2). The third equation comes from applying the outer summation over all combinations on each of the rank pair summations. The first term can be obtained by counting the number of times a specific rank occurs when considering all of the combinations. As for the *RIE* average, the answer is that, if you imagine that you keep fixed the *i*th term, then it becomes clear that you can still position $n-1$ remaining actives within the $N-1$ remaining positions. This explains the first term in the rhs of the third equation. For the second term of the same equation, we are considering all pairs of different ranks without repetition. Again, if you count the number of times a pair occurs by counting the number of remaining possible combinations of $n-2$ actives in the $N-2$ remaining positions, you obtain the second term. The rest of the mathematical development is pure algebra and series calculation.

$$\begin{aligned} \mathrm{Var}[\sum_{i=1}^{n} r_i]_r &= \langle (\sum_{i=1}^{n} r_i)(\sum_{j=1}^{n} r_j) \rangle_r - \langle \sum_{i=1}^{n} r_i \rangle_r^2 \\ &= \frac{1}{C_n^N}\sum_{k=1}^{C_n^N}[\sum_{i=1}^{i\in C_k} r_i^2 + 2\sum_{j\neq l}^{j,l\in C_k} r_j r_l] - \langle \sum_{i=1}^{n} r_i \rangle_r^2 \\ &= \frac{C_{n-1}^{N-1}}{C_n^N}\sum_{i=1}^{N} i^2 + 2\frac{C_{n-2}^{N-2}}{C_n^N}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} jk - \langle \sum_{i=1}^{n} r_i \rangle_r^2 \\ &= \frac{n}{N}\frac{N(N+1)(2N+1)}{6} + 2\frac{n(n-1)}{N(N-1)}\sum_{j=1}^{N-1}\left[\frac{N(N+1)-j(j+1)}{2}\right] - \langle \sum_{i=1}^{n} r_i \rangle_r^2 \\ &= \frac{n}{N}\frac{N(N+1)(2N+1)}{6} + 2\frac{n(n-1)}{N(N-1)} \times \frac{(N-1)(N+1)(2+3N)N}{24} - \langle \sum_{i=1}^{n} r_i \rangle_r^2 \\ &= \frac{n(N-n)(N+1)}{12} \end{aligned}$$

This simple term for the variance of the sum of rank of actives can now be used in the *AUAC* variance equation to give the needed result shown in eq A.23.

$$\mathrm{Var}[AUAC]_r = \frac{(N-n)(N+1)}{12nN^2} \tag{A.23}$$

A.7. Variance of *ROC* Metric. In the case of the *ROC* variance over a uniform distribution of actives, we can reuse the previous work and write the *ROC* metric as a function of the *AUAC* metric for which we already know the desired variance. The first quantity we need to calculate is the average *ROC* over a uniform distribution of actives. This is done below, leading to eq A.24. It might seem surprising to the reader that it is not exactly one-half. However, if we

realize that the actual *ROC* graph is bumpy and never a straight line, then this result is easily understood. Of course, when $N \gg n$, the second term of eq A.24 is negligible.

$$ROC = \frac{AUAC}{R_i} - \frac{R_a}{2R_i}$$

$$\langle ROC \rangle_r = \frac{\langle AUAC \rangle_r}{R_i} - \frac{R_a}{2R_i} = \frac{(N+1)}{2N}\frac{N}{(N-n)} - \frac{n}{2(N-n)} = \frac{1}{2} + \frac{1}{2(N-n)} \quad \text{(A.24)}$$

The *ROC* variance over a random distribution of actives is as simple to obtain:

$$ROC = \frac{AUAC}{R_i} - \frac{R_a}{2R_i}$$

$$\mathrm{Var}[ROC]_r = \frac{\mathrm{Var}[AUAC]_r}{R_i^2} = \frac{N+1}{12n(N-n)} \quad \text{(A.25)}$$

From eq A.25, we can realize that, if n is small and N very large, the standard error of *ROC* is inversely proportional to the square root of n, which makes sense in terms of statistical analysis.

A.8. Variance of *EF*. We calculate the random average and the random variance of the enrichment factor metric the same way as for the precedent examined metrics. We first define a δ function that has a value of 1 if an active i is found before the threshold position defined by χN, which is position 100 if $N = 10\,000$ and $\chi = 1\%$, for instance; the δ function takes a value of 0 otherwise. This way, the *EF* can be written as eq A.26, and the average over random follows.

$$EF = \frac{\sum_{i=1}^{n} \delta_i}{\chi n} \quad \text{where } \delta_i = \begin{cases} 1, & r_i \le \chi N \\ 0, & r_i > \chi N \end{cases} \quad \text{(A.26)}$$

$$\Rightarrow \langle EF \rangle_r = \frac{\langle \sum_{i=1}^{n} \delta_i \rangle_r}{\chi n} \quad \text{where } \delta_i = \begin{cases} 1, & r_i \le \chi N \\ 0, & r_i > \chi N \end{cases}$$

This means that we need to calculate the numerator of the equation above, and we already know how to do this on the basis of previous development in this appendix. Hence, we obtain eq A.27, which contains the lower brackets meaning "the highest integer smaller than".

$$\langle \sum_{i=1}^{n} \delta_i \rangle_r = \frac{1}{C_n^N} \sum_{i=1}^{C_n^N} \sum_{j \in C_i}^{n} \delta_j = \frac{1}{C_n^N} \sum_{i=1}^{N} C_{n-1}^{N-1} \delta_i = \frac{n}{N} \lfloor \chi N \rfloor \quad \text{(A.27)}$$

We write the average *EF* over random in eq A.28.

$$\langle EF \rangle_r = \frac{\lfloor \chi N \rfloor}{\chi N} \quad \text{(A.28)}$$

This in hand, we can start the calculation of the variance in focusing on the first term in the rhs of the equation below.

$$\mathrm{Var}[EF]_r = \langle EF^2 \rangle_r - \langle EF \rangle_r^2$$

We proceed as for the calculation of the *AUAC*, except here it is simpler because we obtain 0 or 1 in the summation. We suppose that χN is $< N$, which should always be the case. Doing the math results in eq A.29, which can be substituted back into the formula of the variance to lead to eq A.30.

$$\begin{aligned} \langle EF^2 \rangle_r \chi^2 n^2 &= \langle (\sum_{j=1}^{n} \delta_j)(\sum_{k=1}^{n} \delta_k) \rangle_r \\ &= \langle (\sum_{j=1}^{n} \delta_j)(\sum_{k=1}^{n} \delta_k) \rangle_r \\ &= \frac{1}{C_n^N} \sum_{i=1}^{C_n^N} (\sum_{j \in C_i}^{n} \delta_j)(\sum_{k \in C_i}^{n} \delta_k) \\ &= \frac{1}{C_n^N} \sum_{i=1}^{C_n^N} (\sum_{j \in C_i}^{n} \delta_j^2 + 2 \sum_{k \neq l}^{k,l \in C_i} \delta_k \delta_l) \\ &= \frac{1}{C_n^N} (C_{n-1}^{N-1} \sum_{j=1}^{N} \delta_j^2 + 2 C_{n-2}^{N-2} \sum_{k=1}^{N} \sum_{l=k+1}^{N-1} \delta_k \delta_l) \\ &= \frac{C_{n-1}^{N-1}}{C_n^N} \sum_{j=1}^{\lfloor \chi N \rfloor} 1 + \frac{2 C_{n-2}^{N-2}}{C_n^N} \sum_{k=1}^{\lfloor \chi N \rfloor} \sum_{l=k+1}^{\lfloor \chi N \rfloor} 1 \\ &= \frac{n}{N} W + \frac{n(n-1)}{N(N-1)} W(W-1) \end{aligned} \quad \text{(A.29)}$$

$$\mathrm{Var}[EF]_r = \frac{W}{nN\chi^2}\left[1 + \frac{n-1}{N-1}(W-1)\right] - \frac{W^2}{\chi^2 N^2} \quad \text{(A.30)}$$

where W is defined as $\lfloor \chi N \rfloor$.

A.9. Variance of *RIE* Metric. The variance of the *RIE* metric is mathematically heavier to calculate although very similar to previous calculations of variances done earlier in this appendix. We start with the definition of the *RIE*

$$RIE = \frac{\sum_{i=1}^{n} e^{-\alpha r_i/N}}{\langle \sum_{i=1}^{n} e^{-\alpha r_i/N} \rangle_r}$$

and we simplify the terms by setting $\alpha/N = \beta$, and we write the variance of the *RIE* when the active distribution is uniform. It is to be noted here that r_i corresponds to the actual rank of the ith active in the ordered list.

$$\mathrm{Var}[RIE]_r = \frac{\mathrm{Var}[\sum_{j=1}^{n} e^{-\beta r_j}]_r}{\langle \sum_{j=1}^{n} e^{-\beta r_j} \rangle_r^2} \qquad \text{(A.31)}$$

The only difficult term is the variance of the sum of exponentials that can be written like this:

$$\mathrm{Var}[\sum_{j=1}^{n} e^{-\beta r_j}]_r = \langle (\sum_{j=1}^{n} e^{-\beta r_j})(\sum_{k=1}^{n} e^{-\beta r_k}) \rangle_r - \langle \sum_{j=1}^{n} e^{-\beta r_j} \rangle_r^2$$

As in previous mathematical developments, we can identify the terms that command more attention, the average of the cross terms. The average can be expanded over all possible combinations of n actives into equally probable N positions. Using the same technique as before, we can write the summation over the combinations into summations over positions by counting how many times each rank contributes. The rest of the development is simply applying mathematical rules. One that we needed to use is the geometric series formula $\sum_{i=m}^{n} r^i = (r^m - r^{n+1})/(1 - r); |r| < 1$, where r is the exponential with a negative exponent prefactor. This leads to eq A.32 that can be substituted back into eq A.31 to lead eq A.33.

$$\langle (\sum_{j=1}^{n} e^{-\beta r_j})(\sum_{k=1}^{n} e^{-\beta r_k}) \rangle_r = \frac{1}{C_n^N} \sum_{i=1}^{C_n^N} (\sum_{j \in C_i}^{n} e^{-2\beta r_j} + 2 \sum_{j \neq k}^{j,k \in C_i} e^{-\beta(r_j + r_k)})$$

$$= \frac{C_{n-1}^{N-1}}{C_n^N} \sum_{j=1}^{N} e^{-2\beta j} + 2\frac{C_{n-2}^{N-2}}{C_n^N} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} e^{-\beta(j+k)}$$

$$= \frac{n}{N} \sum_{j=1}^{N} (e^{-2\beta})^j + 2\frac{n(n-1)}{N(N-1)} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} e^{-\beta(j+k)}$$

$$= \frac{n}{N}\left[\frac{1 - e^{-2\beta N}}{e^{2\beta} - 1}\right] + 2\frac{n(n-1)}{N(N-1)} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} e^{-\beta(j+k)}$$

$$= \frac{n}{N}\left[\frac{1 - e^{-2\beta N}}{e^{2\beta} - 1}\right] + 2\frac{n(n-1)}{N(N-1)} \sum_{j=1}^{N-1} e^{-\beta j} \sum_{k=j+1}^{N} e^{-\beta k}$$

$$= \frac{n}{N}\left[\frac{1 - e^{-2\beta N}}{e^{2\beta} - 1}\right] + 2\frac{n(n-1)}{N(N-1)} \sum_{j=1}^{N-1} e^{-\beta j} \left(\frac{e^{-\beta(j+1)} - e^{-\beta(N+1)}}{1 - e^{-\beta}}\right)$$

$$= \frac{n}{N}\left[\frac{1 - e^{-2\beta N}}{e^{2\beta} - 1}\right] + 2\frac{n(n-1)}{N(N-1)} \sum_{j=1}^{N-1} e^{-\beta j} \left(\frac{e^{-\beta(j+1)} - e^{-\beta(N+1)}}{1 - e^{-\beta}}\right)$$

$$= \frac{n}{N}\left[\frac{1 - e^{-2\beta N}}{e^{2\beta} - 1}\right] + 2\frac{n(n-1)}{N(N-1)} \frac{e^{-\beta}}{(1 - e^{-\beta})} \times \left[\frac{e^{-2\beta} - e^{-2\beta N}}{1 - e^{-2\beta}} - e^{-\beta N}\left(\frac{e^{-\beta} - e^{-\beta N}}{1 - e^{-\beta}}\right)\right]$$

$$= \frac{n}{N}\left[\frac{1 - e^{-2\alpha}}{e^{2\alpha/N} - 1}\right] + 2\frac{n(n-1)}{N(N-1)} \frac{e^{-2\alpha}(e^{\alpha/N} - e^{\alpha})(1 - e^{\alpha})}{(e^{\alpha/N} - 1)^2(1 + e^{\alpha/N})} \qquad \text{(A.32)}$$

$$\mathrm{Var}[RIE]_r = \frac{\left[\frac{1 - e^{-2\alpha}}{e^{2\alpha/N} - 1}\right] + \frac{2(n-1)}{(N-1)} \frac{e^{-2\alpha}(e^{\alpha/N} - e^{\alpha})(1 - e^{\alpha})}{(e^{\alpha/N} - 1)^2(1 + e^{\alpha/N})}}{\frac{n}{N}\left(\frac{1 - e^{-\alpha}}{e^{\alpha/N} - 1}\right)^2} - 1 \qquad \text{(A.33)}$$

A.10. Variance of *wAUAC*. The *wAUAC* variance with an exponential weighting function for a uniform distribution of the actives can be readily calculated using the relationship between the *RIE* and the *wAUAC*, and the weighting function is an exponential PDF, given by eq A.14. The answer is simply eq A.34 where we can replace the $\mathrm{Var}[RIE]_r$ term by eq A.33

$$\mathrm{Var}[wAUAC]_r = \frac{\mathrm{Var}[RIE]_r}{\alpha^2} \qquad \text{(A.34)}$$

Supporting Information Available: We provide C++ and Python (www.python.org) classes with few programs and scripts to make the reproduction of the data presented in this paper easy; more importantly, these tools allow the calculation of the *BEDROC* metric. We also describe a format to store the results of a ranking experiment that was originally designed by R. Sheridan. This information is available free of charge via the Internet at http://pubs.acs.org.

Note Added After ASAP Publication. This paper was published ASAP on February 9, 2007 with errors in the equations (A.6) and (A.33); the corrected version was published ASAP on February 13, 2007.

REFERENCES AND NOTES

(1) Chen, H. M.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. Evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.* **2006**, *46*, 401−415.
(2) Kairys, V.; Fernandes, M. X.; Gilson, M. K. Screening drug-like compounds by docking to homology models: A systematic study. *J. Chem. Inf. Model.* **2006**, *46*, 365−379.

(3) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 2743−2749.
(4) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235−249.
(5) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: Evaluation of current docking tools. *J. Mol. Model.* **2003**, *9*, 47−57.
(6) Muegge, I.; Enyedy, I. J. Virtual screening for kinase targets. *Curr. Med. Chem.* **2004**, *11*, 693−707.
(7) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Med. Chem.* Submitted for publication.
(8) Cornell, W.D. Recent evaluations of high throughput docking methods for pharmaceutical lead finding − Consensus and caveats. In *Annual Reports in Computational Chemistry*; Spellmeyer, D. C., Ed.; Elsevier: Amsterdam, The Netherlands, 2006; Vol. 2, pp 302−328.
(9) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.
(10) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of retrieval in similarity searches of chemical databases: A review of performance measures. *J. Mol. Graphics Modell.* **2000**, *18*, 343−357.
(11) Cleves, A. E.; Jain, A. N. Robust ligand-based modeling of the biological targets of known drugs. *J. Med. Chem.* **2006**, *49*, 2921−2938.
(12) Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein−Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2005**, *47*, 6128−6136.
(13) Klon, A. E.; Glick, M.; Davies, J. W. Application of machine learning to improve the results of high-throughput docking against the HIV-1 protease. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2216−2224.
(14) Webb, G. I.; Ting, K. M. On the application of ROC analysis to predict classification performance under varying class distributions. *Mach. Learn.* **2005**, *58*, 25−32.
(15) Drummond, C.; Holte, R. C. Cost curves: An improved method for visualizing classifier performance. *Mach. Learn.* **2006** [Online].
(16) Seifert, M. H. J. Assessing the discriminatory power of scoring functions for virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1456−1465.
(17) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing scoring functions for protein−ligand interactions, *J. Med. Chem.* **2004**, *47*, 3032−3047.
(18) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(19) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395−1406.
(20) Killeen, P. R.; Taylor, T. J. Symmetric receiver operating characteristics. *J. Math. Psychol.* **2004**, *48*, 432−434.
(21) Swets, J. A. Indexes of Discrimination Or Diagnostic-Accuracy - Their Rocs and Implied Models. *Psychol. Bull.* **1986**, *99*, 100−117.
(22) Ross, S. M. *Introduction to Probability Models*; Academic Press: San Diego, CA, 2002.
(23) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861−874.
(24) Hanley, J. A.; Mcneil, B. J. The meaning and use of the area under a receiver operating characteristic (Roc) Curve. *Radiology* **1982**, *143*, 29−36.
(25) Swets, J. A.; Dawes, R. M.; Monahan, J. Better decisions through science. *Sci. Am.* **2000**, *283*, 82−87.
(26) Swets, J. A. Measuring the Accuracy of Diagnostic Systems. *Science* **1988**, *240*, 1285−1293.
(27) Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.* **1975**, *12*, 387−415.

CI600426E