J Comput Aided Mol Des (2008) 22:239–255
DOI 10.1007/s10822-008-9170-2

What do we know and when do we know it?

Anthony Nicholls

Received: 16 November 2007 / Accepted: 2 January 2008 / Published online: 6 February 2008

Abstract Two essential aspects of virtual screening are considered: experimental design and performance metrics. In the design of any retrospective virtual screen, choices have to be made as to the purpose of the exercise. Is the goal to compare methods? Is the interest in a particular type of target or all targets? Are we simulating a 'real-world' setting, or teasing out distinguishing features of a method? What are the confidence limits for the results? What should be reported in a publication? In particular, what criteria should be used to decide between different performance metrics? Comparing the field of molecular modeling to other endeavors, such as medical statistics, criminology, or computer hardware evaluation indicates some clear directions. Taken together these suggest the modeling field has a long way to go to provide effective assessment of its approaches, either to itself or to a broader audience, but that there are no technical reasons why progress cannot be made.

Keywords Virtual screening · Statistics · AUC · ROC curves · Metrics

Introduction

Virtual screening in the pharmaceutical industry is an essential part of molecular modeling's contribution to lead discovery and, to a lesser extent, lead optimization. This has led to considerable research into what method or approach works best, typically by means of 'retrospective' evaluations, i.e. attempting to predict future, i.e. prospective, behavior by appraising techniques on known systems. Despite this there is no agreed upon theory as to how to conduct a retrospective evaluation. As a consequence, it is very difficult for an outsider to assess if methods are getting better, have stayed the same, or even worsened over time. In a practical enterprise, such as drug discovery, the proposed benefits of virtual screening, i.e. avoiding the cost and time of a real screen, have to be weighed against one simple question: does it actually work? Without proper metrics of success, i.e. ones that go beyond the anecdotal, molecular modeling is not guaranteed a vibrant future.

Observed as a general exercise, there are four elements that ought to be standard for any prediction study, whether of a virtual screen, or any general pattern recognition method. The first is whether the study is well designed. The second is what metrics are used to evaluate the outcome. The third is a consideration of significance, i.e. error analysis. And the fourth is a reliable assessment of whether the results are particular or general. All four aspects are important and yet it is rare for a study in any field to meet all these criteria. Even in the most critical part of drug discovery, i.e. clinical trials, there is considerable room for improvement, as several recent retrospective studies of the medical literature have demonstrated [1, 2]. In reports on virtual screening, in fact in molecular modeling in general, it is rare to find an adequate consideration of any of these issues.

Why is this? Why is the modeling field so poor at the most basic elements of evaluation? A charitable view would be that, as with communication skills, most modelers receive little appropriate formal training. Certainly there is no central resource, whether scholastic review, book or paper. A slightly less charitable view is that journals have not developed standards for publication and as such there is little Darwinian pressure to improve what the community sees as acceptable. It is to be hoped that this is a learning

A. Nicholls (✉)
OpenEye Scientific Software, Inc, 9D Bisbee Crt,
Santa Fe, NM 87508, USA
e-mail: anthony@eyesopen.com

curve, i.e. that editors will eventually appreciate what is required in a study. An extreme view is that we are poor at evaluations because we simply do not matter very much. If large fortunes were won or lost on the results from computational techniques there would be immense debate as to how to analyze and compare methods, on what we know and exactly when we know it. There would be double blind, prospective and rigorously reviewed studies of a scale and depth unknown in our field but common in, for instance, clinical trials. In short, there would be standards.

In the hope that virtual screening is, in fact, worthwhile we provide comment, suggestions and research on two important aspects, namely experimental design and performance metrics. Although the two are intimately linked, i.e. an experiment should be designed with a mind to what is being measured, there are distinguishable aspects. On experimental design, extensive properties, such as number of targets, actives and inactives, need to reflect a statistical understanding of the current unreliability, or high *variance*, of methods [3–5]. So dominant is this variance that it almost renders moot any discussion of other matters, such as decoy design. However, ultimately all aspects of design are important. On decoy selection we suggest the necessity of clarifying design intent and suggest four broad categorizations. In analyzing results, the issue of correlation is considered. This often arises in the context of the 2D similarity of actives from congeneric series, but the general issue also concerns decoys, targets and methods. Research is proposed that would clarify essential and poorly understood issues, such as the transference of predictability between closely related systems. On evaluation metrics we examine the AUC (Area Under the Curve) of ROC (Receiver Operator Characteristic) curves [6–9]. Consideration of why the AUC is a popular measure in many disciplines suggests standards by which virtual screening metrics ought to be judged. Finally, by evaluating average properties of large numbers of systems, and by considering simple cost/benefit examples, we bring into question the validity and utility of metrics proposed to capture 'early' behavior.

Experimental design

In what follows we consider the importance of both intensive and extensive properties of an experiment. An intensive property is something intrinsic to a design, whereas extensive properties change with the size of the system. For example, the type of decoys used in a retrospective study is an intensive property; the number of such is an extensive property. We believe the most overlooked intensive characteristic is the design goal, i.e. what is trying to be proved. This typically falls into a few discrete classes and appropriate labeling would help combine lessons from different studies. For extensive quantities we consider how common statistical approaches can aid the choice of numbers of actives, decoys and targets. Finally, actives, decoys, targets or methods are not always independent and this has to be quantified even in as simple a matter as comparing two programs. Techniques for accounting for correlation within an experimental design are known but rarely applied.

Intensive properties

One of the most basic issues in designing a retrospective screen is how to chose decoys. Typically there are a certain number of active compounds and one wishes to see if a method can distinguish these from a second set, presumed inactive. This is the most basic of classification problems. Is X of type A or type B? The legal system often has the same dilemma, e.g. was X at the scene of a crime or not? A police line-up has all the components of a virtual screen. Usually the number of actives (suspects) is small, usually one. The number of decoys (called 'fillers') has to be sufficient that random selection does not compete with real recognition; a minimum of four is usual. But it cannot be so large that guilt is hidden within the statistical variance of the innocent. The fillers need to be convincing, i.e. not outlandishly dissimilar to the guilty party, but not too similar or even potentially also at the scene (i.e. false false positives). As courtroom verdicts can depend on the appropriateness of a line-up, standard procedures are well known.

We make the argument for four types of virtual screening experiments; each with its own intent. Each of the four designs ultimately consists of a random selection of decoys but after the application of different filters.

(A) *Universal*. Any compound available to be physically screened, typically either from vendors or corporate collections.
(B) *Drug-like*. Available and drug-like, typically by applying simple chemical filters.
(C) *Mimetics*. Available, drug-like and matched to *known* ligands by simple physical properties.
(D) *Modeled*. Available, drug-like and derived using 3D modeling on known ligands or the intended targets.

Although no classification scheme could be perfect, fair comparison of studies requires an alignment of intent. In general, decoys get 'harder' from A to D, although this is not necessarily true on a case-by-case basis and is itself an interesting area of research.

The first, and perhaps least in favor, is the *universal* selection of decoys. A catalogue of compounds from a vendor or set of vendors is treated as a general population from which to draw. An example of a virtual study with *universal* decoys can be found in Rognan et al. [10].

Although this method is now uncommon, it has an interesting intent. Faced with all compounds available for testing, does a method distinguish known actives without using *prior* knowledge of what makes a compound active? Drug discovery has a long and successful history of grinding up exotic plants and animals and screening for activity and so this is a reasonable, if old-fashioned, approach. In the Rognan set, for example, we find I_3, not likely to be a drug but none-the-less an interesting molecule. The problem with *universal* decoys is two fold. First, is it random enough? The space of all possible chemistry is exceedingly vast [11, 12]. The concept that a few thousand compounds, in particular from a vendor database, could act as a thorough sampling is implausible. In fact, there is now evidence suggesting known chemistry is very restrictive [12]. Because of 'inductive bias', a concept frequently highlighted by Jain [13, 14], we tend to make what we know might work, instead of sampling of what can be made. Second, is it possible a universal decoy such as I_3 might stand out pretty much the same way a shady character would stand out against a selection of school children, shop clerks and nuns? Paul Hawkins has described this as the 'dog' test [15], i.e. if your dog could tell the difference between the actives and inactives what have you really proved? Actually, potentially a lot but only if the rest of the experiment is designed with this choice of decoys in mind. The problem here is one of dynamic range. If it is too easy to distinguish an active then the only way to distinguish between the methods is to test many, many times, i.e. to improve the statistical power necessary to rank one method above another. As is well known, and discussed below, the error in any metric depends on both the number of actives and the number of inactives. While it is trivial to increase the number of (presumed) inactives almost without bound, the number of actives is normally very finite. Only in some of the more impressive published studies does the number of actives exceed a hundred [3, 4] and it is this limitation that really hinders random decoys being an effective experimental design. We note that this is only a presumed inadequacy of *universal* decoys; in fact such decoys may prove difficult for some computation methods, the Hawkins dog test not with-standing. The point is that a presumed limitation can be overcome by applying basic error analysis.

A more typical selection procedure is to choose from a decoy set that is 'drug-like'. Of course, there is no rigorous definition of 'drug-like', but this does not stop it being widely used. The intent is to mimic modern physical screens and not test everything but instead be guided by current dogma as to what a drug might look like. The most prevalent of these descriptions is the famous Lipinski Rule-of-Five, but there are many variants [16]. This is not dissimilar to how police line-ups are actually constructed; 'fillers' are normally acquired from local jails. In theory, this should also be a harder test of methods because there are less easily discernable inactives, although this is not proven. Examples of this approach are the studies of McGaughey et al. and Warren et al. One potential advantage of this approach is that because decoys are derived from characterized collections they are more likely to be known to be inactives. This is typically only an assumption for *universal* decoy sets. It is debatable as to how big a problem false decoys are, but clearly they do not help. There are also issues with *drug-like* decoys. Some companies' collections are heavily biased towards certain targets that may or may not be related to the retrospective study at hand. The study by McGaughey et al. reported significant differences in the efficiency of decoys chosen from the MMDR, a kind of 'consensus' drug-like collection, compared to ones from their internal Merck database. This trade of generality for local applicability is a characteristic of many aspects of evaluations. For instance, should targets be chosen to represent all possible systems, a subset of pharmaceutical interest or a class within that subset? What is gained in local applicability is often lost to generality and prospective predictability.

The third approach is to find *mimetic* or *modeled* decoys. These are meant to stress-test methods and should be used to compare approaches, rather than necessarily evaluate real-world performance. *Mimetic* decoys are constructed so that 'simple' methods cannot tell known ligands from decoys. The rationale is utilitarian; why should one chose to use sophisticated methods when simple, ligand-based, ones can do just as well? Approaches include matching physical properties, for instance size, number of hydrogen bond donors and acceptors, lipophilicity, charge or flexibility [17]. An example of this approach is the DUD dataset [18] of Irwin et al. Here, for each target thirty-six decoys are found for each active by matching physical properties, forming a *mimetic* set referred to as DUD-*self*. The combined set, i.e. across all targets is *drug-like* and is referred to as DUD-*all*. *Mimetic* decoys can sometimes be depressingly effective, as illustrated by Irwin et al. However, in not all cases were DUD-*self* decoys harder to distinguish than decoys from DUD-*all*. This at least suggests physical property *mimetics* are not guaranteed to provide a reality check for methods claiming to capture the physics of drug-target interaction.

Modeled decoys go one step further than *mimetics* by eschewing the concept of comparison to simple, practical, methods and instead designing directly against the method under study. As an example, suppose decoys for a docking study were chosen such that every decoy had good shape complementarity with some part of the active site, i.e. it fit well. It is widely known that basic shape complementarity is usually necessary for binding and forms a major component of most scoring functions. Such a set of decoys would them

make for a challenging test of scoring functions. But is this a good test of docking? If essentially random performance is seen an observer might decide docking is without merit, whereas an appropriate conclusion would be that complex scoring functions are useless. The inevitable desire of methods to be *seen* to be useful often prevents *modeled* decoys being chosen, even though they potentially address the most interesting scientific questions. Irwin et al. had intended their *mimetic* decoys to act more as *modeled* decoys, i.e. the aim was to make things harder for docking programs, but, as mentioned, this was not always achieved.

Although *mimetic* and *modeled* decoy selections have virtues, they can also hamper comparison between different studies. In the case of the *mimetic* approach, the definition of 'simple' evolves over time, especially as property calculations improve or change. To arrive at a set of *modeled* decoys the procedure applied must be scrupulously described, e.g. how is the protein prepared, how is the ligand protonated etc, and complete and accurate descriptions of published virtual screening procedures are rare. However, there seems no reason a consensus could not be reached by interested parties. Standard protocols could be developed, shared and used to verify results. The problems are more of will than ingenuity.

Given the above discussion, what is the appropriate decoy set to use? A *universal* set with sufficient actives to enable discrimination between methods? A *drug-like* set built from one group's definition of a corporate collection, but perhaps without general applicability? A *mimetic* set to produce physically similar decoys? Or a *modeled* set defined so as to tease out specific differences between methods, even at perception of poor performance? A suggestion by Geoff Skillman [19] provides a novel framework. Given the speed of modern computers and the cost of storage and transmittal of information, there seems no reason a retrospective study could not contain *all* decoy types, with careful labeling of individual intent. The authors can make of their data what they will, for instance by reporting performance against a subset of decoys. However, if a broader set is included in the supplementary material, others can make use of the data for potentially different purposes. One of the proposals of this paper is for modeling to move beyond the anecdotal towards the systematic. Full reporting of data is essential but a further step would be to include alternate data so that others can construct purposes beyond the original intent.

Extensive properties

In addition to intensive properties, there are the extensive properties such as how many actives, decoys and targets are used. Once again the important consideration is knowing what we want to know. If the purpose is to evaluate a single method on a single target the necessary extensive properties are quite different than for a broad study on the efficacy of several methods on many targets. We illustrate this with some basic error analysis.

The Central Limit Theory (CLT) states that the average of **M** measurements tends towards the true mean with an error proportional to $\sqrt{(\mathbf{V}/\mathbf{M})}$, where **V** is the average squared difference of a quantity from its estimated mean. Thus, the error is an intrinsic quantity, the square root of **V**, divided by an extrinsic quantity, the square root of the number of measurements. Famously, we have to take four times the number of measurements to reduce the error by a factor of two. What does this say as to the number of actives, decoys or targets that should be used to accurately measure the performance of a method? If the performance is similar no matter what actives, decoys or targets are used then the variance is small and **M** can be small. However, this is not the situation for modeling techniques applied to real systems. Instead, it is the ruling zeitgeist that 'performance may vary' [3–5].

Just how variable are virtual screening methods? Figure 1a and b illustrate the extent of the problem by presenting a reanalysis of the Warren et al. study from GSK, with eight different docking methods and our own work on the DUD dataset (DUD-self decoys) using four different virtual screening techniques. The performance metric is the AUC averaged over each dataset. The number of targets for Warren et al. is eight and for DUD forty, i.e. a

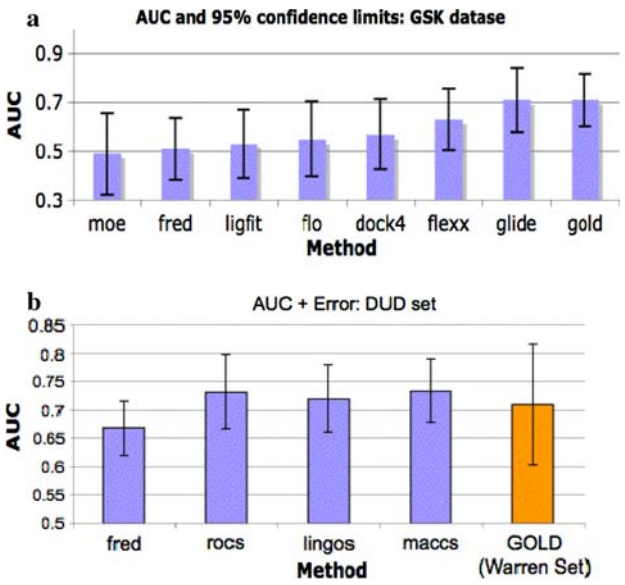

Fig. 1 (**a**) Average AUC values across docking programs in the Warren study, with 95% confidence intervals. Where programs were run in multiple modes the best average AUC was retained. (**b**) Average AUC values plus error bars across the DUD dataset for four in-house methods. Docking: FRED, Ligand-based: ROCS, 2D: Lingos and MACCS Keys [20]. Also included for comparison purposes is the average AUC for GOLD against the Warren set with associated error bars

five fold difference. As expected, the error bars, i.e. the confidence limits at 95%, are slightly more than twice as big for a method assessed against the GSK set than against DUD. In addition it is clear that although methods in the GSK test could be broadly classified as better or worse, this is subject to considerable statistical dispute. The resolving power of DUD begins to be apparent in Fig. 1b, where one can begin to put some significance to the generally held belief that ligand-based methods perform better than docking and about as equivalently as 2D methods [4, 21]. The average AUC and error bars for GOLD from the Warren study are included in for comparison purposes only. A more quantitative analysis of this data will be presented below in the section on correlation between methods.

What is the source of so much variation such that even forty targets are barely able to provide statistically supportable conclusions? In general, given a property measurement that has N *independent* sources of error, the expected error is formed from the root mean square of the individual sources of error, thus:

$$\mathrm{Err} \approx \sqrt{\left(\mathrm{Err}_1^2 + \mathrm{Err}_2^2 + \mathrm{Err}_3^2 \ldots\right)}$$

For our case we can write:

$$\mathrm{Err(method)} \approx \sqrt{\left(\mathrm{Err}_{\mathrm{targets}}^2 + \mathrm{Err}_{\mathrm{actives}}^2 + \mathrm{Err}_{\mathrm{inactives}}^2\right)}$$
$$\approx \sqrt{\left(\mathrm{Var}_{\mathrm{targets}}/\mathrm{N_t} + \mathrm{Var}_{\mathrm{actives}}/\mathrm{N_a} + \mathrm{Var}_{\mathrm{inactives}}/\mathrm{N_i}\right)}$$

The variances are intrinsic properties to 'targets', 'actives' and 'inactives'. How do we know what these variances are? One way is to boot-strap, i.e. leave out a randomly chosen fraction of the targets, or subset of actives or inactives, and measure changes in performance. Repeating this procedure many times gives a statistical sampling of the sensitivity to outliers and the number of measurements. Alternatively, in some cases the variance can be established more precisely. In the case of AUC, for example, it can be shown that for a particular target the variance for both actives and inactives can be approximated by:

$$\mathrm{Var_{active}} = \sum \left(\mathrm{p_i} - \langle \mathrm{p} \rangle\right)^2 / \mathrm{N_{active}}$$

$$\mathrm{Var_{active}} = \sum \left(\mathrm{q_j} - \langle \mathrm{q} \rangle\right)^2 / \mathrm{N_{inactive}}$$

where the sums are over all actives or inactives, $\mathrm{p_i}$ is the probability this active *i* has a higher score than an inactive, $\mathrm{q_j}$ is the probability an inactive *j* has a higher score than an active and $\langle\rangle$ represents the average of a quantity [7]. Typically, the variances of both actives and inactives are roughly equal. This leads to useful insights as to the required ratio of decoys to actives. When this ratio is 100:1, the net error is only larger by 0.5% than if we were to use

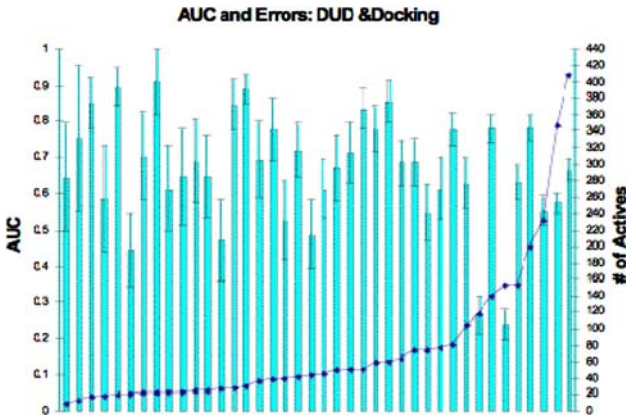

Fig. 2 AUC values ordered from left to right by number of actives for each target in the DUD set. Program used: FRED with Chemscore as the posing and scoring function. Error bars are 95% confidence intervals for each virtual screen

an infinite number of decoys. A ratio of 40:1, roughly that of the DUD-self set, yields an impact about 1%. At 10:1, this impact is about 5% and at 4:1 about 11%. Note that these effects on the error estimates, not on the actual average. What does this look like in practice? Figure 2 shows the AUC values for FRED applied to DUD (self-decoys), along with associated 95% confidence intervals for each system. Given these AUCs and contributions to variance from actives and inactives, we can directly address whether the source of the variance across targets is due to insufficient sampling of actives and decoys, or an intrinsic property of methods. By the CLT,

$$\mathrm{Err(AUC)} \approx \sqrt{\left(\mathrm{Var_{Obs}}/\mathrm{N_t}\right)}$$

where

$$\mathrm{Var_{Obs}} = \sum \left(\mathrm{AUC_i} - \langle \mathrm{AUC} \rangle\right)^2 / \mathrm{N_t}$$

Therefore

$$\mathrm{Var_{targets}} = \mathrm{N_t} * \{(\mathrm{Var_{Obs}}/\mathrm{N_t}) - (\mathrm{Var_{actives}}/\mathrm{N_a}) - (\mathrm{Var_{decoys}}/\mathrm{N_i})\}$$

Table 1 shows contributions to the square of the average error in the mean AUC across DUD for our four methods

Table 1 The contribution to observed variance from actives, decoys and targets over the DUD dataset (DUD-self decoys)

Method	$\langle \mathrm{Err}^2 \rangle$ – Decoys	$\langle \mathrm{Err}^2 \rangle$ – Actives	$\langle \mathrm{Err}^2 \rangle$ – Observed	Est. $\langle \mathrm{Err}^2 \rangle$ – Targets
FRED	0.000048	0.0020	0.023	0.021
ROCS	0.000025	0.0022	0.041	0.039
MACCS	0.00004	0.0017	0.030	0.028
LINGOS	0.000039	0.0017	0.035	0.033

The estimated error (squared) from the variation between targets is estimated from the observed variance and corresponds to that which would be obtained if the number of actives and inactives were infinite

calculated in this manner. First, as expected the contribution from the inactives is about forty times less than that of the actives (because the intrinsic variances are similar and there are thirty-six times more decoys than actives in DUD-*self*). Secondly, it is clear that the errors due to target variability is roughly ten times higher than that due to actives. As independent errors add as squares, this implies only about 5% of the observed confidence limit on the target-averaged AUC is due to the finite number of actives. This leaves 95% of the 95% confidence limit due entirely to the considerable variation from target to target. Comparing DUD-*self* to the careful evaluation of McGaughey et al. from Merck, and Warren et al. from GSK, the latter have roughly four times more actives and four times less systems, i.e. they are more careful studies of particular systems (error bars are 50% smaller per *system*) but substantially less useful for general conclusions (error bars are twice as big per *method*). Remarkably, even if the number of actives in each DUD-*self* set were reduced by a factor of ten, causing a threefold increase in the error estimation per target, the net error of averages over all systems is only increased by about 30%.

The conclusions from this analysis of extensive properties are straightforward.

(1) When calculating the properties of a single system the number of actives is fairly important, but the number of inactives does not have to be substantially larger. A ratio of decoys to actives of 4:1 only has an error 11% higher than the limiting value from an infinite number of inactives. It would be more useful to include sets of inactives designed for different purposes than to attempt to 'overwhelm' the actives with decoys.
(2) If the purpose is to test a method against other methods with 95% confidence then the number of systems required is very large, much larger than even DUD. In our analysis the contributions to the variance from a limited numbers of actives is almost insignificant compared to the target-to-target variation. For example, it would take over 100 test systems to tease apart the difference between the ligand-based method ROCS and the docking program FRED with 95% confidence. (See below.)
(3) The variance between systems is such that the number of actives per target does not need to be very large, perhaps even as low as ten. As such, suggestions to only include representatives of chemical classes, e.g. see Good and Oprea [22], may be statistically quite valid.

Correlations

A key assumption underlying much statistical analysis is the independence of samples, for instance that any two measurements are uncorrelated. This is often a good assumption but it is not hard to find counterexamples. Consider the case where a decoy is included twice. We have gained no new information. Yet N_i, the number of decoys, has increased and so the error goes down. Clearly the error has not really been reduced. Instead, the decoys are no longer *independent*. In the line-up analogy, this would correspond to including identical twins as fillers. While this is an unlikely mistake, what about two individuals that look very similar? How independent are two molecules and what does this even mean? The temptation is to reach for the familiar chemical definition, i.e. 2D similarity. Even though there is no rigorous definition of chemical similarity, it is a major concern in selecting active populations from chemically related (congeneric) series. Methods that either rely on chemical similarity, or are heavily influenced by it, may not be making independent assessments. Clearly 2D methods fall into this category, and sometimes ligand-based 3D methods. Ideally, methods that use protein structure, such as docking, ought to be less affected, but this is far from proven. Suggestions as to how to improve matters include reducing the set of active to a smaller set of representative structures [22], or giving more weight to the first compound discovered in a series [23]. (Application of similar protocols to decoys is seldom discussed, perhaps because they are less likely to be congeneric). These are practical suggestions derived from knowing the nature of drug discovery. There is also a general approach that eschews the particulars. Two compounds are considered *operationally* dependent if their rankings under different tests are correlated. For instance, a method that had a size-bias would tend to rank a pair of molecules of comparable extent similarly, no matter what the target. Even without 2D similarity, this implies a less than perfect independence. Imagine a method where all the decoys are of one size and all the actives another. No matter what the actual number of actives and inactives, there are essentially only two molecules, an active and an inactive, and our ability to extract meaningful statistics is severely compromised. Note that the *operational* part of this definition depends on the nature of the method, i.e. dependence is conditional on the nature of the procedure investigated.

Similar situations occur in assessments of genetic linkage. The degree of dependence amongst a set of markers is evaluated by constructing a matrix where the entries are the correlation of phenotypic scores between any two markers. The eigenvalue spectrum of this matrix is then used to assess the actual number of degrees of freedom [24]. Crucially, though, correlation can only be estimated by knowing the behavior of a pair of samples/compounds over many tests/targets. At first glance this suggests that the same set of decoys should be used across all targets in a

study. If decoy selection is *universal* or *drug-like* this is typically a part of the design, i.e. the decoy set is reused. However, *mimetic* and *modeled* decoys rely on the nature of the actives, which will vary from target to target. This might seem a dilemma, i.e. we want to reuse decoys so we know if they are correlated but we cannot reuse them because one set of decoys may be completely inappropriate for another target. Here Skillman's suggestion is again useful, i.e. there is nothing to stop us including the decoy set for one target in the virtual screen of a second target *not as decoys* but rather to gain information on operational independence. To distinguish the role of decoys from one system applied to a second system for purposes other than assessing performance, we suggest the term 'latent' since these secondary decoys should be hidden from the calculation of performance metrics. In addition to the concept of *latent* decoys, *latent* actives can be used to measure operational independence independent of 2D similarity. Another possibility is to use the actives from one system as *explicit* decoys in other systems. For instance, in the Warren et al. study each set of actives also formed the decoy set for the other targets. The intent was just to produce a set of *drug-like* decoys, but it fortuitously provides the most compact form for a rigorous estimation of decoy/active independence. This work will be presented elsewhere, along with an elaboration of the techniques for assessing operational correlation.

We turn now to the question of target independence. As we have seen, and is widely appreciated, the variation of performance of methods from target to target is considerable. But do certain targets, or classes of targets, behave similarly for certain methods? For instance, one would expect that a docking program parameterized against certain binding motifs would perform similarly across all targets with this motif, if only because of inductive bias. Or one might assume that isoforms of a target are sufficiently akin that docking methods would perform similarly on each. Fortuitously, the Warren study provides one such example in the inclusion of PDFE and PDFS. Figure 3 shows the difference in performance of methods used on isoforms versus the average difference between all other pairs of targets. It is clear there is less variance between the isoforms than unrelated targets. If this were a generalizable result it would have two consequences. On the positive it would mean that methods could be quantified for certain types of problems without requiring large numbers of targets, i.e. because the variance is smaller. On the negative, it would mean that just as considerations need to be made for the true statistical power of closely related actives, or inactives, similar considerations need to be made for targets, increasing the number of targets required to either discern general differences between methods or to reliably gauge progress of

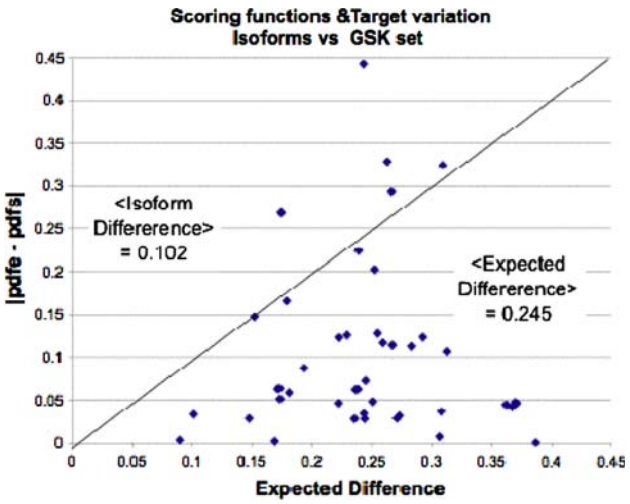

Fig. 3 Docking performance against the two isoforms in the Warren study (PDFS and PDFE), compared to the averaged difference over all other pairs of targets

a single approach. And, as we have seen, to measure global performance on *independent* systems already requires sampling beyond common practice.

The PDFE/S example is a single data point. It is entirely possible that the variation between targets of similar class, or highly conserved isoforms, or even different forms of the same protein structure is not small and that calculating mean properties is still formidable. One might imagine this is a well-researched area, but this appears not to be the case. Retrospective experiments are designed, however poorly, to give an estimate of to what to expect for the next *new* target and so targets are chosen to be diverse. Software would seldom be used in default, out-of-the-box, mode when there is considerable domain knowledge, i.e. within a set of closely related targets. Hence, the question of method variance over similar systems appears to have been overlooked.

The final aspect of independence is correlation between methods. Suppose we have method A and method B, each tested on the same set of targets with the same set of actives and decoys and the results show A is consistently slightly better than B. How can we prove this difference is statistically significant? At first glance this would seem difficult. As illustrated in Fig. 1a and b, the variance of any one computational method over a set of targets in invariably large. As such, the error bars on an average property, such as an AUC or enrichment, are big. So although the average behavior of method A is slightly better than B, this difference would appear statistically insignificant. However, if the test systems are indeed identical this is not the correct assessment. Instead, the CLT is applied to the set of measured *differences* between methods, e.g. for an AUC example the variance becomes:

$$\mathrm{Var_{diff}} = \sum((\mathrm{A_i} - \mathrm{B_i}) - (\langle \mathrm{A} \rangle - \langle \mathrm{B} \rangle))^2/\mathrm{N_t}$$

$$\mathrm{Err(diff)} \approx \sqrt{(\mathrm{Var_{diff}}/\mathrm{N_t})}$$

The formula for $\mathrm{Var_{diff}}$ can be rewritten as:

$$\mathrm{Var_{diff}} = \mathrm{Var_A} + \mathrm{Var_B} - 2{*}\mathrm{Corr(A,B)}$$

$$\mathrm{Corr(A,B)} = \sum(\mathrm{A_i} - \langle \mathrm{A} \rangle)(\mathrm{B_i} - \langle \mathrm{B} \rangle)/\mathrm{N_t}$$

Here Corr(A,B) is a measure of the correlation between methods A and B and is related to the Pearson correlation coefficient, thus:

$$\mathrm{Pear(A,B)} = \mathrm{Corr(A,B)}/\sqrt{(\mathrm{Var_A}{*}\mathrm{Var_B})}$$

If the tests of methods A and B are independent then the correlation is typically assumed zero and the variance is just the sum of the variances of both methods. If variances were roughly equal, this would typically be a *joint* error bar $\sqrt{2}$ larger than the individual error bars. (This also means the common practice of evaluating whether two methods are statistically different by whether their individual error bars overlap is generally incorrect.)

However, if the tests applied to methods A and B are identical, correlation needs to be explicitly included. In the case of A always slightly better than B, we need to assess whether the mean difference is larger than the joint confidence limits generated from the variance of the difference between A and B. In fact, if A is better than B by a constant amount we are guaranteed statistical significance because the variance of the *difference* is zero. In general, methods tend to be positively correlated so that the joint confidence limits are lower than from independent measurements.

Confidence limits are convenient because they give a visual estimate of the possible range of true values, typically at a 95% level of confidence. However, joint confidence limits are less graphical as they pertain to pairs of methods. In addition, there is considerable concern in other fields as to the arbitrariness of the 95% value. The origin of this number is R. A. Fisher, whose work in the 1920s still dominates much of the field of practical statistics. Fisher, primarily an agriculturalist, observed a 10% increase in cabbage yield when manure was used. He also observed that only one in twenty plots without manure showed a yield greater than 10%, and so the 95% cut-off was born! From this, and because of the general utility of Fisher's work on experimental design, a *p*-value of 0.05 dominates many fields, in particular clinical trials. Essentially, a *p*-value is the probability a null hypothesis can be rejected. In our case the null hypothesis would be that method B is in fact better than method A, despite average values suggesting the reverse. Under the assumption the difference in performance between two methods is as predicted by the CLT (i.e. Gaussian), we can assign a (*p*-)value to the probability one method is better only because of random chance. We do this by calculating the area under the normal form for which one appears better than the other. The mathematics of this is shown below:

$$p = (1 - erf(\langle \mathrm{A} - \mathrm{B} \rangle {*} \sqrt{(0.5{*}\mathrm{N_t}/\mathrm{Var_{diff}})}))/2$$

where $\langle$A–B$\rangle$ is the average difference between the methods, the other variables are as defined as above and *erf* is the inverse cumulative Gaussian, or *error*, function. If $\mathrm{N_t}$ is small, i.e. less than twenty, then a slightly different functional form is more accurate (i.e. from the Student t-test) because the CLT only applies in the limit of large N. For practical purposes the difference can be ignored. The smaller *p*, the stronger the case for A being better than B. Note, we are not proving how much better A is than B. The best *estimate* of A's superiority is still the mean difference of whatever property we are measuring. Rather the *p*-value refers to the dichotomous question, is A better than B?

Table 2 illustrates the above concepts for the four methods listed in Fig. 1b applied to DUD (DUD-*self* decoys). The diagonal entries are the mean values of the AUC of each method, followed by the associated 95% confidence limits. The upper triangle of the table contains

Table 2 Statistical measures necessary to accurately assess the relative performance of methods, here applied to the DUD data set (DUD-self decoys)

Method	FRED	ROCS	MACCS	LINGOS
FRED	0.684/ 0.043	0.11/0.08/ 0.07	0.1/0.07/0.06	0.1/0.07/0.065
ROCS	0.17/0.09	0.732/0.065	0.12/0.085/ 0.05	0.125/0.09/0.05
MACCS	0.03/0.05	0.70/0.47	0.734/0.055	0.115/0.08/ 0.055
LINGOS	0.19/0.14	0.65/0.36	0.54/0.31	0.72/0.061

Diagonal terms: average AUC/95% confidence limits. Upper triangle terms: naïve joint confidence limits/joint confidence limits assuming different tests/joint confidence limits assuming same tests and accounting for correlation. Lower triangle terms: Pearson correlation coefficients/ *p*-values that a method has a higher mean AUC by random chance

the naïve joint confidence limits, i.e. by summing individual confidence limits, the joint confidence limits assuming independent tests and the joint confidence limits properly calculated with correlation. The lower diagonal contains the Pearson correlation coefficient for each pair of methods, followed by the *p*-value for the hypothesis that the better method is so only by chance. So, for example, the probability that the 2D and ligand based methods are better than the docking program FRED by chance are around the 10% level, whereas the differences between these three methods themselves is close to 50:50, i.e. in this example we can distinguish ligand-based methods from a docking protocol but not one ligand-based method from another. This is most likely because DUD is not designed to test ligand-based retrieval containing, as it does, many 2D similar actives. Several others have made this point, including the curators themselves [14, 18, 22]. Examples from other fields of how to apply these procedures to differences in AUC can be found in Hanley and McNeil [9].

One of the advantages of *p*-values is the statistical machinery, again developed by Fisher [25], for combining values from different studies. A classic example is the effects of tobacco. It was not one study that convinced the medical profession, but a series of studies and the facility to combine the results that lead to the overwhelming conclusions as to the health risks of smoking. In statistics this is referred to as "Meta-Analysis". Despite this, and the wide application of *p*-values in other fields, they are largely absent from modeling, with a few exceptions [5].

In conclusion, correlation is important in all aspects of virtual screening, but perhaps most important and most easily corrected for in the comparison of pairs of methods. Neglecting the effects of correlation between tests is a frequent problem even in clinical studies [2] and to our knowledge has not been properly applied to comparing methods in virtual screening.

Metrics

Given an experimental design, what quantities should be measured to assess performance? The question suggests a sequential process, i.e. design the experiment and then measure something, whereas good design takes into account what is going to be measured, in particular to what accuracy. However, assuming a given design, how do we extract useful information? In this section we consider what should be measured and why. This is not a quandary specific to virtual screening, in fact is it universal to all prediction exercises. This very commonality can help suggest worthwhile approaches. It also suggests that measures constructed specifically, even uniquely, for chemical virtual screening should be held to a similar standard to those prevalent in the wider world. Is virtual screening really so different from, say, Internet page ranking? In particular, we will consider the issue of 'early' behavior, i.e. measures that reward ranking some active compounds near the top of a list. By considering real-world financial parameters we ask whether 'early' behavior is even *necessarily* to be prized. By looking at a large number of virtual screens, we will ask whether such 'early' measures are necessary and whether they can be predicted from more fundamental and well-understood properties. Finally, the application of accurate error bounds will be shown to suggest at least one way of quantifying the advantage an expert brings to well-studied systems.

Properties of virtual screening metrics

A long list of metrics has been applied to virtual screening. What makes for a good metric? The unfortunate answer with some papers is "any metric that will make my method look good". And if no known metric will suffice, then simply make a one up. This is a typical indicator of an under-regulated and under-developed field. Computer manufacturers used to habitually make up their own measures for the latest processor or operating system, leading to much confusion and annoyance. As a consequence, in 1988 SPEC (Standard Performance Evaluation Corporation) was formed and SPEC Marks became the standard benchmark of anything worth measuring. SPEC had a simple philosophy: "The key realization was that an ounce of honest data was worth more than a pound of marketing hype" [26]. SPEC Marks have evolved over time to now cover CPU, graphics, Java, mail servers, file servers, parallel performance, high performance computing and other aspects. In other words, SPEC is not a single measure because not all users want the same thing, but this does not mean manufacturers can create their own metrics. Rather SPEC is an umbrella organization for a set of open and diverse groups that consider, ratify and develop benchmarks. In this spirit, this section will concentrate on what ought to be general characteristics of a good metric rather than all prevalent quantities.

In a somewhat circular manner, one of the first characteristics of a good measure is that everyone uses it. Clearly one of the problems with a field with diverse measures is incomparability, the "apples and oranges" problem. The most straightforward solution is not imposition of a particular standard but full disclosure of all data. The authors of a study may want to present enrichment at 5%, but if the data is freely available others may calculate the enrichment at 1% or 13% or whatever they wish. This would inevitably lead to standardization as independent parties harvest data from many sources, publishing larger and larger studies on

the advantages and disadvantages of different methods and measures. This would provide another example of meta-analysis described above. Sometimes a valid excuse against disclosure is that compounds or targets are proprietary. However, just providing lists of actives and inactives in rank order with unique, but not necessarily identifying, tags is enough to calculate most of the metrics for a particular virtual screen. Currently the field of modeling lacks even an agreed upon format for the exchange of such rarely available information.

However, if we are going to report a statistic what properties should it have? From considering measures that have become standard in other fields, what characteristics define a good measure? We suggest the following short list:

(i) Independence to extensive variables
(ii) Robustness
(iii) Straightforward assessment of error bounds
(iv) No free parameters
(v) Easily understood and interpretable

Take for example the very popular "enrichment" measure. Everyone understands the concept of enrichment: swirl a pan of water and gravel from the Klondike river in 1896 in just the right way and you ended up with mostly gold. In virtual screening you look at the top few percent and see whether there are more actives than you would expect by chance. As a mathematical formula this is typically presented as:

$$EF(X\%) = (100/X) * (\text{Fraction of Actives Found})$$

The problem with this measure is that the enrichment becomes smaller if fewer inactives are initially present. Imagine panning for gold with all the sand removed. There would still be the same gold in the pan, along with maybe some pebbles and small rocks, but the eventual relative *improvement* after 'panning', i.e. 'enrichment', is reduced. The problem is that the (Fraction of Actives Found) contradicts requirement (i), i.e. is a function of extensive quantities, the number of actives and inactives. This means that enrichment is not actually a measure of a method; it is a measure of a method and a *particular* experiment. If the ratio of inactives to actives becomes very large it is assumed this problem disappears, i.e. that the limiting behavior obeys (i). This is not true if the enrichment at a given percent is large, i.e. at precisely the points of most interest. Also, enrichment does not meet requirements (ii). At a small enough percentage the enrichment becomes an unstable function of the exact positions of actives in a list. There is also no agreed upon percentage, making this an adjustable parameter (often abused). Finally, other than by bootstrapping, the author knows of no simple assessment of error bounds. However, it is an intuitive measure, easily understood, passing rule (v), and so almost uniquely to this field is the most common metric reported.

Some have been aware of the lack of robustness of enrichment and proposed metrics that average over all percentages with weighting schemes. Before we consider these measures we point out that a simple fix to the common variant of enrichment is to make the enrichment refer to the fraction of inactives, not to the fraction of all compounds. This simple change makes the enrichment independent of extensive quantities, more robust, accessible to analytic error approximation [27] and yet suffers only a slight reduction in interpretability. Only a few, such as Jain [14] have used this alternate form. Perhaps the only important failing of this measure is that it lacks a specific name. For the purpose of this paper it will be referred to as the ROC enrichment to distinguish it from the widely abused variety.

ROC enrichment has better properties because is related to an even better metric, the AUC, defined as the area under a ROC curve. A ROC curve is simply a plot of the discovered active fraction versus the discovered inactive fraction. (Each point on the ROC curve can be translated to a ROC enrichment by dividing by the fraction of inactives). The AUC is the average of this property over all inactive fractions. Many excellent treatises can be found [6–9] and it has become a standard for classification performance in many disciplines (medical diagnostics, radiology, clinical testing, criminology, machine learning, data mining, psychology and economics to name a few). It satisfies all of the criteria listed above as a metric, including (v), ease of interpretation. The AUC is simply the probability that a randomly chosen active has a higher score than a randomly chosen inactive. The main complaint against the AUC is that is does not directly answer the questions some want posed, i.e. the performance of a method in the top few percent. This is akin to complaining that SPEC Marks do not do a good job of evaluating mobile phone processors; a fair complaint perhaps but hardly justifying creating a new benchmark without the strengths of existing standards. The AUC ought to at least be held as such a standard against which new measures are judged.

Early performance in virtual screening

Figure 4 illustrates the supposed limitations of the AUC as a measure of performance. The graph shows two ROC curves, each with an AUC of exactly 0.5. Overall this means that an active is as equally likely to out-rank an inactive than the other way around. However, clearly in the case of the solid line a certain fraction of actives is being scored significantly higher than most inactives, while another fraction is being scored worse, i.e. it is only the

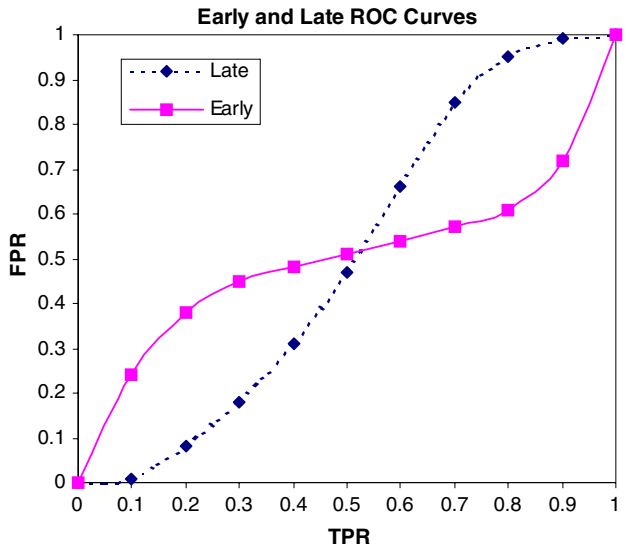

Fig. 4 Example ROC plots for "early" and "late" methods

average behavior that appears even-handed. Similarly, the dashed curve illustrates the case where the actives are all scored better then a certain fraction of the inactives but worse than another fraction. The solid and dashed curves are instances of bimodal score distributions for the actives and inactives respectively. Since the goal of a virtual screen is to save us the trouble of actually screening all the compounds it is entirely reasonable to prefer good 'early' behavior. And yet the AUC does not distinguish between such curves and so, it is claimed, is not appropriate.

It is against this backdrop that metrics such as RIE [28] and BedROC [29] were developed. In both cases the essential idea is to give early rankings of actives more weight than late rankings. In RIE/BedROC actives are given a weight depending on their position in the list using an exponential function running from 1.0 for the top ranked compound to a number typically close to zero for the lowest ranked. The exponential factor, beta, determines how fast this exponential dies away from the top rank and controls how much the RIE/BedROC parameter focuses on the top of the list. The larger beta the greater the early focus. In RIE the sum of active weights is normalized by the RIE of a random distribution of actives. In BedROC normalization is by the maximum dynamic range, i.e. the score with all the actives ranked at the top minus the score with all the actives ranked at the bottom. In addition, by first subtracting the score of the worst-case scenario, BedROC has the elegant property of running from 0.0 to 1.0. The rational behind using these approaches is to give precedence to actives ranking early but not to fall into the trap of choosing a single enrichment value, i.e. be more robust to perturbations in the rank ordering. Impressively, Truchon and Bayly also derive analytic estimations of the error bounds for BedROC and give some suggested values for the beta parameter. Some incorrect statements regarding the AUC mar their work, for instance that random scores do not give an AUC of 0.5, and that the AUC is dependent on the number of actives and inactives, but overall the work is an interesting attempt to answer a perceived need. Applied to the examples in Fig. 4, the ratio of the BedROC score of the solid line to the dashed is about two for a beta of ten and about ten for a beta of twenty.

So does BedROC or RIE qualify as a good metric for virtual screening? Comparing against the five criteria listed above, both are more robust than enrichment, and the error protocols for BedROC satisfies criteria (iii). RIE suffers from having an ill-defined numerical interpretation (i.e. how good is an RIE of 5.34?). BedROC attempts to overcome this by scaling between 0.0 and 1.0, but does this qualify as being understandable? There is no absolute, interpretable meaning to a BedROC (or RIE) number, only a relative meaning when ranking methods.

Unfortunately, neither BedROC nor RIE satisfy criteria (i) or (iv), i.e. both are dependent on extrinsic variables and have an adjustable parameter, the exponential factor beta. The former, as we have seen, means that scores can only be compared in the limiting case of an excess of inactives and, as in the case of enrichment, this excess has to persist even when the enrichment of actives is very high, i.e. it is exactly when the actives are predominantly at the top of the list that both BedROC and RIE (and enrichment) are most sensitive to the total number of inactives. Interestingly, it would be possible to reformulate both metrics to avoid this problem. Just as ROC enrichment is a better metric than enrichment, an exponential weighting across the ROC curve, rather than to the individual rankings of actives amongst inactives, would remove the sensitivity of these measures to extensive properties. However, there would still remains the issue of the arbitrariness of the exponential factor beta. Just as with enrichment thresholds there is nothing intrinsically wrong with the freedom to select a threshold that is of interest to the particular research group. Some companies might have the facility to physically screen ten percent of their database, another only one percent. However, as a characteristic of a method, or a class of methods, it is a disadvantage. A proponent of a method has a free parameter with which to make their method look favorable or even just less unfavorable. In the example in Fig. 4, a factor of two in BedROC between the methods (beta = 5.0) does not sound anywhere near as bad as a factor of ten (beta = 20.0).

Cost structures of virtual screening

There is no fundamental meaning to BedROC or RIE. Neither gets to the real heart of why the solid curve in

Fig. 4 represents a better method than the dashed curve. In what follows we will argue that this can only be stated with respect to a set of assigned costs, assumed but never stated. We start by noting that the current focus on early enrichment is actually a change in values for the industry. This author recalls conversations in the mid-1990s wherein the concept of missing *any* potential lead compound was deemed unacceptable. By contrast, a preference for early behavior implies it is acceptable to miss a significant fraction of potential actives in favor of finding a few good leads. There is merit in this approach. Often a chemistry team can only follow up on a small number of leads. High throughput screens can take several months to design and bring on-line, time chemists could use to explore initial leads from a smaller focused set designed by a virtual screen [30]. What are not made explicit in this shift are the costs of the four components of any virtual screen: true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN). Not wanting to miss anything is equivalent to assigning an infinite cost to a false negative. This was never sensible, but reflected a 'lottery' mentality prevalent at the time. The reality is that virtual screening never finds drugs; at best it can find things that might, after considerable effort, become drugs. In addition, the attrition rate at many stages in the drug design process means any lead-like compound is at best a bet that will often fail, costing many millions of dollars. A lottery ticket is potentially worth millions; the expected value, i.e. averaged over all contingencies, is usually less than the cost of the ticket. The assumption behind virtual screening is that the value of a true positive similarly averaged is worth the cost of computers and modelers. This is an unproven conjecture.

The assignment of a cost structure to the components of a screen is common in the field of medical diagnostics. Here the costs can be estimated with some reliability. A true positive represents the successful diagnosis of a condition that will save money when treated. A false positive means further, costly, tests will need to be performed. A false negative might cost a lot if a more severe condition develops. Finally, a true negative can be set to the cost of the test or a small saving if compared to a more expensive test. If these values are assigned to each "truth table" component (TP, FP, TN, FN), a ROC curve can be transformed into a cost curve. A small caveat is that the ROC curve deals with true and false positive *rates* and so to transform to real costs the expected number of actives and inactives is required, or at least the ratio of the two. Suppose we apply a cost structure to Fig. 4 as follows:

(i) $TP = 8.0$
(ii) $FN = -2.0$
(iii) $FP = -0.16$
(iv) $TN = 0.02$

Positive numbers are favorable, for instance the cost assigned to a true negative is the saving from not physically screening a compound. At any point in the curve the cost of progressing with all compounds higher than a given threshold t depends on the False Positive Rate (FPR) and True Positive Rare (TPR):

$$\begin{aligned}\text{Cost}(t) = {} & \text{TPR} * N_a * (8.0) + (1 - \text{TPR}) * N_a * (-2.0) \\ & + \text{FPR} * N_i * (-0.16) + (1 - \text{FPR}) * N_i * (0.02)\end{aligned}$$

Let us assume $N_a/N_i = 1/100$, then:

$$\begin{aligned}\text{Cost}(t)/Ni = {} & (\text{TPR} * (8.0 + 2.0) - 2.0)/100 \\ & - \text{FPR} * (0.16 + 0.02) + 0.02 \\ = {} & 0.10 * \text{TPR} - 0.18 * \text{FPR}\end{aligned}$$

This is a simple linear scaling of the graphs in Fig. 4, as shown in Fig. 5a. As expected, the best approach is to take the method with early performance over the later

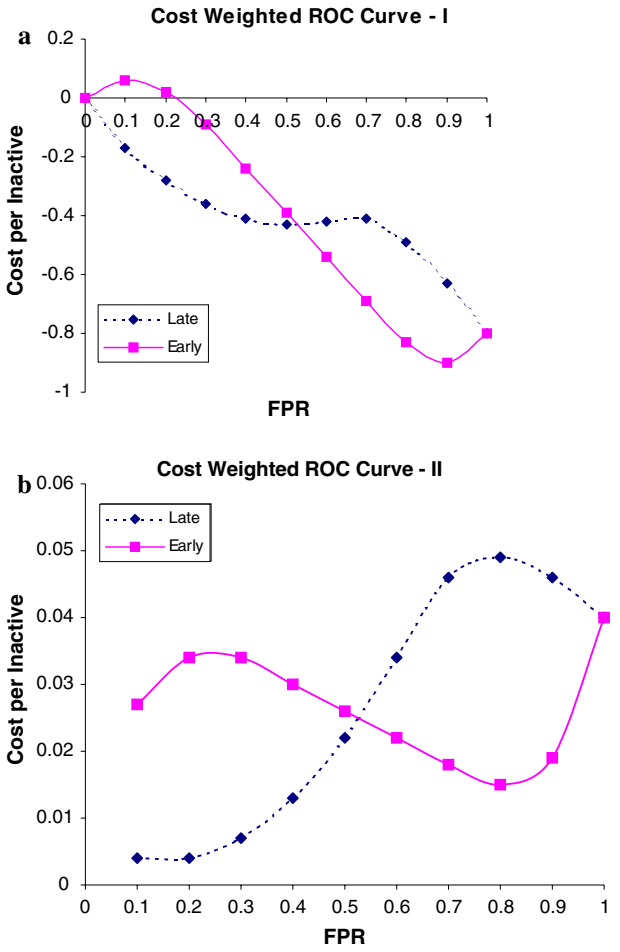

Fig. 5 (**a**) Cost weighted versions of the curves in Figure 4 as per the first description in the text. (**b**) Cost weighted versions of the curves in Fig. 4 as per the second description in the text

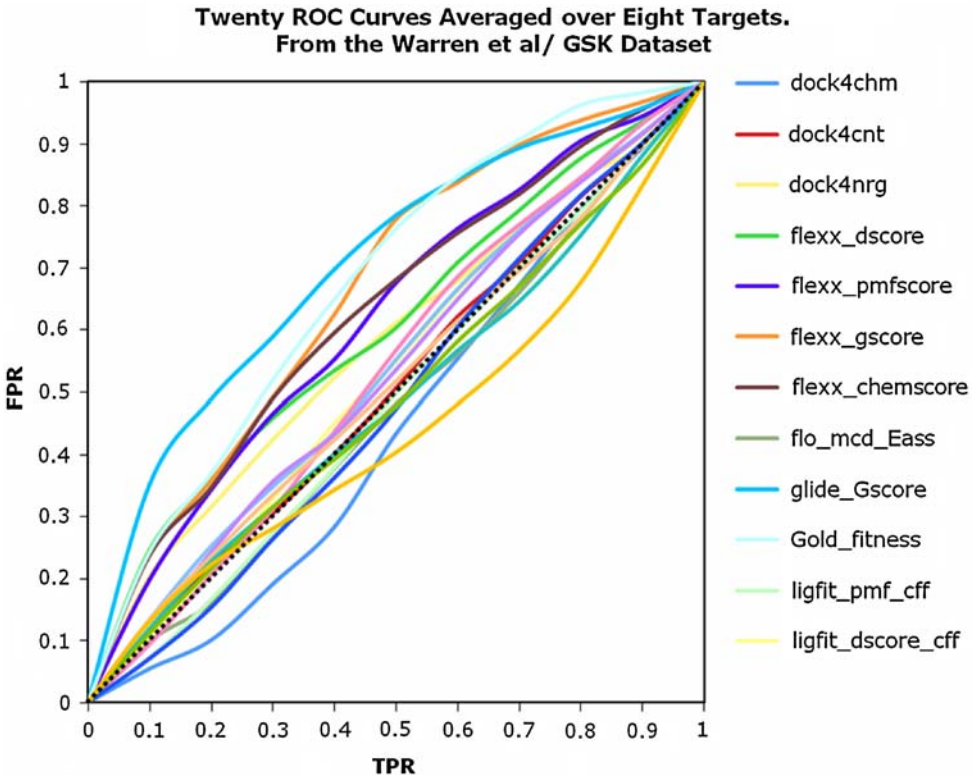

Fig. 6 Averaged ROC curves for twenty methods in the Warren study for which scores for all eight targets where available. Programs and scoring functions listed to the right of the graph

performance. Notice that the late performing method is never cost effective and even the early method is only cost effective for a narrow range of rankings.

Now consider a slightly difference weighting:

(i) $TP = 8.0$
(ii) $FN = -2.0$
(iii) $FP = -0.04$
(iv) $TN = 0.03$

$$\text{Cost(t)}/N_i = 0.1 * TPR - 0.07 * FPR + 0.01$$

Figure 5b illustrates the effect of these new weightings. By reducing the cost of a false positive by 75%, i.e. to around the savings of a true negative, both methods are always cost effective. Furthermore, although the early method has a clear maximum at around 20% of the database, it is actually worth physically screening about 75% of the database.

These examples are obviously only illustrative, but the point they make is real. Early enrichment is important only because of an assumed cost structure. Clearly much more complicated models could be constructed, possibly with real data, as with medical tests. However, to the author's knowledge this has never been published, presented or even discussed within the industry. It is an assumption that early enrichment is better. Likewise, it is also an assumption that virtual screening itself is a productive exercise compared to physical screening.

Averaged properties of virtual screening

Suppose the cost structure of virtual screening does favors early enrichment. Can we at least say metrics such as RIE and BedROC, perhaps reformulated to be independent of

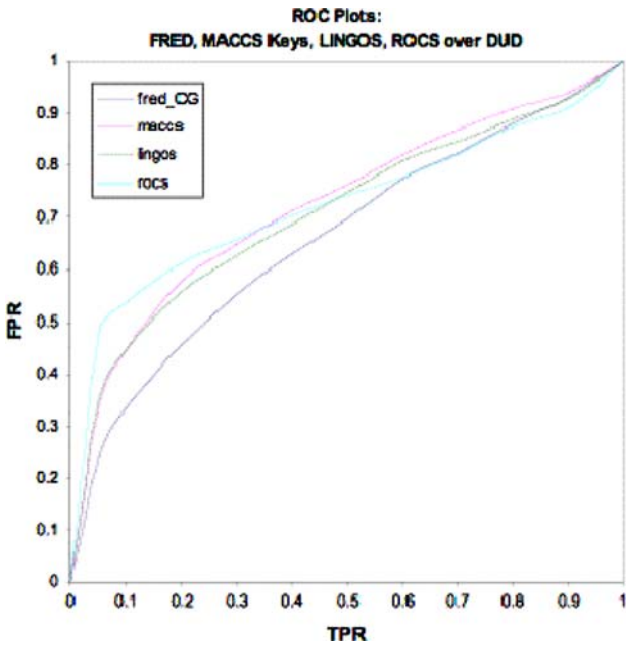

Fig. 7 Average ROC curves for FRED, ROCS, MACCS keys and LINGOS over DUD, with DUD-self decoys. FRED was run with the ChemGauss3 scoring function

extensive variables, are superior to AUC? If the early behavior shown in Fig. 4 were indeed repeated from system to system then clearly this would be the case. In Fig. 6 we show data from Warren et al. for twenty docking procedures averaged over all eight targets in the study. Examination of these curves reveals nothing that resembles the biphasic nature anticipated from Fig. 4. Individual curves might occasionally suggest biphasic behavior but there is little evidence for this in *target averaged* ROC curves. Figure 7 shows similar curves for the four methods in Fig. 1b averaged across the DUD set. The curves in Fig. 7 are smoother because the averaging across forty targets in DUD is more extensive than the eight from GSK and show even less evidence of biphasic behavior. There are two possibilities for these observations. Either the individual curves are not biphasic or the averaging dilutes this characteristic. It is possible to imagine a technique that would rank one type of actives well, perhaps hydrophobic moieties, but ranks others badly, e.g. hydrophilic ones, but that the proportions of each set differ target to target such that the total behavior appears monophasic. To see if this might be the case we examined two hundred and seventy virtual screens from the Warren study, looking for a divergence between BedROC, with exponential parameter 5.0, and AUC, i.e. an abnormally low AUC and a high BedROC, although possibly the reverse. The results are shown in Fig. 8. Clearly there is a strong correlation between BedROC and AUC. Similar correlations were also seen when higher exponential factors were employed and suggest no evidence for biphasic behavior. A better AUC

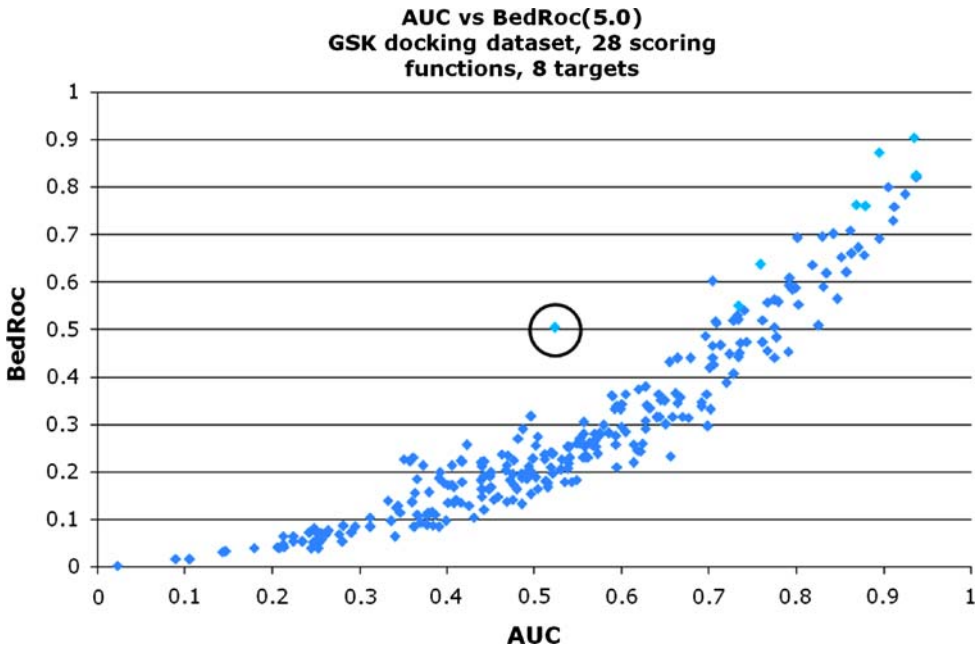

Fig. 8 BedROC scores with an exponential factor of 5.0 versus the AUC for 270 virtual screens from the Warren study

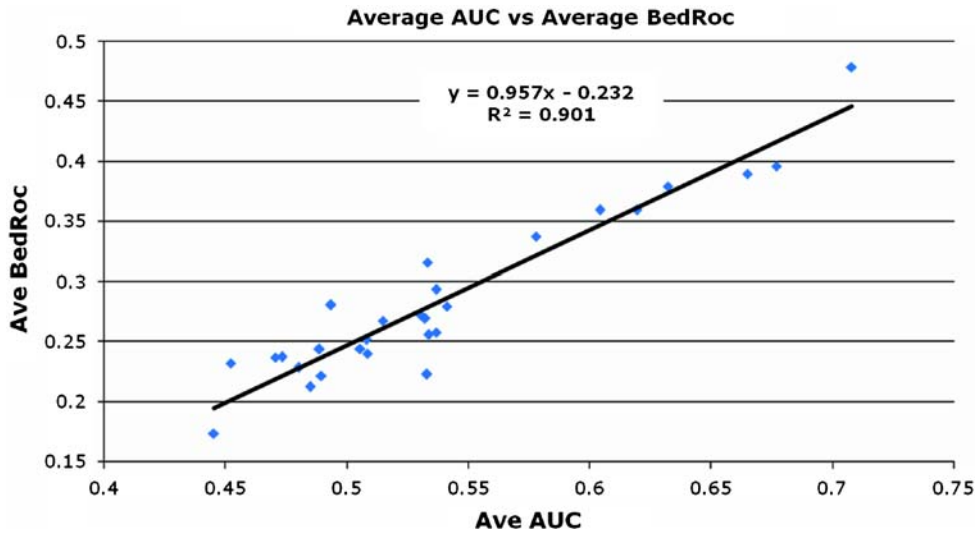

Fig. 9 The average AUC for each method run against all eight targets in the Warren study versus the averaged BedROC score for each such method

naturally leads to a better BedROC value and does so with surprisingly little variation. It might still be possible that one of the methods has average biphasic behavior, for instance if all the BedROC/AUC points for a method trended higher. Figure 9 shows this is not the case. Here, the average AUC per method is compared to the average BedROC value for that method. In addition, the correlation is stronger the better the BedROC value, so that methods that have a good AUC will also have a (strongly correlated) good BedROC score.

There is one exception to monophasicity, shown circled in Fig. 9. This point lies outside the 95% confidence limits of both AUC and BedROC. Its BedROC score is significantly higher than expected and its AUC is around 0.5, i.e. random ranking. The target protein represented by this point is PPAR-δ and the method is MVP, a program developed in-house at GSK by Mill Lambert. In conversation Lambert freely admitted that not only did he have extensive knowledge of this target, he used all of this information to tune MVP. Unfortunately, because of certain aspects of the target he could only select one of three chemical classes for this 'hands-on' treatment at a cost to the two other classes. Hence, MVP had to be biphasic. It seems interesting that out of two hundred and seventy virtual screens the only outlier from the BedROC-AUC correspondence is an example of expert intervention. An unintended consequence of this study might be a method to spot and quantify expert contributions to virtual screening, i.e. by comparing early behavior, either with BedROC or other metrics, to that predicted from the fundamental measure of AUC.

Conclusions

In this study we have considered several aspects of experimental design and performance metrics for virtual screening. There is clearly interest in doing things the right way, not least because of a popular, if unproven, belief that virtual screening saves the pharmaceutical industry money. As with many relatively young endeavors, molecular modeling has been long on promises and short on standards, and it is standards that ultimately deliver the proof that our field is useful. For many years the computer industry suffered from similar growing pains. Not only were there few, if any, reliable comparison metrics for different processors, operating systems, compilers and so forth, the proposed benefits of computers were more assumed than quantified. These days no one doubts the impact of the computing revolution. It is to be hoped that a similar statement can one day be made for molecule modeling. It is with this in mind that the following observations and recommendations are made.

On the issue of experimental design we propose:

(i) Decoy selection needs to be properly labeled as to intent to facilitate inter-study comparison. We have suggested four classifications, *universal*, *drug-like*, *mimetic* and *modeled* based on examples from the literature and on typical use-case analysis.

(ii) Providing access to primary data would allow the field to gain cumulative knowledge. The field of modeling has almost no "meta-analysis", i.e. research combining the results from studies, largely because of a lack of standards as to procedures and measures, but also due to the lack of primary data. A comprehensive format for virtual screening information would be useful.

(iii) The inclusion of multiple decoy sets of different design and intent for each target in an evaluation would, in combination with (i) and (ii) above, greatly increase the cumulative value of published studies.

(iv) The number of targets, actives and inactives need to be carefully considered with respect to the purpose of the experiment and the required accuracy of the results. These can be derived from simple statistical methods that are almost never applied.

(v) The effects of correlation between actives or inactives can be generally defined as an operational quantity. This could be investigated if actives and inactives for one target were included as *explicit* or *latent* decoys for all other targets. Warren et al. provides an example of the first, i.e. decoy sets were made from the actives of other targets. The second is an extension of point (iii), i.e. include multiple sets of decoys in a study but for different purposes. In conjunction with (ii) above, this would provide material for a rigorous analysis of operational correlation in virtual screening.

(vi) Correlation between targets needs further research, in particular the question of the variance of computational methods on closely related systems.

(vii) Differences between methods, especially within a single study over multiple targets, should only be reported if the effects of correlation are included. Editors of journals should never publish papers that suggest one method is better than another if these basic statistics are not employed. At a minimum it is recommended that the method variances along with correlation-corrected joint confidence limits be reported. This would allow the estimation of *p*-values for any assessment of method superiority.

On the issue of performance metrics we propose:

(i) Deciding on the metrics to be reported should be a community effort, although access to primary data to

encourage "meta-analysis" would aid the autonomous adoption of metrics.

(ii) There are good reasons metrics such as the AUC are popular in other fields and any new or additional measures for virtual screening need to be assessed against the characteristics that have made such metrics successful. Five characteristics required for a metric to be of similar heft to the AUC are proposed: independence to extensive variables, robustness, error bounds, no adjustable parameters and ease of interpretation. As an illustration, an improvement to the common enrichment measure is described. We propose the term "ROC enrichment" for this new measure. Similar improvements to early measures are proposed.

(iii) Currently, it would seem that providing AUCs and a few ROC enrichment values for the early part of a screen, e.g. 1% and 2%, would capture most *average* behavior of interest.

(iv) The assumption that 'early' behavior is necessarily a benefit is based on an assumed cost structure that may or may not hold. Similar statements are true for virtual screening in general. A rigorous attempt to assign real-word costs would be of use to the field.

(v) We have found very little evidence that suggests *average* behaviors cannot be accurately predicted by AUC or obvious extensions there of. Those suggesting otherwise need to provide clear-cut, statistically valid, evidence.

(vi) Divergence from (v) may be an indicator of local or domain knowledge, i.e. knowing the right answer and/or extensive knowledge of the system under study. A potential future area of research is whether this is also an indicator of over-parameterization, posterior system preparation or other reliance on retrospective knowledge. Interestingly, 2D methods applied to DUD, showed no evidence of such a divergence.

In conclusion, there is no reason it is not possible to establish standards in the field of molecular modeling necessary to enhance the quality of publications and allow a reliable assessment of methods and progress. However, there are also powerful incentives not to be rigorous. As one invested scientist was heard to pronounce, "livelihoods are at stake". This is true; we suggest the livelihood of the entire field. Whether the modeling community has the will to enact such measures may well determine whether future generations of scientists look back and see a field that became essential to drug discovery or one that became a mere footnote.

Acknowledgements The author wishes to thank Ajay Jain for discussions and his efforts on the "Evaluation of Computational Methods" symposium at the 234th American Chemical Society meeting, Martha Head for sharing the AUC and rank orderings of the GSK docking study, and Geoff Skillman, Paul Hawkins and Mark McGann for ideas, discussions and, most importantly, data. Finally, Christopher Bayly and Jean-Francois Truchon for spirited discussions on BedROC that, even if they did not convince the author, provided the spur for much of the work presented here.

References

1. Ioannidis JPA (2007) Contradicted and initially stronger effects in highly cited clinical research. JAMA 294(2):218–228
2. Obuchowski NA, Lieber ML, Wians FH (2004) ROC curves in *Clinical Chemistry*: Uses, misuses and possible solutions. Clin Chem 50(7):1118–1125
3. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49(20):5912–5931
4. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon J-F, Cornell WD (2007) Comparison of topological, shape, and docking methods in virtual screening. J Chem Inf Model 47(4):1504–1519
5. Sheridan RP This issue
6. Hanley JA, McNeil BJ (1982) The meaning and use of the area under an ROC curve. Radiology 143:29–36
7. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44: 837–845
8. Metz CE (1978) Basic principles of ROC analysis. Semin Nuclear Med VIII(4):283–298
9. Hanley JA, McNeil BJ (1983) A method of comparing the areas under the receiver operator characteristic curves derived from the same cases. Radiology 148:839–843
10. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluations of diffferent docking/scoring combinations. J Med Chem 43(25):4759–4767
11. Weininger D (1998) Combinatorics of small molecular structures. Encyclopedia of computational chemistry, vol 1. Wiley, New York, pp 425–430
12. Fink T, Bruggesser H, Reymond J-L (2005) Virtual exploration of the small molecule chemical universe below 160 Daltons. Angew Chem Int Ed 44:1504–1508
13. Jain AN (2004) Ligand-based structure hypothesis for virtual screeninG. J Med Chem 47:947–961
14. Cleves AE, Jain AN This issue
15. Hawkins P (2007) On how not to do an evaluation. Presentation at CUP8, Santa Fe, New Mexico, February 26th–28th 2007
16. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Del Rev 46:3–26
17. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WTM, Murray CW, Taylor RD, Watson P (2004) Virtual screening using protein-ligand docking: Avoiding artificial enrichment. J Chem Inf Comput Sci 44:793–806

18. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49(23):6789–6807
19. Geoffrey Skillman A (2007) Personal communication
20. ROCS, FRED are products of OpenEye Scientific Software, www.eyesopen.com. LINGOS are based on the OEChem SMILES canonicalization, also from OpenEye. MACCS Keys are fingerprints of molecular 2D features
21. Hawkins PC, Skillman AG, Nicholls A (2007) Comparison of shape-matching and docking as virtual screening tools. J Med Chem 50(1):74–82
22. Good A, Oprea T This issue
23. Clark RD This issue
24. Cheverud JM (2001) A simple correction for the multiple comparisons in interval mapping genome scans. Heredity 87:52–58
25. Fisher RA (1948) Combining independent tests of significance. Am Stat 2(5):30
26. SPEC website:.www.spec.org/spec
27. Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford Statistical Science Series. Oxford University Press. Chapter 5: Empircal Estimation
28. Sheridan RP, Singh SB, Fluder EM, Kearsley SK (2001) Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. J Chem Inf Comput Sci 41(5):1395–1406
29. Truchon J-F, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model 47(2):488–508
30. Bayly CI, Brideau C, Liaw A, Svetnick V (2006) Iterative focused screening using Random Forest: A comparison with HTS/random screening for two extreme cases. Thomas Kuhn Paradigm Shift Award Competition, 231st ACS National Meeting. March 26–30, 2006