Alternative Global Goodness Metrics and Sensitivity Analysis: Heuristics to Check the Robustness of Conclusions from Studies Comparing Virtual Screening Methods

Robert P. Sheridan*

Molecular Systems, Merck Research Laboratories, RY50SW-100, Rahway, New Jersey 07065

Received October 22, 2007

We introduce two ways of testing the robustness of conclusions from studies comparing virtual screening methods: alternative “global goodness” metrics and sensitivity analysis. While the robustness tests cannot eliminate all biases in virtual screening comparisons, they are useful as a “reality check” for any given study. To illustrate this, we apply them to a set of enrichments published in McGaughey et al. (*J. Chem. Inf. Model.* **2007**, *47*, 1504−1519) where 11 target protein/ligand combinations are tested on 2D and 3D similarity methods, plus docking. The major conclusions in that paper, for instance, that ligand-based methods are better than docking methods, hold up. However, some minor conclusions, such as Glide being the best docking method, do not.

INTRODUCTION

Virtual screening (VS) is an essential part of molecular modeling. In VS, one has some kind of model for a specific biological activity, and one searches a database for molecules likely to be active. Topological similarity, 3D similarity, docking, QSAR (also called machine learning), and so forth are examples of VS methods. They differ mainly in how the “model” is constructed. For instance, in topological similarity, it is assumed that molecules that resemble one or more target molecules based on their connection tables are most likely to be active. In docking, it is assumed that molecules most complementary to a protein binding site are most likely to be active.

An almost universal observation in VS is that the apparent goodness of a method varies greatly among potential targets, and it is almost impossible to predict in advance how well a method will do with a particular target/database combination. Also, different VS methods retrieve different molecules from the database. Thus, it is a common practice for a molecular modeler to run several VS methods for any given target and combine the results. Still, any research group can afford to maintain and run only a small set of potentially available methods, so we are forced to choose among them. Thus, while it is a futile endeavor to look for a universally best method, it is still valid and useful to ask which methods work better on more targets.

There are many studies in the literature comparing two or more VS methods by running several methods on several targets.[1−18] A number of these compare docking methods;[1−9] others compare ligand-based methods,[14−18] and still others compare docking with ligand-based methods.[10−13] Conclusions are often drawn from such studies stating the superiority of some methods over others, but how robust are the conclusions? One major difficulty is that the results are sensitive to the targets, database, and so forth, and no two investigators use the same targets or database, so one cannot easily compare methods across different studies. Also, if one is allowed to pick the targets, it is too easy to make any method look better than any other, especially when then number of targets is small. One type of solution to this problem is to settle on an agreed set of standard databases, targets, and sets of actives, so different methods across different studies can be sensibly compared. Some small progress has been made in the area of docking,[19,20] but currently there is not much agreement in the wider field of VS so far. A separate approach is to apply more rigor to individual studies, by testing the robustness of conclusions derived from the results. That is what we attempt here. We examine four potential “global goodness” metrics and suggest two types of sensitivity analysis to test how robust these metrics are for a given set of targets and methods. As an illustration, we apply these to one of our own studies (McGaughey et al.) that was recently published.[13]

METHODS

Enrichment Metrics. How “good” a method is usually depends on a series of retrospective VS experiments. We have a target *t*, a method *m*, and a database consisting of known actives and known (or presumed) inactives. Let us assume *n* actives in a database of *N* molecules. The target can be a single molecule, in the case of similarity methods, or it can be a protein-active site known at atomic resolution in the case of docking methods. One scores each entry in the database against the target and sorts the database entries in order of the score (descending or ascending, depending on whether high scores are more or less likely to be associated with activity). One then “tests” the database entries in that order and notes the total number of actives found as a function of the number of database entries tested. If the method is useful, the front of the list is enriched in actives relative to a list where the actives are randomly scattered. There are many metrics to measure the enrichment,[21−24] for example, area under the receiver-operator (ROC) curve,[21] the enrichment factor (EF) at a certain fraction *f* of the database, robust initial enhancement (RIE),[22] Boltzmann-

* Corresponding author e-mail: sheridan@merck.com.

10.1021/ci700380x CCC: $40.75
Published on Web 02/02/2008

enhanced discrimination of ROC (BEDROC),[23] η-squared,[24] the number of inactives that score above actives,[8] and so forth. Truchon and Bayly[23] discuss the various merits and drawbacks of these metrics. They argue that metrics that pay particular attention to the very front of a sorted list (i.e., those suitable for "early recognition" like EF, RIE, and BEDROC) are more appropriate for VS. In this paper, we use the EF at 1% of the database, that is, how many times more actives are found after 1% of the database ($f = 0.01$) is tested than would be expected if the actives were randomly distributed in the database. We make this choice because EF is popular in the literature, it is easy to understand, and it is the one used in McGaughey et al. However, the formulations below could use any enrichment metric.

One caveat: A characteristic of most enrichment metrics is that they are "saturatable"; that is, there is a highest possible value. In the case of EF it is $1/f$ or N/n, whichever is smaller. Truchon and Bayly[23] discuss the impact of saturation on enrichment metrics. Comparing methods is less meaningful if EF is at saturation for two or more values. None of the formulations here can overcome saturation effects.

Possible Global Goodness Metrics. Assume we have a matrix of EFs for T targets on M methods. The expression EF(t,m) will refer to the enrichment metric of method m on target t. Often, authors of studies compare methods by looking at individual targets and forming a qualitative impression. A single metric that represents all targets simultaneously might be called a "global goodness." A note on nomenclature:

Here, mean$_t$EF(t,m) will mean the EF(t,m) averaged over all targets for any given method, and mean$_m$EF(t,m) will refer to EF(t,m) averaged over all methods for any given target. stdev$_m$EF(t,m) will be the standard deviation over all methods for any given target.

Four possibilities for global goodness we examine are as follows:

1. meanEF. This is the simplest, the one used often in the literature, and the one we used in McGaughey et al. The global goodness for method m is simply the average over all targets:

$$\text{meanEF}(m) = \text{mean}_t\text{EF}(t,m)$$

The higher the meanEF(m), presumably the better the method. The advantage is that the numerical value of this metric is not dependent upon which set of methods is being compared. A significant weakness with meanEF is that the targets that have the largest differences between methods dominate the meanEF. Figure 1 shows the stdev$_m$EF(t,m) versus the mean$_m$EF(t,m) for the 11 targets, demonstrating that the targets with the largest differences are the ones where the EFs are highest. In effect, meanEF(m) will be disproportionately affected by the subset of targets on which the methods are doing well.

2. medianEF. This uses the median value instead of the average value over all targets for method m. We also used this in McGaughey et al. The larger medianEF(m) is, presumably the better the method. Again, the numerical value does not depend on the methods being compared, and since the median is considered a "robust" metric, we would expect medianEF(m) to be less sensitive to extreme values. medianEF(m) still has the drawback that targets with higher EFs dominate the metric, since one can show that the variance between methods, as measured by the median absolute difference, increases with the median.

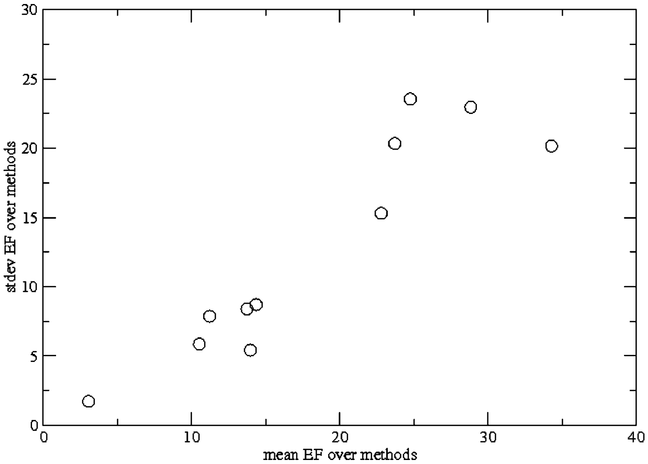

Figure 1. The standard deviation over all methods (stdev$_m$EF(t,m)) vs mean EF over all methods (mean$_m$EF(t,m)) for the data set in Table 1. Each circle represents a target. This demonstrates that the targets that have the largest EFs also have the largest variations between methods.

3. meanZscore. For each target, mean$_m$EF(t,m) and stdev$_m$EF(t,m) are calculated. Each EF(t,m) is re-expressed as a normalized score called Zscore:

$$\text{Zscore}(t,m) = [\text{EF}(t,m) - \text{mean}_m\text{EF}(t,m)]/\text{stdev}_m\text{EF}(t,m)$$

This transformation is illustrated in Figure 2A. The global goodness for method m will be

$$\text{meanZscore}(m) = \text{mean}_t\text{Zscore}(t,m)$$

Sometimes, stdev$_m$EF(t,m) is zero (i.e., all methods do equally well on the target), in which case the Zscore(t,m) cannot be calculated and the target must be omitted from the meanZscore(m). The higher the meanZscore(m) is, the better the method. This metric gives approximately equal weight to all the targets because the targets have been normalized relative to each other by their variation among methods. The disadvantage is that the numerical values of meanZscore(m) depend on the set of methods that are being compared.

4. meanRank. For each target t, the methods are ranked. This is illustrated in Figure 2B. Let us say method m5 has the highest EF(t,m) for target t. Then, Rank(t,m5) = 1. Method m1 is the next highest, so Rank(t,m1) = 2, and so forth. Ties in EF allow for ties in rank. For instance, if two methods are tied for second place, the ranks for both are 2.5. The ranking is done for all targets. The global goodness for method m will be the average rank over all targets:

$$\text{meanRank}(m) = \text{mean}_t\text{Rank}(t,m)$$

The lower the meanRank(m), the better the method. A method that was always the best would have a meanRank(m) of 1; a method that was always the worst would have a meanRank(m) = M. In practice, VS methods rarely have a meanRank(m) near 1 or M but are closer to the middle. Again, this metric weights all targets approximately equally, because the ranks do not depend on

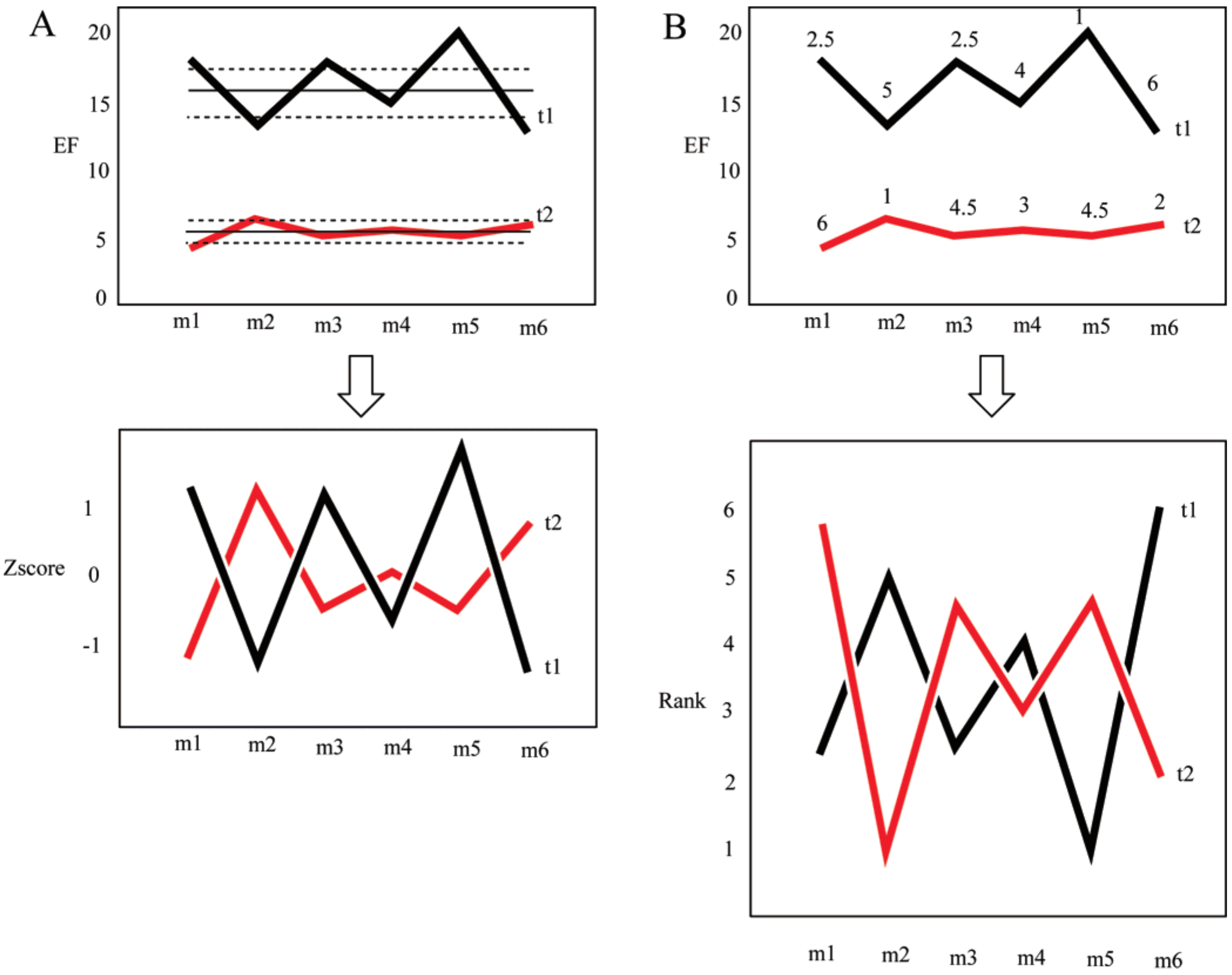

Figure 2. A simplified example with six methods (m1 to m6) and two targets (t1 and t2). The lines between methods are there so the eye can easily follow trends; no particular order among methods is implied. (A) The transformation of EFs for each target on the basis of the mean (horizontal solid lines) and standard deviation (horizontal dotted lines) of EFs over methods. (B) The transformation of EFs for a target on the basis of the relative rank of methods, 1 being the best method for that target, 2 being the second best, and so forth.

the absolute value of the EFs. Again, the value depends on the set of methods that are being compared.

Sensitivity by Database Perturbation. Individual EF(*t*,*m*) values are sensitive to the exact composition of the database, and therefore the global goodness measures may be also. Truchon and Bayly[23] proposed that one could monitor the variance of a goodness metric (in their case, BEDROC) by randomly removing 20% of the molecules in the database (regardless of whether they are active or inactive). They showed that the variance is very high if the number of actives is low. Here, we will test the robustness of the global goodness measures by generating 20 versions of the database where 20% of the molecules are deleted. For every version of the database, we recalculate EF(*t*,*m*) on the reduced database for every combination of *t* and *m*. To do this, it is necessary to know the ranks of the actives for every combination of *t* and *m* on the unperturbed database. Start with an array **a** with elements a[1] to a[*N*] set to "0". For each active with rank *i*, set a[*i*] to "1". For the molecules that are deleted, set a[*i*] to −1. One then moves upward from rank 1, incrementing the rank with every nondeleted molecule, and noting the new rank of the remaining actives. Given all the new EF(*t*,*m*)'s, we then calculate the global goodness measures for that version of the database. We monitor the standard deviation in global goodness for each method over the 20 versions of the database to create a type of "error bar" on the method.

Sensitivity by Target Set Perturbation. Whatever the global goodness metric, it will depend on the selection of targets. This will be especially true when the number of targets is small. One way of measuring the robustness of the conclusions of a study is to perturb the set of targets and see how much the global goodness changes. One would do this by randomly selecting a subset of targets from the full set and recalculating the global goodness metrics. How many targets should one include? A reasonable definition of "perturbation" might mean we omit 10−20% of the targets. If one does this for a number of perturbation trials, say 20, one can generate another type of "error bar" for the global goodness for each method.

Data Set. In McGaughey et al., a set of 11 protein structures and their corresponding ligands were chosen as targets. We applied three docking methods (FLOG, Glide, and FRED) using the protein, four 3D similarity methods

Table 1. EFs at 1% for the MDDR Database

target no. actives	TOPOSIM-	TOPOSIM-	TOPOSIM-				SQW-		ROCS-			
PDB data set	AP	TT	TTDT	Daylight	LINGO	SQW	shape	ROCS	color	FLOG	Glide	FRED
CA_I 80 1azm	56.4	62.7	36.4	50.2	45.1	6.3	3.8	11.3	31.4	1.3	1.3	2.5
CDK2 77 1aq1	22.2	11.7	20.8	10.4	9.1	9.1	11.7	11.7	18.2	1.3	1.3	9.1
COX2 257 1cx2/1cvu	21.1	8.6	19.1	5.1	14.1	11.3	15.2	17.2	25.4	0.4	4.7	0.4
DHFR 26 3dfr	34.7	42.4	46.3	27.0	19.3	46.3	0.0	3.9	38.6	15.4	7.7	15.4
ER 74 3ert	14.9	9.5	20.3	12.2	12.2	23.0	16.3	10.8	21.7	4.1	10.1	19.0
HIV-pr 136 1hsh	31.7	14.0	24.3	11.8	29.5	5.9	5.9	8.9	12.5	7.4	10.3	13.3
HIV-rt 149 1ep4	2.7	3.4	3.4	3.4	4.7	5.4	3.4	4.0	2.0	0.0	0.0	4.7
NEURAMINIDASE 12 1a4q	33.4	41.8	25.1	25.1	33.4	25.1	16.7	16.1	92.0	0.0	25.1	8.4
PTP1B 8 1c87	50.2	50.2	50.2	50.2	51.0	50.2	12.5	12.5	12.5	12.5	62.7	12.5
THROMBIN 200 1dwc/1mu6	28.6	22.6	32.6	13.0	10.5	27.1	4.5	6.0	21.1	4.0	14.0	6.5
TS 31 2bbq	22.7	51.8	64.7	61.5	31.0	48.5	6.5	0.0	6.5	9.7	9.7	12.9
meanEF	29.0	29.0	31.2	24.5	23.6	23.5	8.8	9.3	25.6	5.1	13.4	9.5
medianEF	28.6	22.6	25.1	13.0	19.3	23.0	6.5	10.8	21.1	4.0	9.7	9.1
meanZscore	0.80	0.46	0.96	0.13	0.30	0.43	−0.58	−0.54	0.52	−1.26	−0.69	−0.50
meanRank	3.9	4.8	3.4	6.0	5.1	5.2	8.6	8.3	5.3	10.8	8.6	8.0

(ROCS, ROCS-color, SQW, and SQW-shape) using the ligand, and two topological similarity methods (TOPOSIM and Daylight 3/10) using the ligand. We did the VS over two databases: a diverse subset of the MDL Drug Data Report (MDDR) (www.mdli.com, last accessed November 16, 2007) and a diverse subset of the MCIDB (Merck's proprietary database) and noted the EF at 1%. For the purposes of this paper, we added three new topological methods: TOPOSIM-TT, TOPOSIM-TTDT, and LINGO. Here, we will use the notation "TOPOSIM-AP" for what we called "TOPOSIM" in McGaughey et al.: topological similarity with the Carhart atom pair descriptor and Dice similarity index.[25] TOPOSIM-TT is a topological similarity method using the TT (topological torsion) descriptor[26] and the Dice similarity index. TOPOSIM-TTDT is a "combination descriptor" of topological torsions and "binding property torsions" presented in Kearsley et al.[27] (It is called "ttbt" in that publication). Recently, in other studies, we found that TTDT is a near-optimal combination for lead-hopping, at least for the MDDR. LINGO was first described by Vidal et al.[28] We generated LINGO descriptors using the OEChem (from OpenEye Scientific Software, www.eyesopen.com) version of canonical SMILES. Although McGaughey et al. looked at two databases, for illustrative purposes, we will look only at the MDDR results. The EFs are in Table 1. There is only one EF (NEURAMINIDASE/ROCS-color) close to the maximum possible value ($100 = 1/f$ or $96 = N/n$ for COX2).

RESULTS

Alternative Global Goodness Metrics. The global goodness metrics for each method are listed in Table 1. Selected graphs of one global goodness metric versus the others are in Figure 3. The various metrics are highly correlated, but there are some differences. The most obvious difference is with Glide, which has a higher meanEF than expected from the meanZscore and meanRank. This is probably because meanEF is heavily influenced by PTP1B, which has a very high $\text{mean}_m\text{EF}(t,m)$. The other three metrics feel the influence of that particular target much less. Daylight is an outlier of medianEF versus the other metrics, having a lower medianEF than expected. Since both metrics involve rescaling, it is not surprising that meanZscore and meanRank are more anticorrelated with each other than either is correlated with meanEF or medianEF, but they can still be slightly different from each other because meanRank is concerned only with the relative order of the methods, while meanZscore pays attention to the quantitative differences between methods.

By any of the global goodness metrics, the methods fall into three categories:

1. good: topological methods (TOPOSIM-AP, TOPOSIM-TT, TOPOSIM-TTDT, Daylight, and LINGO) and atom-typed 3D similarity methods (ROCS-color and SQW)
2. medium: some docking methods (Glide and FRED) and shape-only 3D similarity methods (ROCS and SQW-shape)
3. poor: FLOG, a docking method.

This is generally consistent with the major conclusions in McGaughey et al. However, the fact that Glide appears better than FRED only in meanEF weakens the conclusion of McGaughey et al. that it is the best docking method.

Sensitivity Analysis through Perturbation. Results of the perturbation studies are shown in Figure 4 for each global goodness metric. In each case, the values for the unperturbed metrics are shown as an open circle. We include two types of perturbation with 20 trials each: database perturbation where the database is reduced by 20% (red) and target set perturbation where we include 9 out of 11 targets (black). For these, we show the mean (solid circle) and error bars at ±1 standard deviation. If we assume that two types of perturbations have the same mean (see below) and are uncorrelated, it is possible to estimate the overall standard

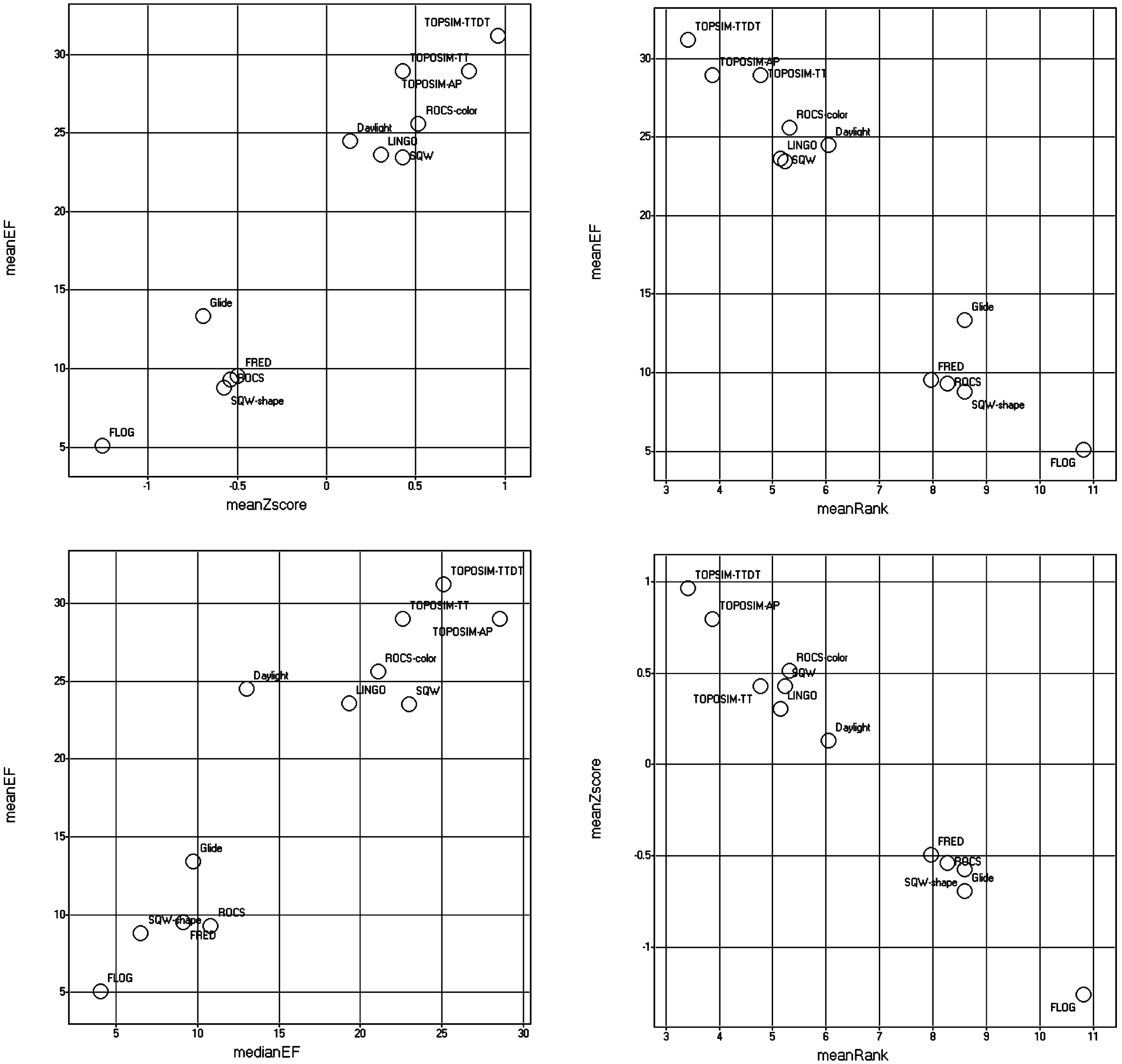

Figure 3. Comparison of three global goodness metrics for the 11 targets.

deviation[29] with the following simple formula:

$$\mathrm{stdevO} = \mathrm{sqrt}(\mathrm{stdevD}^2 + \mathrm{stdevT}^2)$$

where stdevD and stdevT are the standard deviations for the database and target set perturbations, respectively. The stdevO's are shown in green in Figure 4 and are assumed to center around the unperturbed value. Henceforth, we shall refer to the means ± 1 standard deviation as the "error bars".

We make the following observations:

1. stdevD and stdevT are of comparable size for this data set. stdevO is usually not much larger than the larger of the two. This is not surprising, since the maximum possible value of stdevO is sqrt(2) = ~1.4 × max(stdevD,stdevT), and it will usually be closer to 1.0 × max(stdevD,stdevT).

2. The standard deviation can be a large fraction of the distance between some methods, which generally means that many methods are not really distinguishable. For instance, the difference between Daylight, LINGO, and SQW are always smaller than the standard deviations whatever global goodness one picks. Similarly, this is so for SQW-shape, ROCS, and FRED. Glide could be considered a member of that group even when viewed by meanEF if one takes into account the large value of stdevT. This is perhaps not surprising since, as noted before, Glide appears good in meanEF because of the influence of a single target PTP1B, and removing that target has a big effect. Again, therefore, the conclusion of McGaughey et al. that Glide is the best docking method seems dubious.

3. The relative size of the standard deviations seems especially large for medianEF, in particular for stdevT, less large for meanEF (particularly for the methods with high meanEF), and small for meanZscore and meanRank. An anonymous reviewer pointed out that it is expected that stdevT for medianEF would be large because it is known that the median is less efficient at approaching the "true" value than the mean as a function of sample size.[30] Thus,

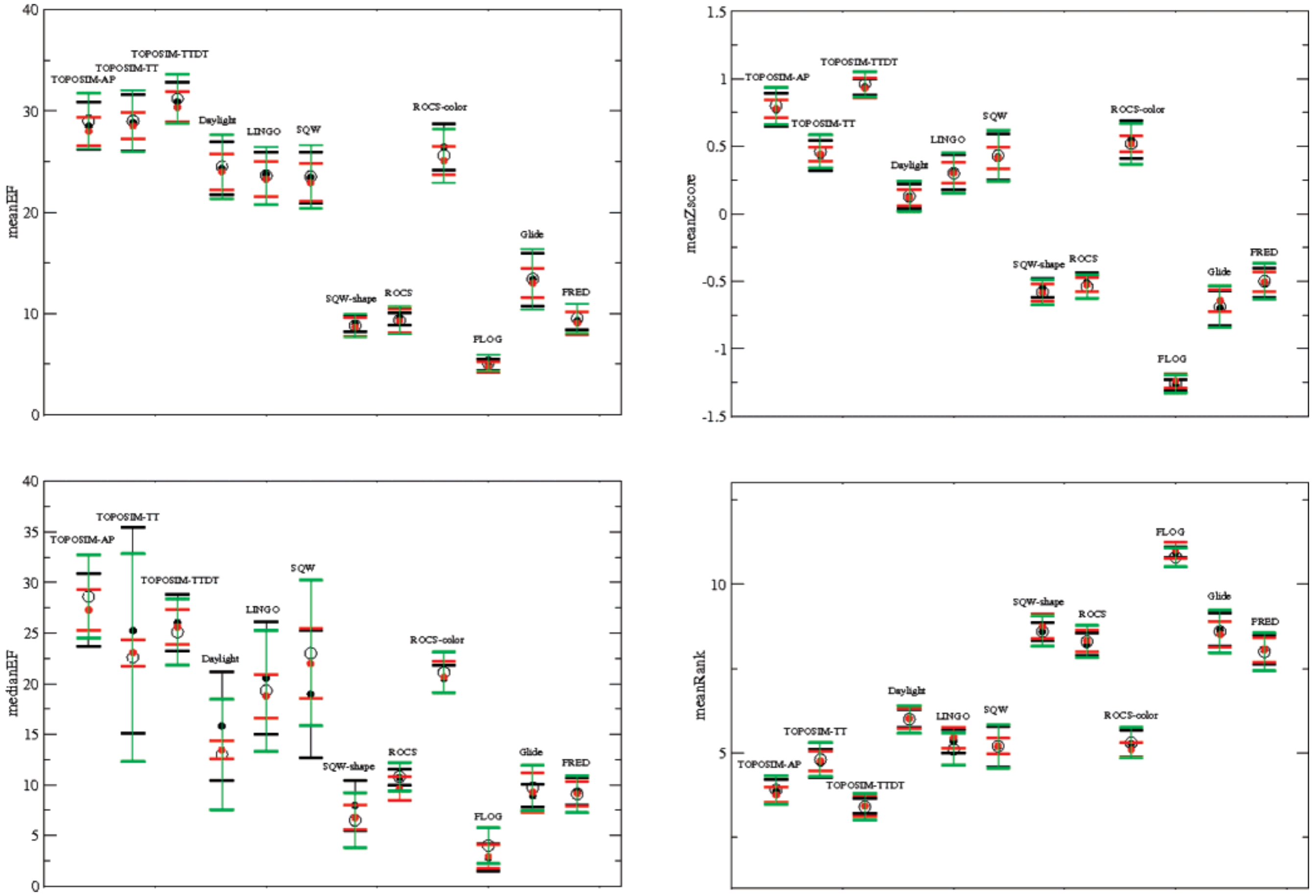

Figure 4. Perturbation effects on global goodness metrics. The unperturbed metrics for all 11 targets are shown as open circles. The mean and standard deviations are shown for 20 trials of "database perturbation" (red) and 20 trials of "target set perturbation" (black). In each trial of database perturbation, 20% of the molecules in the database are randomly omitted, the individual EFs for each combination of target and method are recalculated, and the global goodness metrics are recalculated from the individual EFs. In each trial of target set perturbation, 9 targets out of 11 were randomly selected and the global goodness metrics recalculated. Error bars that take into account both the database and target set perturbations simultaneously are shown in green. They are assumed to center around the unperturbed value.

we feel that medianEF is not a very good metric to use with perturbation, at least with a sample size as small as nine targets. The fact that the standard deviations for meanEF are large compared to meanZscore and meanRank probably can be accounted for by the fact that meanEF takes into account effectively fewer targets, and the EF of those targets is high, and therefore meanEF is more sensitive to the removal of a few targets during target set perturbation.

4. For the most part, the unperturbed values of global goodness (open circles) and the mean of the perturbed values (solid circles) are very close most of the time, at least compared to the size of the standard deviations, indicating that either type of perturbation is equally likely to produce changes in either direction. The exception seems to be in the medianEF, but we have already argued why medianEF might not be suitable for perturbation.

5. There is some justification for considering some of the TOPOSIM methods to be better than Daylight and LINGO, since for most of global goodness metrics the error bars do not significantly overlap. The conclusion from McGaughey et al. that TOPOSIM-AP is better than Daylight still appears sound. It is not clear that TOPOSIM-TTDT is better for this data set than TOPOSIM-AP.

DISCUSSION

We introduced two proposals to help test the robustness of conclusions from comparisons of VS methods. The first is the introduction of alternative global goodness metrics. Although we formulated them with EF, these global goodness metrics are applicable to any metric for enrichment (area under the ROC curve, BEDROC, etc.). One expects that they are equally applicable to other measures of merit one wants to compare between methods, such as diversity of actives or, in the case of docking, how close predicted poses come to the X-ray pose. The point of having more than one global goodness metric is that, if one VS method is shown to be better than another by more than one of these metrics, one has more confidence that this relationship is true, and not just an accident of one particular definition.

Two new metrics introduced with this paper are meanZscore and meanRank. MeanEF and medianEF, the simplest metrics, are overly influenced by particularly good targets. In contrast, meanZscore and meanRank normalize EF for each target and allow for approximately equal weighting of targets. It has been a matter of philosophical debate in our group whether good targets should dominate, but the availability of the new metrics allows a choice. Another concern with combining EFs directly is that perhaps some targets may have actives that are very different in physical properties (e.g., molecular weight) than the database as a whole, and therefore some EFs will be artificially high, while other targets will not. meanZscore and meanRank normalize each target beforehand, so those effects should be reduced. It should be re-emphasized though that, unlike meanEF and medianEF, the numerical values of meanZscore and meanRank are dependent on the set of methods being compared and indicate only the relative value of the methods among themselves, not the absolute value. For instance, had we looked at only topological methods, all the meanEF's would have been fairly high on an absolute basis, but some would have had very negative meanZscores.

The second concept we propose is to test how sensitive the conclusions are to small changes. The first method is "database perturbation," wherein one tests the effect of variation in individual EF(*t*,*m*)'s on the global goodness metrics, based on reducing the database. The second method is "target set perturbation", wherein one tests the sensitivity of the global goodness metrics on the basis of including only a subset of targets. One can think of one standard deviation of the global goodness metrics under these two types of perturbations as two kinds of heuristic "error bar". Considering the two types of error bar together does not change the conclusions that could be drawn by looking at the larger error bar. Generally, the size of the standard deviation for database perturbation will depend strongly on the number of actives associated with the targets; the smaller the number of actives, the larger it is. On the other hand, the standard deviation for target set perturbation will depend on the number of targets; the smaller the target set, the larger it is.

For the purposes of inspection, it would be equally valid to define the error bars as ∼2.0 standard deviations (i.e., as a 95% confidence interval) instead of 1.0. It should be remembered that the standard deviations depend on the somewhat arbitrary amount of "perturbation" one introduces. Therefore, we are not suggesting that it is possible to calculate an absolute probability that the differences between methods are real, only that we take the error bars as a qualitative heuristic that lets one judge which differences between methods are more likely to be "real" than others.

It should be noted that target set perturbation can be applied retrospectively to VS studies in the literature where only the matrix of EFs has been provided. In contrast, database perturbation as we have implemented it here requires the ranks of actives used to calculate each EF(*t*,*m*), and most of the time these are not published. However, even in the absence of such information, it might be possible to estimate the error bars on individual EF(*t*,*m*)'s without doing any explicit perturbations on the database. For instance, Truchon and Bayly have shown, and we can confirm, that the maximum variation in EF(*t*,*m*) is well-correlated with 1/sqrt(8*n*), where *n* is the number of actives for target *t*.

Having examined four global goodness metrics, our preference is for meanZscore and meanRank over the simpler meanEF and medianEF. We believe the extra complexity is justified because of the follwing:

1. meanZscore and meanRank give more uniform weighting to all targets and compensate for some nonequivalencies between targets.

2. meanZscore and meanRank are well-behaved under both types of perturbation and give error bars that are small relative to the difference between methods.

There are many biases that can come in when one is comparing VS methods: the selection of targets, the selection of decoys, the adjustable parameters one uses for the individual methods, and so forth. The concepts introduced here cannot correct for every bias and are meant to apply only within a particular study. Therefore, having one's conclusions be robust to both alternative global goodness metrics and sensitivity analysis means only that the conclusions within that particular study are valid. One cannot say the conclusions are generally "true" until one uses a new database, a separate set of targets, and so forth. However, the ability to quickly eliminate false conclusions from any given study is very useful, and methods for doing so should be consistently applied.

ACKNOWLEDGMENT

The author thanks the other authors of the previous study: Georgia McGaughey, Christopher Bayly, Chris Culberson, Constantine Kreatsoulas, Stephanie Lindsley, Vladimir Maiorov, Jean-Francois Truchon, and Wendy Cornell. The author especially thanks Christopher Bayly for insisting on the inclusion of "database perturbation." Viktor Hornak wrote the scripts to generate the LINGO descriptors and provided useful discussions. Discussions with Simon Kearsley provided the impetus for this work.

Supporting Information Available: Ranks of actives for all the searches are supplied.

REFERENCES AND NOTES

(1) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Struct. Funct. Bioinf.* **2004**, *56*, 235−249.

(2) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct. Funct. Bioinf.* **2004**, *57*, 225−242.

(3) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P.; Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* **2005**, *48*, 962−976.

(4) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11−22.

(5) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2005**, *49*, 5912−5931.

(6) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.* **2006**, *46*, 401−415.

(7) Onodera, K; Satou, K.; Hirota, H. Evaluations of molecular docking programs for virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1609−1618.

(8) Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. *J. Chem. Inf. Model.* **2007**, *47*, 1599−1608.

(9) Andersson, C. D.; Thysell, E.; Lindstrom, A.; Bylesjo, M.; Raubacher, F.; Linusson, A. A multivariate approach to investigate docking parameters' effects on docking performance. *J. Chem. Inf. Model.* **2007**, *47*, 1673−1687.

(10) Zhang, Q.; Muegge, I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: Ranking, voting, and consensus scoring. *J. Med. Chem.* **2006**, *49*, 1536−1548.

(11) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *Med. Chem.* **2007**, *50*, 74−82.

(12) Moro, S.; Bacilieri, M.; Deflorian, F. Combining ligand-based and structure-based drug design in the virtual screening arena. *Expert Opin. Drug. Disc.* **2007**, *2*, 37−49.

(13) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C. K.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504−1519.

(14) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 529−536.

(15) Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J. Med. Chem.* **2004**, *47*, 6144−6159.

(16) Kogej, T.; Engkvist, O.; Blomberg, N.; Muresan, S. Multifingerprint based similarity searches for targeted class compound selection. *J. Chem. Inf. Model.* **2006**, *46*, 1201−1213.

(17) Ewing, T.; Baber, J. C.; Feher, M. Novel 2D fingerprints for ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 2423−2431.

(18) Sperandio, O.; Andrieu, O.; Miteva, M. A.; Vo, M.-Q.; Souaille, M.; Defaud, F.; Villoutriex, B. O. MED-SuMoLig: a new ligand-based screening tool for efficient scaffold hopping. *J. Chem. Inf. Model.* **2007**, *47*, 1097−1110.

(19) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, 49, 6789−6801.

(20) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W.; Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(21) 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.

(22) 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.

(23) Truchon, J.-F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the early recognition rroblem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.

(24) Seifert, M. H. J. Assessing the discriminatory power of scoring functions for virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1456−1465.

(25) Carhart, R. E.; Smith, D. H.; Ventkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(26) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.

(27) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−27.

(28) Vidal, D.; Thormann, M.; Pons, M. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386−393.

(29) Hocking, R. R. In *Methods and Applications of Linear Models*; John Wiley and Sons: New York, 1996; Chapter 2, p 40.

(30) Venables, W. N.; Ripley, B. D. In *Modern Applied Statistics with S*, 4th ed.; Springer: New York, 2002; Chapter 5, p 121.

CI700380X