

# Boosting Virtual Screening Enrichments with Data Fusion: Coalescing Hits from Two-Dimensional Fingerprints, Shape, and Docking

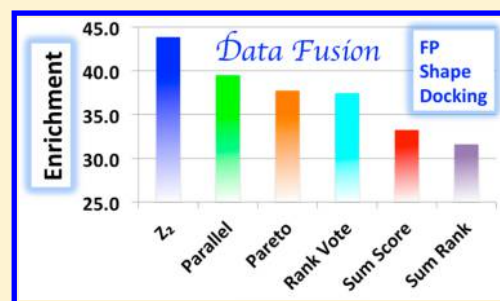
G. Madhavi Sastry,<sup>†</sup> V. S. Sandeep Inakollu,<sup>†</sup> and Woody Sherman<sup>\*,§</sup>

<sup>†</sup>Schrödinger, Sanali Infopark, 8-2-120/113, Banjara Hills, Hyderabad 500034, Andhra Pradesh, India

<sup>§</sup>Schrödinger, 120 West 45th Street, New York, New York 10036, United States

## Supporting Information

**ABSTRACT:** Virtual screening is an effective way to find hits in drug discovery, with approaches ranging from fast information-based similarity methods to more computationally intensive physics-based docking methods. However, the best approach to use for a given project is not clear in advance of the screen. In this work, we show that combining results from multiple methods using a standard score (Z-score) can significantly improve virtual screening enrichments over any of the single screening methods. We show that an augmented Z-score, which considers the best two out of three scores for a given compound, outperforms previously published data fusion algorithms. We use three different virtual screening methods (two-dimensional (2D) fingerprint similarity, shape-based similarity, and docking) and study two different databases (DUD and MDDR). The average enrichment in the top 1% was improved by 9% for DUD and 25% for the MDDR, compared with the top individual method. Improvements of 22% for DUD and 43% for MDDR are seen over the average of the three individual methods. Statistics are presented that show a high significance associated with the findings in this work.



## ■ INTRODUCTION

Virtual screening has been established as an effective way to find hits in drug discovery. While pitfalls exist,<sup>1</sup> there are many examples of successful virtual screening campaigns in the literature using a variety of methods.<sup>2–6</sup> On one end of the spectrum are one- and two-dimensional (1D and 2D) ligand-based approaches that rely solely on knowledge derived from the connectivity information of one or more active ligands, such as 2D fingerprints,<sup>7–11</sup> substructure matching,<sup>12–15</sup> and even text-based similarities.<sup>16,17</sup> These methods tend to be very fast and have proven to be successful, although the extent of their success depends heavily on the similarity of active database compounds to the query. On the other end of the spectrum are docking methods, which tend to rely on physics-based scoring of protein–ligand complexes.<sup>18–26</sup> Docking generally takes more computational time but offers the opportunity to find new and diverse actives that are unrelated to existing active compounds.<sup>27–30</sup> In the middle of these extremes lies 3D ligand-based methods such as pharmacophore<sup>31,32</sup> and shape screening.<sup>33–35</sup> These methods take advantage of 3D information from active ligands and search for other ligands that match the 3D properties of a query molecule.

In most published examples, the different virtual screening approaches mentioned above have been used independently. However, a number of groups have shown that information can be combined from multiple computational methods to improve results.<sup>36–39</sup> For example, researchers at Vertex explored two different docking methods and thirteen scoring functions and showed that consensus scoring (interchangeably called data fusion) can improve results.<sup>40</sup> Researchers at Boehringer

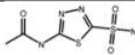
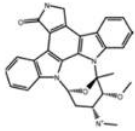
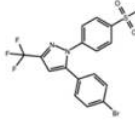
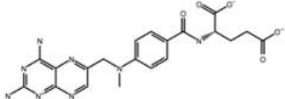
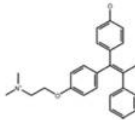
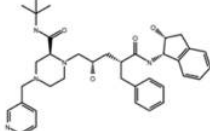
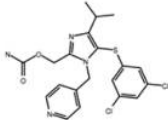
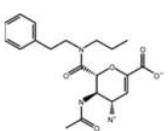
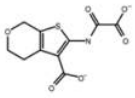
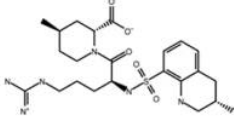
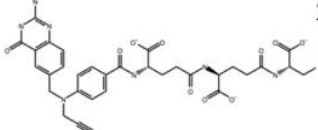
Ingelheim later showed that several ligand-based methods could be combined to improve scaffold hopping and database enrichments.<sup>41</sup> Additional work has been done to lay out a theoretical framework for determining if data fusion techniques will improve results when the performance of the individual methods is known.<sup>42</sup> In that work, the approach was applied to docking with five scoring functions and two evolutionary algorithms on three targets. The authors showed an improvement in virtual screening enrichment compared with a single scoring function and docking method. A statistical analysis revealed that the improvement imparted by data fusion can be understood based on the fact that, given an error in the predicted property of interest (i.e., active or inactive), the combination of multiple statistical measurements tends to be closer to the real value.<sup>43</sup> In that same study the authors established that three or four scoring functions are sufficient to realize the statistical gains of consensus scoring and data fusion. Several other papers have described the value of consensus scoring,<sup>44–49</sup> including applications beyond virtual screening, such as pose assessment and binding energy prediction.<sup>50–52</sup> Extensions to data fusion that incorporate belief theory can provide confidence estimates on predictions and offer an attractive approach when enough data is available about the performance of the methods of interest.<sup>53</sup>

In the work here, we take the general principles from the previous data fusion studies and apply them to a more diverse set of virtual screening methods than previously published (2D fingerprints, 3D shape, and docking). Furthermore, we expand

Received: September 28, 2012

Published: June 19, 2013

Table 1. Targets Studied in This Work<sup>a</sup>

Target	Co-crystallized Ligand	PDB Code	#Actives
Carbonic Anhydrase I (CA)		1azm	80
Cyclin-dependent Kinase 2 (CDK2)		1aq1	77
Cyclooxygenase 2 (COX2)		1cvu/1cx2	257
Dihydrofolate Reductase (DHFR)		3dfr	26
Estrogen Receptor Alpha (ER)		3ert	74
HIV Protease (HIVpr)		1hsh	136
HIV Reverse Transcriptase (HIVrt)		1ep4	149
Neuraminidase (NA)		1a4q	12
Protein Tyrosine Phosphatase 1B (PTP1B)		1c87	8
Thrombin (Throm)		1mu6(1dwc)	200
Thymidylate Synthase (TS)		2bbq	31

<sup>a</sup>Abbreviated target names are shown in parentheses. The ligands shown (extracted from the targets) were used for Shape Screening and Canvas 2D fingerprint methods. The COX2 ligand was extracted from a different PDB structure (1cx2) than the structure used for the docking calculations (1cvu), as in the original work by McGaughey et al.<sup>56</sup>

Table 2. Database Enrichment Values Obtained from the Individual Methods Used in This Work<sup>a</sup>

method	DUD				MDDR			
	BEDROC ( $\alpha = 20$ )		EF1%		BEDROC ( $\alpha = 20$ )		EF1%	
	mean	median	mean	median	mean	median	mean	median
D	0.47	0.46	18.9	17.3	0.45	0.48	33.6	31.0
M	0.46	0.41	19.0	17.7	0.46	0.40	35.1	23.5
R	0.47	0.47	19.6	18.3	0.45	0.51	33.2	27.0
C	0.43	0.38	17.4	15.7	0.44	0.47	32.0	29.5
X	0.43	0.38	17.7	16.5	0.44	0.46	33.2	28.0
G	0.36	0.33	13.1	11.6	0.29	0.27	17.4	9.5

<sup>a</sup>Mean and median values across the targets are shown. D, M, and R designate dendritic, Molprint2D, and radial fingerprints, respectively. C and X designate the ConfGen and X-ray crystal conformation, respectively, of the query ligand used for Shape Screening calculations. G denotes Glide docking.

Table 3. Mean and Median BEDROC ( $\alpha = 20$ ) Enrichment Values for the DUD and MDDR Sets Using the Two Z-Score Data Fusion Methods on Six Combinations of Fingerprints, Shape, and Docking<sup>a</sup>

	DUD				MDDR			
	$Z_2$		$Z_3$		$Z_2$		$Z_3$	
	mean	median	mean	median	mean	median	mean	median
DCG	0.53	0.53	0.52	0.53	0.53	0.60	0.53	0.61
MCG	0.49	0.47	0.48	0.48	0.51	0.53	0.51	0.54
RCG	0.52	0.52	0.51	0.49	0.54	0.58	0.53	0.60
DXG	0.53	0.52	0.52	0.52	0.54	0.61	0.54	0.63
MXG	0.50	0.49	0.48	0.49	0.52	0.54	0.51	0.56
RXG	0.52	0.49	0.51	0.48	0.55	0.59	0.54	0.61

<sup>a</sup> $Z_2$  is the average of the best two Z-scores from the three screening methods whereas  $Z_3$  is the average of all three Z-scores.

upon the previous works by developing a new data fusion algorithm that considers only the best two of three scores from the diverse methods. We show that database enrichments are improved over other data fusion approaches and present statistics to assess the significance of the improvements. We find that database enrichments can be improved considerably with data fusion and show examples with explanations for the improvements, especially in the cases where only subsets of the individual methods perform well. Finally, we discuss the limitations of this approach and future research directions.

## MATERIALS AND METHODS

**Data Sets.** Two data sets were used for the calculations, MDDR<sup>54</sup> and DUD.<sup>55</sup> For the MDDR, the actives and decoys were obtained from previous work by McGaughey et al.<sup>56</sup> The targets and number of active ligands are shown in Table 1. A total of 24 116 ligands and 11 targets from the MDDR were used. The DUD release 2 data set includes 40 targets with a 33:1 average decoy to active ratio and ligands that have been carefully selected to ensure diversity within the active compounds along with per-target curated decoy molecules that have similar properties to the active compounds for that target (<http://dud.docking.org/r2>). The MDDR proteins were prepared using the Protein Preparation Wizard<sup>57</sup> in Maestro with default settings, and the DUD proteins were used as prepared from the DUD Web site. The H-bond network was optimized by flipping the terminal chi angle for Asn, Gln, and His side chains. Neutral and protonated states of Asp, Glu, and His states were varied and hydroxyl/thiol hydrogens were sampled. Finally, an all-atom minimization with a 0.3 Å heavy-atom RMSD criteria for termination was performed using the Impref module of Impact<sup>58</sup> and the OPLS\_2005 force field.<sup>59–61</sup> These settings were determined

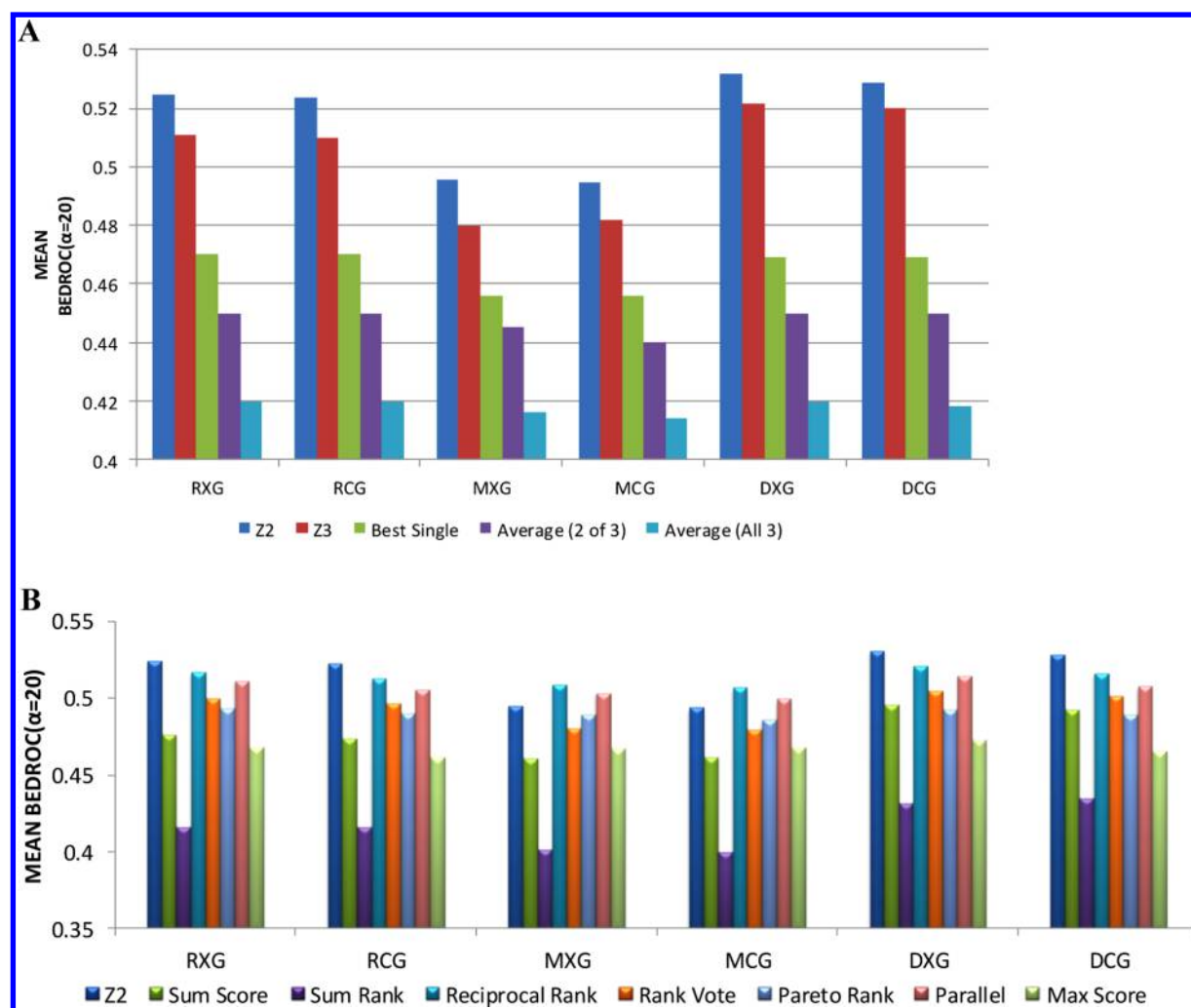
to be appropriate for virtual screening structure preparation based on previous work.<sup>57</sup>

Three-dimensional coordinates were generated for all ligands with LigPrep.<sup>62</sup> Ligand ionization and tautomeric states were generated with Epik.<sup>63,64</sup> In addition to calculating reasonable ligand states, Epik also estimates a state penalty to quantify the energetic cost it takes to generate each state. The Epik state penalty is computed in units of kilocalories per mole and was added to the GlideScore (also in kilocalories per mole) to get the final DockingScore.

**Screening Methods.** 2D molecular fingerprints were generated with Canvas<sup>7,8</sup> in the Schrödinger Suite for the MDDR and DUD database and query molecules. The best three fingerprint methods from our earlier study were used (Molprint2D with element + aromatic + cyclic atom types, dendritic with Daylight atom types, and radial with Daylight atom types).<sup>8</sup> Enrichments were computed based on the computed similarities between the database and query fingerprints. Tanimoto similarity was used for dendritic and radial fingerprints, and Buser similarity was used for Molprint2D.

For shape-based screening, we used Shape Screening in the Schrödinger Suite with the “pharm” option, which combines shape with pharmacophoric feature types and was shown to be the best virtual screening setting in our earlier study.<sup>33</sup> Two separate shape screens were performed: one with the crystallographic conformation for the query and the other with the lowest energy conformation from a ConfGen conformational search.<sup>65</sup> The same query molecules were used as for the fingerprint screening (see above).

Finally, docking calculations were performed with Glide<sup>18,19</sup> using the high-throughput virtual screening (HTVS) mode and Epik state penalties. All water molecules were removed before docking. Receptor grids were generated prior to docking using all



**Figure 1.** Mean BEDROC ( $\alpha = 20$ ) enrichments from different scoring and data fusion methods on the DUD data set. (A) Comparison between the Z-score data fusion methods presented here ( $Z_2$  and  $Z_3$ ) and the individual screening methods. Blue is for average of the top two Z-scores ( $Z_2$ ), red for the average of all three Z-scores ( $Z_3$ ), green for the best single method within the set of three methods, purple for the average of the best two of three methods, and cyan for the average of all three methods. Refer to Materials and Methods for a definition of the three-letter abbreviations. (B) Comparison between multiple data fusion approaches. In addition to  $Z_2$  (blue), we also show Sum Score (green), Sum Rank (purple), Reciprocal Rank (cyan), Rank Vote (orange), Pareto Rank (gray), Parallel Selection (pink), and MaxScore (light green).

default options. Final ranking from the docking was based on the DockingScore, which combines the Epik state penalty with the GlideScore.<sup>63</sup>

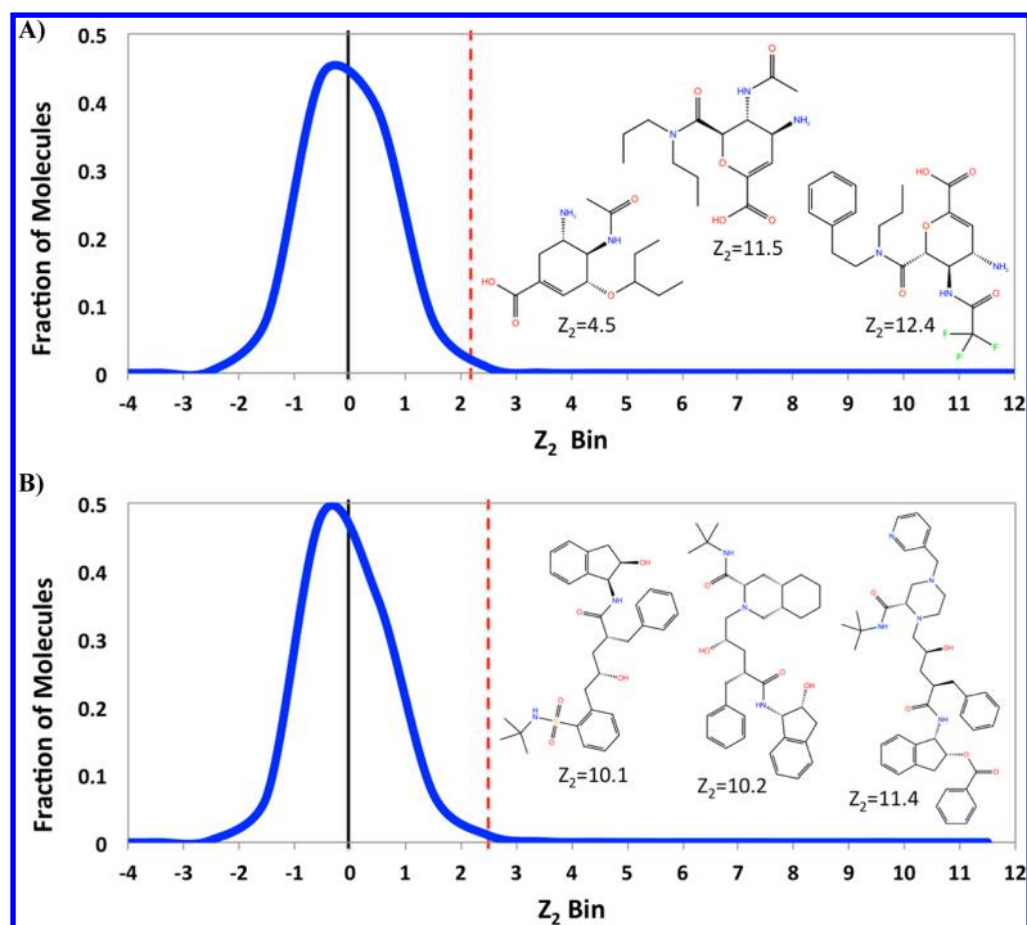
**Data Fusion.** The scores obtained from the three independent methods prevent directly adding or averaging the values to obtain a single score, since GlideScore is provided in units of kilocalories per mole (with more negative being better) while the fingerprint and shape similarity scores lie on the range [0, 1] with more positive being better. Furthermore, simply scaling the scores from the methods does not properly account for the variability and dynamic range of the different methods. While many possible ways exist to combine data from multiple screens, here we converted the scores for each ligand to a standard score (i.e., Z-score) as defined in eq 1. We find Z-scores to be particularly useful because they indicate by how many standard deviations a value is above or below the mean of a distribution. The Z-score is a dimensionless quantity derived by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. In addition, the Z-score can be directly converted into a percent

rank within the distribution, where  $1\sigma$  above the mean ( $\sigma$  is a standard deviation) represents the top 84.1% of the distribution,  $2\sigma$  represents the top 97.7%,  $3\sigma$  represents the top 99.9%, etc. We then calculated the average Z-score from either all three methods ( $Z_3$ ) or only the best two of the three methods for each ligand ( $Z_2$ ). The motivation for the  $Z_2$  score was based on empirical observations that, for a given screen, we often see only two of the three methods (fingerprint, shape, and docking) scoring a given active well.

$$Z_{mi} = \frac{S_{mi} - \mu_m}{\sigma_m} \quad (1)$$

In eq 1, " $Z_{mi}$ " is the Z-score obtained for ligand " $i$ " from method " $m$ ". " $S_{mi}$ " is the score of the  $i$ th ligand in the database for method  $m$ , " $\mu_m$ " is the mean score of all the compounds (actives plus decoys) in the database, and  $\sigma$  is the standard deviation of the distribution of the scores obtained from the method  $m$ . The sign on the docking scores was inverted to maintain consistency with fingerprints and shape, where a more positive score is better. The average Z-scores for all three methods ( $Z_3$ ) or only the best two





**Figure 2.**  $Z_2$  probability distribution curves for two targets. The red dotted line corresponds to the point where 1% of the database is retrieved. (A) Neuraminidase with RXG scoring. (B) HIV protease with DXG scoring. The compounds shown are the top three  $Z_2$  scores for each target (all six are actives). The  $Z_2$  value is given below each structure.

of the three methods for each ligand ( $Z_2$ ) were then sorted to rank the compounds and calculate the enrichment factors. For the  $Z_2$  score, the two methods used for the final score are based solely on the Z-score of each compound for each method, thus no prior information is needed to use this method and it is not biased by knowing in advance which method performs better for a given compound. Finally, we extended the application of Z-scores to all six virtual screening protocols from this study (three fingerprint, two shape, and one docking) and explored the effect of using anywhere between two ( $Z_2$ ) to all six ( $Z_6$ ) of the best Z-scores for each ligand.

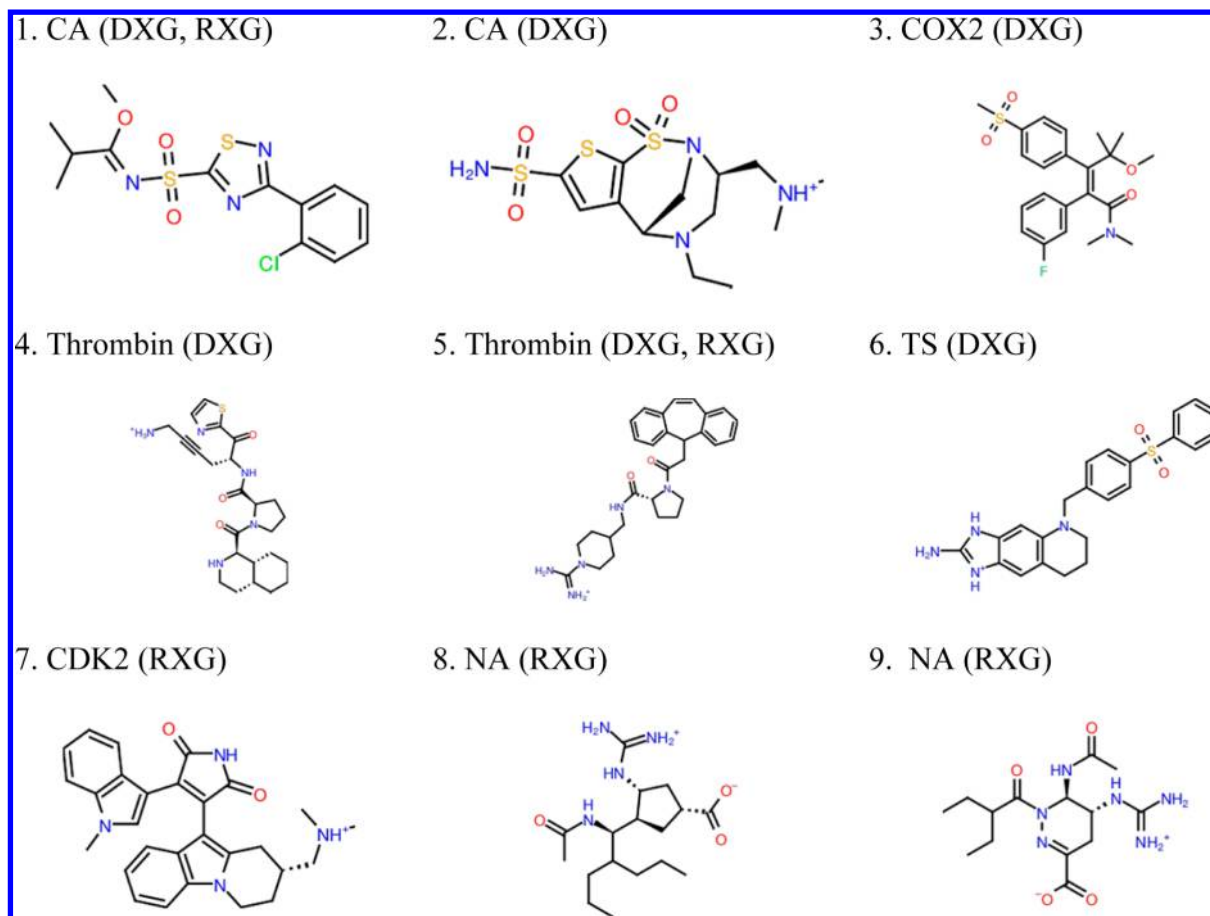
To compare the  $Z_2$  and  $Z_3$  scores with other data fusion approaches, we used five methods recently described by Svensson et al.<sup>66</sup> and two additional methods recently shown by Willett to perform well in virtual screening.<sup>36</sup> Parallel Selection retrieves the top compound from each method in turn until the desired number of compounds is reached. If a compound that would have been selected has already been selected by another method then the next compound from that method is chosen. Sum Score normalizes the scores for each method by dividing by the best score any compound acquires from that method, then adds the normalized scores. Sum Rank adds together the ranks from each of the different methods, thus no normalization or standardization of the scores is needed. Ties are broken by the Sum Score. Pareto Ranking scores a compound based on how many other compounds are better than it in all screening methods. Ties are broken using Sum Rank. Rank Vote

**Table 4.** Mean and Median EF(1%) Enrichment Values for the DUD and MDDR Sets Using the  $Z_2$  Data Fusion Method for Six Combinations of Fingerprints, Shape, and Docking

	DUD		MDDR	
	mean	median	mean	median
DCG	20.5	19.9	40.5	37.5
MCG	18.8	17.6	37.4	32.5
RCG	21.6	23.1	43.4	36.1
DXG	20.5	19.1	43.5	36.0
MXG	19.3	18.3	38.0	33.4
RXG	21.4	21.3	43.8	37.5

adds a vote each time a compound is scored within the top 1% of the screen and the final ranking is based on the number of votes each compound receives. Reciprocal Rank<sup>36</sup> sums the reciprocal of the rank for each compound in each screen. Finally, MaxScore<sup>67</sup> simply takes the best individual score from the three methods after appropriate normalization of the scores.

For the results presented in this work, each data fusion combination is abbreviated by the three-letter code FSG, where  $F \in [D, M, R]$  denotes the fingerprint type,  $S \in [C, X]$  denotes the shape query used in the shape screen, and G denotes Glide docking. D, M, and R designate dendritic, Molprint2D, and radial fingerprints, respectively. C and X designate the ConfGen and X-ray crystal shape, respectively, of the query ligand used for Shape calculations. For example, the combination DCG represents



**Figure 3.** Examples of active ligands ranked in the top 1% by the  $Z_2$  scoring but not by any of the individual methods.

dendritic fingerprints, shape screening with the ConfGen query conformation, and Glide HTVS docking. The combination MXG represents Molprint2D fingerprints, shape screening with the X-ray query conformation, and Glide HTVS docking.

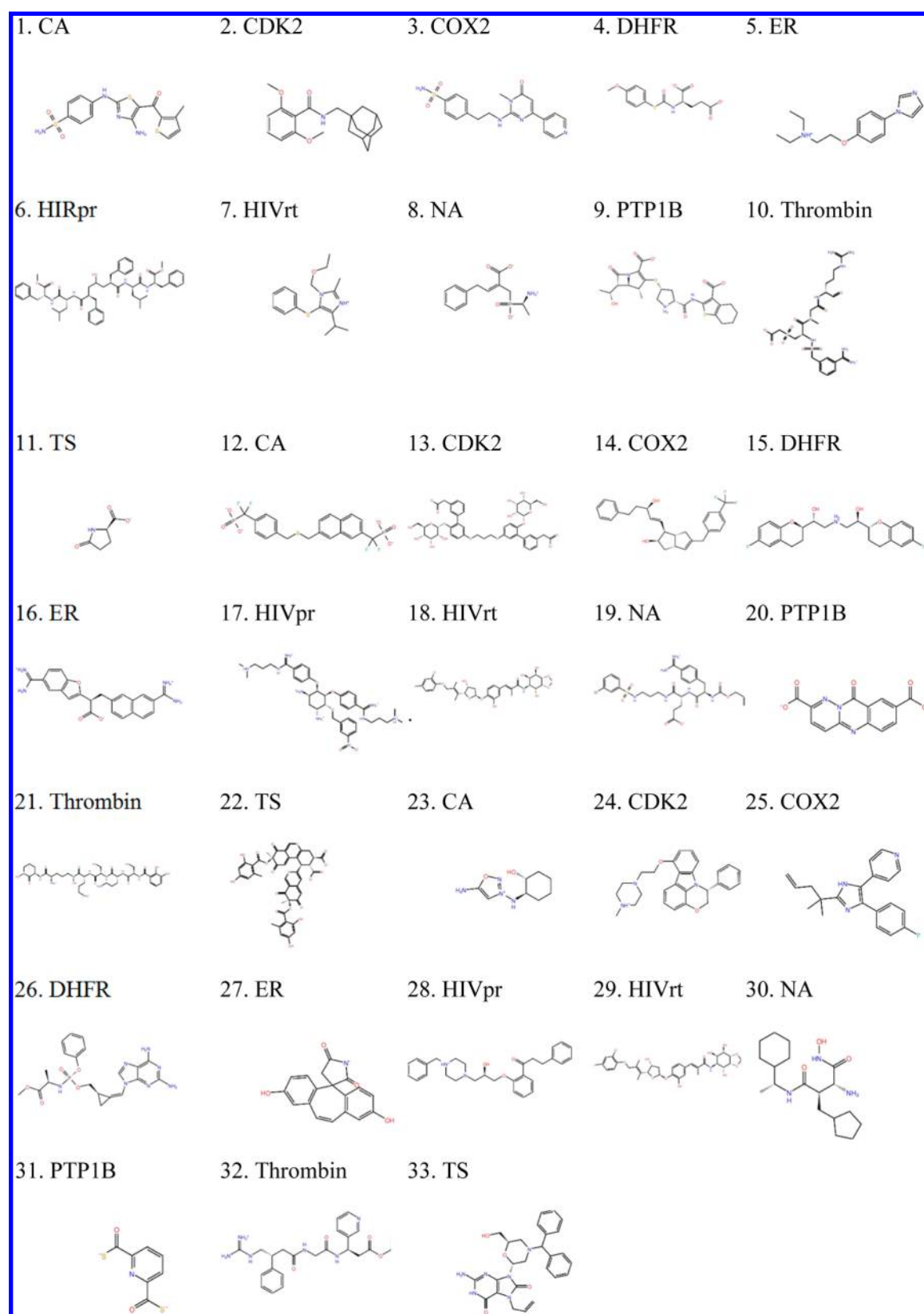
**Enrichment Calculations.** We report both the enrichment factor for the top 1% of the database (EF(1%)) and the Boltzmann-enhanced discrimination of the receiver operating characteristic (BEDROC).<sup>68</sup> We use  $\alpha = 20$  for the BEDROC calculations, which corresponds to 80% of the BEDROC score being accounted for in the top 8% of the database screen. Although BEDROC is a desirable enrichment metric for various reasons (see discussions by Truchon and Bayly<sup>68</sup>), for some of this work we use the more common EF(1%) metric, since it can be compared directly to results from other papers and is intuitive to understand. Both EF(1%) and BEDROC ( $\alpha = 20$ ) have the desirable characteristic that they emphasize the early part of the receiver operating characteristic (ROC) curve, which is important for real-world applications where only the early part of a database screen gets carried forward for experimental testing.

**Computational Times.** Glide HTVS calculations require an initial generation of receptor energy grids, which takes approximately 5 min per receptor structure. Docking calculations take approximately 1–2 s per ligand, which includes the generation of conformations during the docking calculations. The search speed for shape screening is approximately 500 pregenerated conformers per second. Running shape screening on a database of precomputed conformations with an average of 50 conformations per ligand took an average of about 0.1 s per ligand (i.e., 10 ligands per second). Canvas fingerprint similarity

searches took an average of about 0.0001 s per ligand (i.e., 10 000 ligands per second). All calculations were run on 2.4 GHz AMD Opteron processors.

## RESULTS AND DISCUSSION

We performed virtual screening of the 11 MDDR targets shown in Table 1 and the 40 DUD targets using each of the individual methods (three 2D fingerprints, two shapes, and one docking). As seen from the enrichment results in Table 2, the highest average enrichments for individual methods are obtained with the fingerprint methods (D, M, and R), followed closely by shape (C and X) and then docking (G). It should be noted that for docking only the HTVS mode of Glide was used here because it is the fastest, but the SP<sup>18,19</sup> and XP<sup>20</sup> modes tend to produce higher enrichments at the expense of speed. Looking at the enrichments for each of the individual targets (see the Supporting Information), there are apparent complementarities between the methods. For example, in the case of the MDDR, docking performs poorly on carbonic anhydrase (CA) whereas both fingerprints and shape perform very well. On the other hand, in the case of estrogen receptor (ER), both docking and shape perform well whereas fingerprints perform poorly. Similarly, for the DUD targets, docking performs poorly on ACE while fingerprints and shape perform well, while docking and shape perform better than any of the fingerprint methods for fXa and ER-agonist. Given this observation and analysis of the individual ligands that perform well with some methods and poorly with others, we set forth to determine whether the



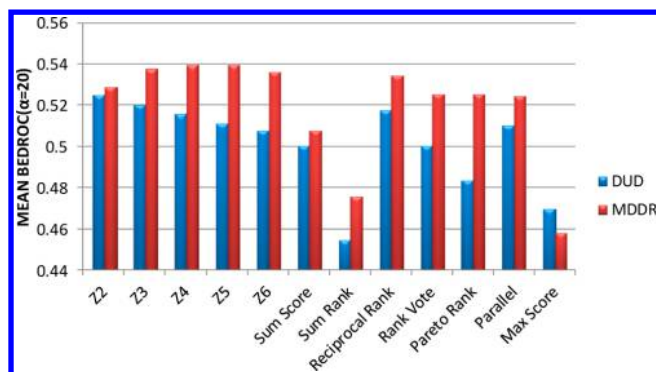
**Figure 4.** Decoys that score in the top 1% by one method but scored poorly with the  $Z_2$  score. Compounds 1–11 are scored well by fingerprints, 12–22 are scored well by docking, and 23–33 are scored well by shape.

individual methods could be combined to enhance the overall enrichment results.

The scores obtained from each of the methods (fingerprints, shape, and docking) were combined using various data fusion

approaches, as described in Materials and Methods. One fingerprint (D, M, or R) was combined with one shape (C or X) and HTVS Glide docking (G). Table 3 shows the  $Z_2$  (best two of three Z-scores) and  $Z_3$  (average of all three Z-scores) average





**Figure 5.** Data fusion algorithms applied to the full set of six virtual screening protocols used in this work (three fingerprints, two shape-based, and one docking). The algorithms based on Z-score outperform the other data fusion algorithms when a sufficient number of screening methods are included. In this case, all Z-score algorithms with at least three screening methods outperform the other data fusion algorithms.

enrichment results for the 11 MDDR targets and 40 DUD targets. We find that the  $Z_2$  score and  $Z_3$  score both perform better than any of the single methods shown in Table 2. In the cases where the  $Z_2$  score outperforms the  $Z_3$  score, active compounds tend to have two good scores but one bad score, thereby explaining the observed advantage of  $Z_2$  compared with  $Z_3$ . For example, a compound very similar to the query might have high 2D fingerprint and shape scores but a poor docking score, which would still produce a good  $Z_2$  score. Likewise, an active compound topologically very different from the query molecule could still receive good shape and docking scores,

thereby producing a better  $Z_2$  score than  $Z_3$  score. In analyzing the hit lists, we found that decoys seldom get good scores from more than one screening method, which is why the Z-score approach (and data fusion in general) outperforms the individual methods.

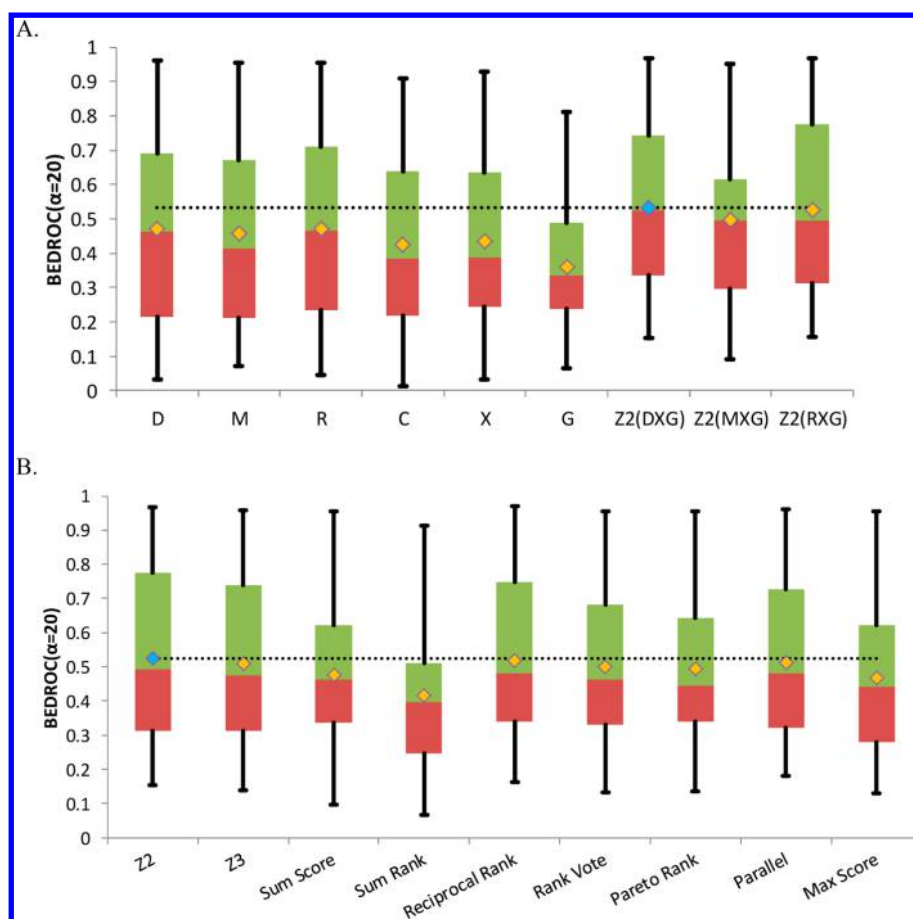
Figure 1a shows the average enrichment results across the 40 DUD targets for  $Z_2$ ,  $Z_3$ , the individual methods, and the average enrichment of the three methods. Figure 1b shows the  $Z_2$  results compared with seven other data fusion approaches. Parallel Selection, Pareto Rank, Rank Vote, and Reciprocal Rank all perform relatively well, with higher average enrichments than any of the individual methods. However, they perform worse than  $Z_2$  for all six scoring combinations shown in Figure 1b, with the exception of Rank Vote and Reciprocal Rank that perform better with the Molprint2D fingerprints. In the case of Sum Score and Max Score, the data fusion approach does not significantly outperform the single best screening method. Finally, Sum Rank has the poorest performance of the data fusion methods, significantly underperforming the best single screening method. On the basis of these findings, we shall discuss results from the  $Z_2$  data fusion method in the subsequent sections.

Figure 2 shows the distribution of  $Z_2$  scores for neuraminidase and HIV protease with representative active compounds and their  $Z_2$  score. The  $Z_2$  values approximate a normal distribution, although they have a peak to the left of zero and a positive skewness (i.e., a fat tail to the right). The fat tail, which extends well beyond the score representing 1% of the database (red dotted line in Figure 2), contains the compounds that are many standard deviations above the mean of the  $Z_2$  scores. The top three scoring compounds for both neuraminidase and HIV



**Figure 6.** BEDROC ( $\alpha = 20$ )  $p$ -value matrix between different screening approaches on the DUD and MDDR sets. Smaller values represent more statistical significance, with 0.05 representing the 95% confidence level. Cell coloring changes linearly from green (high statistically significant difference) to red (low statistically significant difference). Radial fingerprints, X-ray shape, and HTVS Glide (RXG) screening methods were used for data fusion methods (rows in the table above/before dendritic). The  $p$ -value corresponds to the significance that the method listed on the left is better than the method listed at the top of the matrix.





**Figure 7.** Box plots of BEDROC enrichment results from the 40 DUD targets. (A) Individual screening methods compared with the Z-score data fusion methods. (B) Z-score data fusion methods compared with other data fusion methods. The diamond symbols show the location of the mean value for each method. The diamond colored in blue and the associated horizontal black dashed line show the highest mean. The data fusion methods in B are from RXG screening methods.

protease, all of which are active, are also shown in Figure 2. All targets explored in this work follow a similar normal distribution (data not shown). Table 4 shows average and median EF(1%) enrichment results using the  $Z_2$  score for all targets. The average enrichment using the  $Z_2$  score across the 11 MDDR and the 40 DUD targets improves considerably over the best single method (compare with Table 2).

Interestingly, there were several active ligands that did not score in the top 1% for any of the individual methods but were in the top 1% with the  $Z_2$  score. These actives are generally structurally dissimilar from the input query compound. For example, active compound 2 of the CA screen with the DXG method is shown in Figure 3 and has a very different chemical topology from the query (refer to Table 1 for the query), with a central aliphatic fused heterocyclic ring system. However, the  $Z_2$  score ranks it within top 1% of the database because of the combination of moderately good fingerprint and shape scores (top 4% and 2%, respectively). Similarly, actives 4 and 5 of thrombin show less branching than the query and are missing the sulfonyl group in the query. While the fingerprint ranking for these compounds is poor, they still get good shape and docking scores (both in the top 2%), resulting in a combined  $Z_2$  score in the top 1%. Thymidylate synthase active 6 has small R-groups whereas the query has long linear polyglutamyl moieties, resulting in neither the fingerprint or shape scores in the top 1% (both in the top 2%), but the combined  $Z_2$  score is in the top 1%. Finally, CDK2 active 7 shows a different chemical topology

and ring system when compared to the smaller query. Shape and fingerprint scores are both in the top 2% but the combined  $Z_2$  score places it within the top 0.3%. Comparing the hits in Figure 3 with the query molecules in Table 1 shows that diverse actives can be found with the data fusion approach presented here.

Another driving factor in the improved enrichments is the cases where a decoy molecule scores very highly by one of the methods but the  $Z_2$  score produces a much worse rank for the decoy because it scores poorly with the other two methods. Figure 4 shows examples of decoys ranked very high from only one of the methods but not in the top 1% of the database based on  $Z_2$ . For example, decoys 1–11 in Figure 4 receive good fingerprint scores based on their high similarity to the query molecule (compare with the query molecules in Table 1), but shape and docking score these decoys poorly, resulting in poor  $Z_2$  scores for these decoys and an improved enrichment for these targets. Similarly, decoys 12–22 score well with docking but poorly with fingerprint and shape. Many of these compounds have multiple charged groups, which can be difficult to score well with a docking program but the fingerprint and shape dissimilarity from the query results in their not being ranked in the top 1% based on  $Z_2$ . Finally, decoys 23–33 in Figure 4 are examples where shape produces a good score but fingerprints and docking score the compounds poorly. Here, the compounds can adopt a similar shape to the query and even have some of the same chemical features, but the similarity is not high in terms of 2D fingerprints and the compounds do not dock well with Glide.

Finally, to test the potential value of including more virtual screening protocols, we applied the data fusion algorithms to all six screening results (three fingerprints, two shapes, and one docking) and show the result in Figure 5. Including more screening protocols improves the results for most of the data fusion algorithms on the MDDR data set, suggesting that the similarity-based methods have high predictive capabilities for this set. This might be expected, given the relatively high similarity among a large number of the actives in the MDDR. On the other hand, including more methods does not improve results for the DUD set, which has a more challenging set of active compounds. The average EF(1%) for the DUD set when combining all six screening methods ( $Z_6$ ) is 21.0, which is slightly lower than the best  $Z_2$  value of 21.4 obtained with just three screening protocols.

To assess the significance of the differences observed between the data fusion methods presented in this work, the  $p$ -value matrix, computed as described by Nicholls,<sup>69</sup> is shown in Figure 6 for the DUD and MDDR sets. Here, lower values indicate higher significance that the method listed on the left side of the table is better than the method on the top of the table. A value of 0.05 represents the 95% significance cutoff, 0.01 represents 99%, etc. As seen, the  $Z_2$  method is statistically better than all other data fusion methods for both the DUD and MDDR sets, with most cases showing significance at the 99% confidence level. The second best method for the DUD set is Reciprocal Rank and for the MDDR is  $Z_3$ . Finally, we show box plots of BEDROC enrichments from the 40 DUD targets in Figure 7 (A and B), again indicating the strong performance of the  $Z_2$  data fusion method.

## CONCLUSION

The data fusion virtual screening approach presented in this work has been shown to produce high enrichments and retrieve diverse actives. The  $Z_2$  score described here, which combines the best two of three  $Z$ -scores for a given compound, was shown to outperform any of the individual virtual screening protocols and other data fusion algorithms. The combination of three dissimilar virtual screening methods used in this work (2D fingerprints, shape, and docking) allows for the strengths of some approaches to complement the weaknesses of others. For example, an active compound that is highly similar to the query compound in terms of fingerprints and shape will get a good  $Z_2$  score even if docking does not score it well. The poor docking score could result from a variety of effects not included in the docking methodology, such as receptor movements that are not treated by the rigid receptor docking approach or other limitations within the docking method. Along the same lines, if an active compound gets a very good docking score and shape score but has a highly dissimilar fingerprint from the query molecule, the  $Z_2$  score will still be high.

The successful application of data fusion in this work to diverse targets and actives suggests that this approach should be applicable in virtual screening campaigns on other targets. Nonetheless, there are several potential limitations of this approach and opportunities for improvements. First is the assumption that the scores for each method are normally distributed, which is the basis for the  $Z$ -score calculations. While scores are indeed normally distributed in the cases studied here, it is possible that other methods or targets will not behave in the same way and the implications of a strong deviation from a normal distribution are not known in the context of the  $Z_2$  score. Next, while the average results show significant improvements in enrichment using data fusion, there are individual targets that perform better with a

single method, especially when the active compounds are highly similar to the query. In addition, data fusion results will improve as each of the individual methods improves. In the work presented here, we used the best fingerprint and shape methods in the Schrödinger Suite. However, for docking we only used the fastest mode of Glide (HTVS), which is not the most accurate Glide docking method. In addition to more ligand sampling, improvements in docking can come from including receptor sampling,<sup>70,71</sup> ensemble docking,<sup>72,73</sup> and accounting for waters.<sup>74,75</sup> Finally, the work here only considered targets for which crystal structures were available. While one could use data fusion with the ligand-based methods for targets without crystal structures, recent results suggest that high enrichments can be achieved with docking to a homology model when the models are sufficiently refined.<sup>76</sup> In this work we have not explored virtual screening and data fusion when docking to a homology model. Further improvements and generalizations of the  $Z$ -score data fusion algorithm will be realized with future work exploring the inclusion of more virtual screening approaches.

## ASSOCIATED CONTENT

### Supporting Information

Enrichment results for each target in the MDDR and DUD sets. The enrichments using each of the data fusion methods are provided in addition to the enrichments for each of the individual screening methods. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [woody.sherman@schrodinger.com](mailto:woody.sherman@schrodinger.com). Phone: +1 212 295 5800. Fax: +1 212 295 5801.

### Notes

The authors are employed by Schrödinger, Inc., which has developed Maestro, Impact, Ligprep, Epik, Canvas, Shape Screening, ConfGen, and Glide.

## REFERENCES

- (1) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing pitfalls in virtual screening: A critical review. *J. Chem. Inf. Model.* **2012**, *52*, 867–881.
- (2) Tikhonova, I. G.; Sum, C. S.; Neumann, S.; Engel, S.; Raaka, B. M.; Costanzi, S.; Gershengorn, M. C. Discovery of novel agonists and antagonists of the free fatty acid receptor 1 (ffar1) using virtual screening. *J. Med. Chem.* **2008**, *51*, 625–633.
- (3) Mizutani, M. Y.; Itai, A. Efficient method for high-throughput virtual screening based on flexible docking: Discovery of novel acetylcholinesterase inhibitors. *J. Med. Chem.* **2004**, *47*, 4818–4828.
- (4) Lu, Y.; Nikolovska-Coleska, Z.; Fang, X.; Gao, W.; Shangary, S.; Qiu, S.; Qin, D.; Wang, S. Discovery of a nanomolar inhibitor of the human murine double minute 2 (mdm2)-p53 interaction through an integrated, virtual database screening strategy. *J. Med. Chem.* **2006**, *49*, 3759–3762.
- (5) Grüneberg, S.; Stubbs, M. T.; Klebe, G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: Strategy and experimental confirmation. *J. Med. Chem.* **2002**, *45*, 3588–3602.
- (6) Trosset, J. Y.; Dalvit, C.; Knapp, S.; Fasolini, M.; Veronesi, M.; Mantegani, S.; Gianellini, L. M.; Catana, C.; Sundström, M.; Stouten, P. F. W. Inhibition of protein-protein interactions: The discovery of druglike  $\beta$ -catenin inhibitors by combining virtual and biophysical screening. *Proteins: Struct., Funct., Bioinf.* **2006**, *64*, 60–67.
- (7) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening

performance using eight fingerprint methods. *J. Mol. Graphics Modell.* **2010**, *29*, 157–170.

(8) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771–784.

(9) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189–1202.

(10) Tovar, A.; Eckert, H.; Bajorath, J. Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. *ChemMedChem* **2007**, *2*, 208–217.

(11) Wild, D.; Blankley, C. Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using ward's clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155–162.

(12) *The daylight toolkit*; Daylight Chemical Information Systems: Aliso Viejo, CA, 2008.

(13) Brown, R. D.; Martin, Y. C. The information content of 2D and 3d structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.

(14) Willett, P.; Winterman, V.; Bawden, D. Implementation of nonhierarchical cluster analysis methods in chemical information systems: Selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109–118.

(15) Sheridan, R. P. Finding multiactivity substructures by mining databases of drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1037–1050.

(16) 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.

(17) Klekota, J.; Roth, F. P.; Schreiber, S. L. Query chem: A google-powered web search combining text and chemical structures. *Bioinformatics* **2006**, *22*, 1670–1673.

(18) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750–1759.

(19) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(20) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.

(21) McGann, M. Fred pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578–596.

(22) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499–511.

(23) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: Applications of autodock. *J. Mol. Recognit.* **1996**, *9*, 1–5.

(24) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43–53.

(25) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(26) Totrov, M.; Abagyan, R. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins: Struct., Funct., Bioinf.* **1997**, *Suppl 1*, 215–20.

(27) Pierce, A. C.; Jacobs, M.; Stuver-Moody, C. Docking study yields four novel inhibitors of the protooncogene pim-1 kinase. *J. Med. Chem.* **2008**, *51*, 1972–1975.

(28) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K.

Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1b. *J. Med. Chem.* **2002**, *45*, 2213–2221.

(29) Tervo, A. J.; Kyrylenko, S.; Niskanen, P.; Salminen, A.; Leppänen, J.; Nyrönen, T. H.; Järvinen, T.; Poso, A. An in silico approach to discovering novel inhibitors of human sirtuin type 2. *J. Med. Chem.* **2004**, *47*, 6292–6298.

(30) Luzhkov, V. B.; Selisko, B.; Nordqvist, A.; Peyrane, F.; Decroly, E.; Alvarez, K.; Karlen, A.; Canard, B.; Åqvist, J. Virtual screening and bioassay study of novel inhibitors for dengue virus mrna cap (nucleoside-2'-o)-methyltransferase. *Bioorg. Med. Chem.* **2007**, *15*, 7795–7802.

(31) Dixon, S.; Smondryev, A.; Knoll, E.; Rao, S.; Shaw, D.; Friesner, R. Phase: A new engine for pharmacophore perception, 3d qsar model development, and 3d database screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647–671.

(32) Kurogi, Y.; Guner, O. F. Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr. Med. Chem.* **2001**, *8*, 1035–1055.

(33) Sastry, M.; Dixon, S.; Sherman, W. Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *J. Chem. Inf. Model.* **2011**, *51*, 2455–2466.

(34) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-d scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48*, 1489–1495.

(35) Ballester, P. J.; Finn, P. W.; Richards, W. G. Ultrafast shape recognition: Evaluating a new ligand-based virtual screening technology. *J. Mol. Graphics Modell.* **2009**, *27*, 836–845.

(36) Willett, P. Combination of similarity rankings using data fusion. *J. Chem. Inf. Model.* **2013**, *53*, 1–10.

(37) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How to optimize shape-based virtual screening: Choosing the right query and including chemical information. *J. Chem. Inf. Model.* **2009**, *49*, 678–692.

(38) Willett, P. Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR Comb. Sci.* **2006**, *25*, 1143–1152.

(39) Williams, C. Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Mol. Divers.* **2006**, *10*, 311–332.

(40) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100–5109.

(41) 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.

(42) Yang, J.-M.; Chen, Y.-F.; Shen, T.-W.; Kristal, B. S.; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134–1146.

(43) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422–1426.

(44) Feher, M. Consensus scoring for protein-ligand interactions. *Drug Discovery Today* **2006**, *11*, 421–428.

(45) Klon, A. E.; Glick, M.; Davies, J. W. Combination of a naive bayes classifier with consensus scoring improves enrichment of high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 4356–4359.

(46) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The use of consensus scoring in ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 277–288.

(47) Teramoto, R.; Fukunishi, H. Supervised consensus scoring for docking and virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 526–534.

(48) Salim, N.; Holliday, J.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* **2002**, *43*, 435–442.

(49) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New methods for ligand-based virtual screening: Use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* **2006**, *46*, 462–470.



- (50) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281–295.
- (51) Oda, A.; Tsuchida, K.; Takakura, T.; Yamaotsu, N.; Hirono, S. Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. *J. Chem. Inf. Model.* **2005**, *46*, 380–391.
- (52) Costanzi, S.; Tikhonova, I. G.; Ohno, M.; Roh, E. J.; Joshi, B. V.; Colson, A. O.; Houston, D.; Maddileti, S.; Harden, T. K.; Jacobson, K. A. P2y1 antagonists: Combining receptor-based modeling and qsar for a quantitative prediction of the biological activity based on consensus scoring. *J. Med. Chem.* **2007**, *50*, 3229–3241.
- (53) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941–948.
- (54) *Mdl drug data report*; Accelrys: San Diego, CA, 2005.
- (55) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
- (56) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
- (57) Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221–234.
- (58) Banks, J. L.; Beard, H. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. Integrated modeling program, applied chemical theory (IMPACT). *J. Comput. Chem.* **2005**, *26*, 1752–1780.
- (59) Jorgensen, W. L.; Tirado-Rives, J. The OPLS potential function for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.
- (60) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparametrization of the OPLS-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.
- (61) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field. *J. Chem. Theory Comput.* **2010**, *6*, 1509–1519.
- (62) *Ligprep v2.5*; Schrödinger, Inc.: Portland, OR, 2011.
- (63) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A software program for pk a prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681–691.
- (64) *Epik v2.2*; Schrödinger, Inc.: Portland, OR, 2011.
- (65) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. Confgen: A conformational search method for efficient generation of bioactive conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534–546.
- (66) Svensson, F.; Karlén, A.; Sköld, C. Virtual screening data fusion using both structure-and ligand-based methods. *J. Chem. Inf. Model.* **2011**, *52*, 225–232.
- (67) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of data fusion methods in virtual screening: Theoretical model. *J. Chem. Inf. Model.* **2006**, *46*, 2193–2205.
- (68) Truchon, J.-F.; Bayly, C. I. Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
- (69) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
- (70) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49*, 534–553.
- (71) Sherman, W.; Beard, H. S.; Farid, R. Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug Des.* **2006**, *67*, 83–84.
- (72) Osguthorpe, D. J.; Sherman, W.; Hagler, A. T. Generation of receptor structural ensembles for virtual screening using binding site shape analysis and clustering. *Chem. Biol. Drug Des.* **2012**, *80*, 182–193.
- (73) Osguthorpe, D. J.; Sherman, W.; Hagler, A. T. Exploring protein flexibility: Incorporating structural ensembles from crystal structures and simulation into virtual screening protocols. *J. Phys. Chem. B* **2012**, *116*, 6952–6959.
- (74) Repasky, M. P.; Murphy, R. B.; Banks, J. L.; Greenwood, J. R.; Tubert-Brohman, I.; Bhat, S.; Friesner, R. A. Docking performance of the Glide program as evaluated on the astex and dud datasets: A complete set of Glide sp results and selected results for a new scoring function integrating watermap and Glide. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 787–799.
- (75) Knegtel, R.; Robinson, D. D. A role for hydration in interleukin-2 inducible T cell kinase (itk) selectivity. *Mol. Infor.* **2011**, *30*, 950–959.
- (76) Pala, D.; Beuming, T.; Sherman, W.; Lodola, A.; Rivara, S.; Mor, M. Structure-based virtual screening of mt2 melatonin receptor: Influence of template choice and structural refinement. *J. Chem. Inf. Model.* **2013**, *53*, 821–835.