

## Considerations in Compound Database Preparation—"Hidden" Impact on Virtual Screening Results

Andrew J. S. Knox,<sup>†</sup> Mary J. Meegan,<sup>†</sup> Giorgio Carta,<sup>‡</sup> and David G. Lloyd<sup>\*,‡</sup>

School of Pharmacy and Pharmaceutical Sciences, Trinity College Dublin, Dublin 2, Ireland, and  
Molecular Design Group, School of Biochemistry and Immunology, Trinity College Dublin, Dublin 2, Ireland

Received May 5, 2005

Structure-based virtual screening (SBVS) utilizing docking algorithms has become an essential tool in the drug discovery process, and significant progress has been made in successfully applying the technique to a wide range of receptor targets. In silico validation of virtual screening protocols before application to a receptor target using a corporate or commercially available compound collection is key to establishing a successful process. Ultimately, retrieval of a set of active compounds from a database of inactives is required, and the metric of enrichment ( $E$ ) is habitually used to discern the quality of separation of the two. Numerous reports have addressed the performance of docking algorithms with regard to the quality of binding mode prediction and the issue of postprocessing "hit lists" of docked ligands. However, the impact of ligand database preprocessing has yet to be examined in the context of virtual screening and prioritization of compounds for biological evaluation. We provide an insight into the implications of cheminformatic preprocessing of a validation database of compounds where multiple protonated, tautomeric, stereochemical, and conformational states have been enumerated. Several commonly used methods for the generation of ligand conformations and conformational ensembles are examined, paired with an exhaustive rigid-body algorithm for the docking of different "multimeric" compound representations to the ligand binding site of the human estrogen receptor alpha. Chemgauss, a shapegaussian scoring function with intrinsic chemical knowledge, was combined with PLP as a consensus-scoring scheme to rank output from the docking protocol and enrichment rates calculated for each screen. The overheads of CPU consumption and the effect on relative database size (disk requirement) for each of the protocols employed are considered. Assessment of these parameters indicates that SBVS enrichments are highly dependent on the initial cheminformatic treatment(s) used in database construction. The interplay of SMILES representations, stereochemical information, protonation state enumeration, and ligand conformation ensembles are critical in achieving optimum enrichment rates in such screening.

### INTRODUCTION

A typical drug discovery research program involves the testing of every available compound in a corporate or commercial compound library using high-throughput biological screening techniques (HTS). Such an approach inevitably leads to high cost and large time scales.<sup>1</sup> One of the major problems emerging in the pharmaceutical industry is that biologically assaying many thousands of compounds for receptor affinity is unlikely to yield the potential number of druglike molecules that realistically exist in a random set, because the screening set is not enriched with compounds that have previously been categorized according to specific "druglike" parameters.

Evidence for this becomes apparent when we see that R&D production of new drugs has remained constant over the past number of years, with major pharmaceutical companies each launching roughly one new drug per year.<sup>2</sup> A recent report also suggested that HTS has generated no lead compounds when used as a sole technique by a large number of major R&D companies.<sup>3</sup>

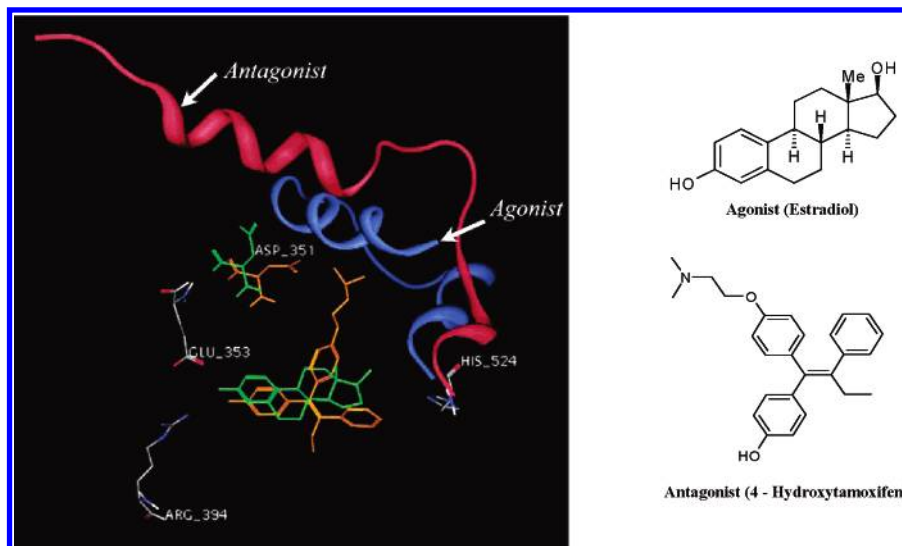
An increase in the number of highly resolved X-ray crystal structures of pharmaceutically relevant biological targets<sup>4</sup> has prompted the use and development of computational techniques to predict ligand affinity to such macromolecules.<sup>5,6</sup> Structure-based virtual screening (SBVS), the most prominent computational technique employed, involves the use of a docking program to generate ligand poses in the active site of a receptor and identification of the optimally docked pose using a scoring function.<sup>7</sup> The scoring function should reflect the complementarity or binding affinity of the ligand for the receptor.<sup>8,9</sup> In the context of a virtual screen, this method should discriminate compounds that bind to the receptor from nonbinders and, where possible, subsequently rank the binders according to their potency.

Virtual screening of compound libraries against therapeutic targets for a particular disease state has become integrated in the drug discovery process and provides a low-cost, rapid, and effective method of enriching a random compound library with the possibility of identifying active species directly. This technique has been applied successfully to compound libraries against a number of targets using various docking programs.<sup>10–12</sup> It is still at an early stage in its development, and the improvement of sampling methods and scoring functions will undoubtedly advance both the reli-

\* To whom correspondence should be addressed. Phone: +353-1-6082904. Fax: +353-1-677240. E-mail lloydg@tcd.ie.

<sup>†</sup> School of Pharmacy and Pharmaceutical Sciences.

<sup>‡</sup> Molecular Design Group, School of Biochemistry and Immunology.



**Figure 1.** Binding of agonist (green) and antagonist (orange) induces different helix-12 conformations.

ability and the efficacy of the technique. The significance of these factors will manifest themselves in the form of increased differentiation between active and inactive compounds in a corporate compound collection. The impact of ligand database preprocessing prior to SBVS on enrichment (*E*) has yet to be examined.

In this study, we seek to examine the effects of preprocessing on the prioritization of known active ligands from a database containing both known actives and inactives, where protonation, tautomeric, stereochemical, and conformational states are represented. Research in our group is focused on the identification of novel modulators of human nuclear hormone receptors.<sup>13–17</sup> In this light, we have applied a virtual screening protocol to a validation target of therapeutic importance, estrogen receptor (ER) alpha, where several different preprocessing techniques were used to generate the database of ligands to be screened.

The ER alpha is a nuclear hormone receptor<sup>18</sup> with a buried lipophilic binding site where liganding is highly dependent on hydrogen bonds as well as lipophilic contacts. The binding site is enclosed upon ligand binding, and liganding results in reorganization of the receptor. In particular, helix-12 encapsulates the receptor if an agonist (estradiol) is bound but is prevented from attaining this orientation when an antagonist (4-hydroxytamoxifen) is bound,<sup>19</sup> as in Figure 1. The main differences between the antagonist 4-hydroxytamoxifen and the endogenous agonist ligand (estradiol) are that the antagonist lacks a second hydroxyl group, which prevents hydrogen bonding with His524, but has an extended side chain to accommodate additional interaction with Asp351. The large amount of crystallographic data available and our understanding of the mechanism of action make the ER a viable and therapeutically important target for virtual screening. More specifically, estrogens are mitogenic for ER-positive breast cancer cells, and as 50% of primary breast cancers contain ER alpha,<sup>20</sup> we deemed this to be an important target for application to the optimization of virtual screening approaches.

This study is split into three main stages, where stage 1 involved the assessment of the effect on the prioritization of the actives from the compound collection where protonation, tautomeric, stereochemical, and conformational states

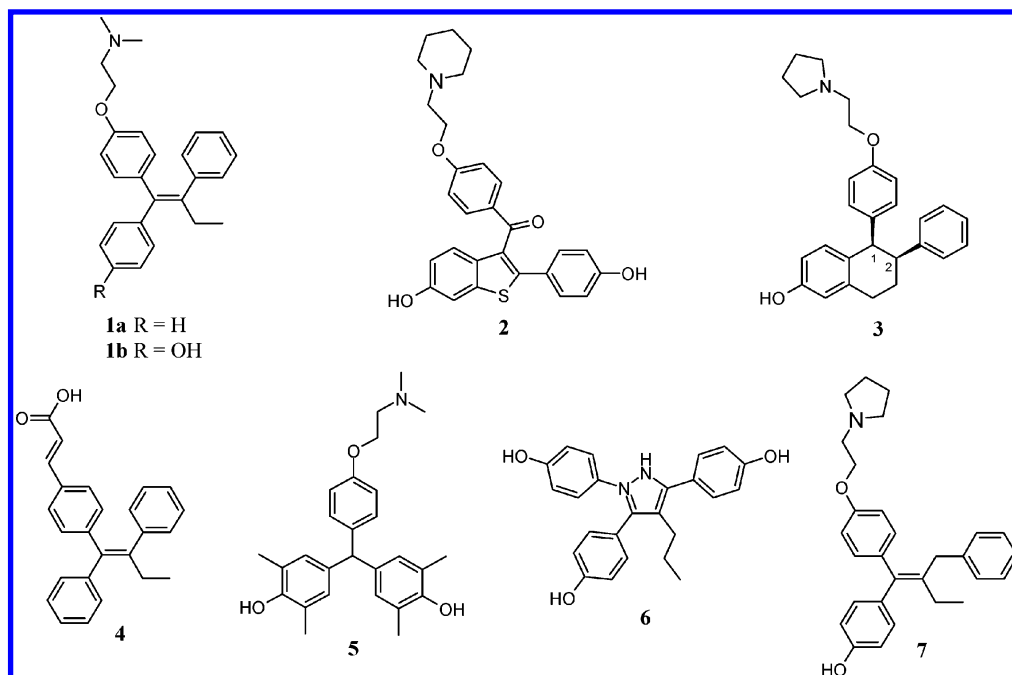
are enumerated. To quantify these effects, we report *E* values for each level of preprocessing where a compound database consisting of 1000 compounds (haystack) seeded with 40 ER known actives (needles) was utilized as input. Figure 2 depicts an indicative portion of several of the ER-alpha-active ligands that were included in the needle set.

The second stage consisted of a more rigorous test using 10 000 compounds (claimed inactive, with disclosed biological data) seeded with a single known active, in this case, a previously disclosed potent flexible antagonist,<sup>21</sup> illustrated in Figure 2. Using the same preprocessing protocols, the impact on ranking of the known active was assessed. Application of these processes using a training database allows one to effectively calibrate the docking and scoring protocol prior to deployment on a large corporate or commercial compound library. We also determined the quality of the binding poses produced under each protocol compared with those of the crystal structure 3ERT.<sup>22</sup>

Finally, the third stage divulged an analysis of the same 10 000 compounds seeded with the set of 40 actives used in stage 1. The impact of preprocessing here was assessed through the calculation of false positive (FP) rates for 50% of the true positives.

Why is it necessary to preprocess a database prior to virtual screening? Upon expansion of a database of 2D structures to one of 3D structures, accounting for hydrogen-bond donor and acceptor capabilities is imperative, as changes in the positions of hydrogens in the 2D format denotes conformational changes in the 3D format.<sup>23</sup> Accurate representation of the correct tautomeric and protonated forms of a compound, depending on the ultimate physiological environment of a compound, is extremely important in this case. It is computationally expensive and difficult to assign the most probable state of a compound; thus, all relevant states are typically enumerated as a representation. Similarly, it is necessary to generate all physically relevant stereoisomers of a compound arising from ambiguous stereocenter descriptions.

Conformational changes must also be accounted for as ligands rarely bind to a receptor in their lowest energy form and usually experience some strain upon binding that induces an increase in the energy of the ligand.<sup>24</sup> Computational



**Figure 2.** (1a) Tamoxifen, (1b) 4-hydroxytamoxifen, (2) raloxifene, (3) lasofoxifene, (4) GW5638, (5) sumimato biphenol, (6) pyrazole antagonist, and (7) flexible antagonist.

methodologies that account for ligand flexibility can be divided into two types. First, generation of conformational ensembles prior to a docking experiment, or “on the fly”,<sup>25</sup> can account for conformational changes upon ligand binding. Second, and less widely used, an evolutionary algorithmic approach is taken where core fragments of molecules are “grown” in the binding site of a receptor.<sup>26</sup> The former method of ensemble generation is addressed in this study. Conformer ensemble generation is also shown to be imperative in achieving optimum enrichment for the ER.

For the docking algorithm, this is CPU costly, as multiple forms of the same ligand are docked; however, this overhead is usually only a fraction of the time needed to predict the individual most dominant and relevant state of a molecule using software modules currently available. To further reduce docking CPU time, we chose to use a docking tool with extremely rapid docking times.<sup>27</sup> FRED (Fast Rigid Exhaustive Docking)<sup>28</sup> from OpenEye Scientific was used in combination with a consensus scoring scheme, Chemgauss and PLP, to prioritize the compounds in our docking studies in the ER. The use of FRED integrated with several scoring functions has been previously reviewed in screening experiments for seven targets with different active site characteristics, for example, lipophilic buried cavities, intermediate polarity, and very polar solvent-exposed binding sites. Chemscore was found to be the most applicable general scoring function for SBVS.<sup>29</sup> We take this scoring a step further by implementing a consensus scoring scheme to rank the database postdocking, where a shape-based function with terms accounting for chemical potentials, most significantly hydrogen-bonding interactions, is combined with PLP, an empirical scoring function shown to correlate well with protein–ligand binding affinities. Preliminary docking and scoring studies in our laboratory have highlighted the efficacy of this method combined with FRED over other scoring methods when the ER is used as a target. This study also highlights the unanticipated impact on enrichment rate. Very

different in silico enrichments can be achieved depending on the initial SMILES string representation used.

## COMPUTATIONAL METHODS

**Preparation of Estrogen Receptor (ER) Alpha.** Protein target coordinates were extracted from the PDB entry corresponding to the crystal structure of the estrogen receptor complexed with 4-hydroxytamoxifen (3ERT).<sup>22</sup> Structural waters were removed from the monomer, and Macromodel 6.5<sup>30</sup> was used to subsequently re-establish the correct connections in the PDB file. MOE<sup>31</sup> was used to add hydrogens to the protein, and a minimization protocol using a MMFF94 force field was implemented to adjust the positions of hydrogens and keep the heavy atoms fixed at their respective crystallographic positions. The complexed ligand was extracted and utilized as the search box for docking in the receptor. Protein and ligand files were saved in the mol2 format using MOE.

**Preparation of Validation Set (1000).** Verdonk et al. recently described “virtual enrichments” achieved using decoy sets that are dissimilar from the active set.<sup>32</sup> Here, we implement the same strategy of a “focused” decoy set with similar properties to those of the active set using a validation set (haystack) of 1000 compounds. The haystack was built as follows.

A subset of the Derwent World Drug Index (WDI)<sup>33</sup> was selected using Lipinski’s rules,<sup>34</sup> by removing compounds with intrinsically nondruglike properties such as those with molecular weights < 200 or > 550, the number of hydrogen bond donors  $0 < x < 6$  and acceptors  $0 < x < 10$ , and calculated log *P* values < 7, using an MCL script implemented in the Daylight toolkit.<sup>35</sup> Additional compound filtering was carried out with FILTER.<sup>36</sup> A Perl script was used to select a random subset of compounds from this filtered data set. Over half of all known marketed drugs contain chiral centers, and it was deemed of importance to



represent this in the data set. To this end, 500 molecules with their respective specific active chiral and isomeric data were taken from the WDI using the Daylight toolkit. Subsequently, 460 molecules whose active chirality and isomers were unspecified or “ambiguous” were also selected and added to the data set. This prevented any sources of imbalanced results where the decoy compounds are not representative of active species. The data set was retained in SMILES format.

A set comprising 40 ligands (needles) active for ER alpha was selected from the literature where binding and antiproliferative data were experimentally determined, with activities ranging from nanomolar to low micromolar potency. This active ligand set was then added to the validation set to bring the total number of compounds up to 1000.

**Preparation of Stages 2 and 3 Decoy Sets (10 000).** To further, more comprehensively test the protocols, a decoy database of 9999 inactive compounds was built from the WDI and CHEMBANK<sup>37</sup> and seeded with a single potent flexible estrogen alpha antagonist<sup>21</sup> (shown in Figure 2) to bring the total database size to 10 000 ligands for stage 2. In the generation of this dataset, all known estrogen actives were excluded, and as with the preparation of the 1000 ligand dataset, all compounds with nondruglike properties were removed using MCL scripting in Daylight and the application of FILTER. As before, a Perl script was used to randomly select the final 9999 compounds from a larger filtered population before seeding with the active ligand. The data set was stored in SMILES format.

Also, a database of 10 000 compounds combining the decoy set from stage 2 and the 40 active ligands (needles) for ER alpha used in stage 1 was prepared for stage 3. Similarly, the procedure was as above for stage 2. Stage 3 was carried out to make sure that the results obtained from stage 2 were not biased because of the incorporation of only a single active antiestrogen.

**Preprocessing of Validation Set. Generation of SMILES.** The Daylight toolkit<sup>35</sup> was used to export a data set of actives in tab format, allowing retention of the correct assignment of isomeric where specified and “ambiguous” SMILES where stereochemistry was not defined in the compound records. All structures were stored in SMILES format using two methods. MOL2SMI (Daylight Toolkit) and CONVERT (Molecular Networks GmbH)<sup>38</sup> were used to produce two alternate SMILES representations, A and B, respectively.

**Generation of Tautomeric and Protonated States.** The utility TAUTOMER (Molecular Networks GmbH)<sup>39</sup> was used to generate relevant tautomeric states of each molecule in the database. Conversion of SMILES strings to SDF format using UNITY<sup>40</sup> was necessary as strings lose their stereochemical information through canonicalization. To preserve the effects of input SMILES formatting differences, tautomatically processed SDF files were reconverted to SMILES strings using either MOL2SMI or CONVERT as required. This procedure was repeated using TAUTOMER (Openeye Scientific Software)<sup>41</sup> to facilitate a direct comparison of two commercially available and widely used tautomer generators.

The computational utility QUACPAC<sup>42</sup> was used to enumerate physiologically relevant protonation states of the validation set. Again, to preserve stereochemical information,

SDF files were used as input and reconverted to SMILES strings using either MOL2SMI or CONVERT as required.

Options to limit enumeration to a specified maximum number of protonated and tautomeric states are possible using both QUACPAC and TAUTOMER; however, for the purposes of this study, all calculable protonation and tautomeric states were enumerated in the pH range 2–14.

**Generation of Stereoisomers.** Different conformations exist for enantiomers, and it is necessary to manifest this molecular conformational space in a virtual screen as compound libraries often have inadequate stereochemical information denoted. STERGEN<sup>43</sup> identified ligand stereocenters and generated a set of isomeric structures where none was explicitly specified. This step enumerates only the multiple possible stereocenters for “ambiguous” SMILES strings in the validation set as the actives and approximately 50% of the selected haystack strings had explicit (correct) stereochemistry defined. As before, MOL2SMI and CONVERT were used to produce the alternate sets of SMILES strings following stereoisomer generation. This procedure was repeated using FLIPPER<sup>44</sup> as an alternative stereochemistry tool to facilitate a direct comparison of two widely used stereochemical generators.

**Conformer Generation.** To account for the fact that a molecule can adopt several 3D conformations by rotation about single and acyclic bonds, four 2D to 3D conformer generators were considered in this study, CORINA,<sup>45</sup> OMEGA,<sup>46</sup> RUBICON,<sup>47</sup> and CATALYST.<sup>48</sup>

In all cases, with the exception of CORINA, a single-conformer database and a multiple-conformer database (10 conformers) were generated. In the case of CORINA, generation of only one true conformer (when one discounts ring-flipping variants) is possible and, thus, OMEGA was subsequently used to expand the data to 10 conformers from the original input conformer passed by CORINA.

CORINA uses monocentric fragments with standard bond lengths, angles, and dihedral angles to form a 3D representation of a molecule. Sadowski has shown that CORINA reproduced the correct conformation of bound ligands for almost half of a data set of 639 X-ray structures.<sup>49</sup> OMEGA uses a torsion-driving beam rule-based method to generate conformational ensembles. A SMILES string is reduced to fragments with rotatable bonds, and rules are then applied to regenerate the ensembles. Application of the MMFF force field to refine input geometries allows any high-energy constructs to be minimized. RUBICON uses distance-geometry methods to randomly sample conformations. A rule-based method for establishing geometric constraints based on SMARTS<sup>50</sup> is utilized. CATALYST employs two methods of conformer sampling using a poling algorithm, FAST and BEST. CPU time is a contributing factor to the choice of preprocessing protocol in SBVS, so for the purpose of this study, the FAST option was chosen.

**Structure-Based Virtual Screening Protocol.** FRED 2.01<sup>28</sup> was utilized in this study to dock all preprocessed compound sets. FRED 2.01 uses a systematic, nonstochastic algorithm to ensure that reproducible results are attained. FRED rigidly and exhaustively examines all poses in an active site, filters by shape complementarity, then ranks by “fitness” prior to scoring using Gaussian functions which have chemical awareness incorporated (e.g., Chemgauss and Shapegauss). The final poses can be scored simultaneously

**Table 1.** Classification of Database Preprocessing Protocols Applied

level	preprocessing protocol
LEVEL1_X_A/B	SMILES – 1 conformer
LEVEL2_X_A/B	SMILES – 10 conformers
LEVEL3_X_A/B	SMILES – protonation – 1 conformer
LEVEL4_X_A/B	SMILES – protonation – 10 conformers
LEVEL5_X_A/B	SMILES – stereoisomers – 1 conformer
LEVEL6_X_A/B	SMILES – stereoisomers – 10 conformers
LEVEL7_X_A/B	SMILES – tautomers – 1 conformer
LEVEL8_X_A/B	SMILES – tautomers – 10 conformer

utilizing a number of scoring functions such as Shapegauss, PLP, Chemgauss, Chemscore, Screenscore, and Zapbind.

For this study, default operational values were applied and the docking of separately generated input conformers was enabled. Following rigid-body optimization of the ligands in the docking, ranking of the ligand poses using several scoring functions is possible. In internal validation studies using rigid-body docking algorithms and scoring with several scoring functions, either separately or as a consensus<sup>51</sup> scoring function, we have found Chemgauss and PLP<sup>52</sup> used as a consensus score to be the most efficacious scoring method for ranking the docked poses of a lipophilic binding site such as that of ER alpha.<sup>17</sup> The Chemgauss scoring function accounts most significantly for hydrogen-bond interactions. PLP scoring accounts for both simple and steric hydrogen-bond interactions. A recent report reviewed a set of screening experiments for seven targets with different active site characteristics, for example, lipophilic buried cavities, intermediate polarity, and very polar solvent-exposed binding sites. Chemscore emerged as the most applicable general scoring function for SBVS.<sup>27</sup> In previous studies using Chemscore as implemented with FRED 1.1, we have produced results in agreement with this report. However, in this work, we have chosen to utilize the most recent code release—FRED 2.01—where shape docking with a chemical knowledge function (Chemgauss) in combination with PLP delivers superior enrichment results using the same data sets in comparative trials.

**Computational Overheads—CPU Time Consumption and Database Size.** Despite steady drops in the cost of computational equipment, in parallel with increases in processing power, all computational SBVS experiments have associated overheads in terms of the time required for processing and the resultant physical database size produced. We therefore examine the time involved in producing the various “multimeric” databases examined, and their relative sizes.

**Stage 1. Impact of Preprocessing Levels on Enrichment Rate (1000 Compounds).** The two different representations of SMILES strings generated according to section 1 of the Preprocessing of Validation Set section above were used as the input for all subsequent enumerations of protonation, stereochemical, and tautomeric states shown as A (MOL2SMI) or B (CONVERT). In the presentation of the data, a qualifier “X” denotes which of four 2D–3D toolkits (CORINA, OMEGA, RUBICON, and CATALYST) was used for the validation set. Table 1 outlines the various preprocessing levels considered in this study. Each level was repeated for each of the SMILES generated; that is, LEVEL1–8\_X\_A/B is run for both MOL2SMI and CONVERT SMILES string representations, to furnish 64 individual protocols in total, when all four conversion tools are employed.

**Table 2.** Classification of 10 000 Compound Database Preprocessing Protocols Applied

level	protocol
9	single conformer
10	10 conformers
11	100 conformers
12	protonation + single conformer
13	protonation + 10 conformers
14	stereoisomers + single conformer
15	stereoisomers + 10 conformers
16	tautomers + single conformer
17	tautomers + 10 conformers

Following SBVS in each of the preprocessed databases outlined, enrichment rates for the first 0.5%, 1%, 1.5%, 2%, and 4% of the screen population were calculated. *E* indicates the ratio of the yield of actives in the hit list (post-screen ranked database population) relative to the random yield of actives as distributed throughout the unranked database and is calculated as follows:

$$E = \left( \frac{H_a}{H_t} \right) \frac{D}{A} = \frac{H_a}{H_t} \times \frac{D}{A}$$

where  $H_t$  is the total number of compounds in the hit list,  $H_a$  is the number of known actives in the hit list,  $A$  is the number of active compounds in the database, and  $D$  is the number of compounds in the database. Thus, if only five actives ( $H_a$ ) are retrieved in the first 1% of the ranked database ( $H_t$ ),  $D = 1000$ , and  $A = 40$  actives, the enrichment would be 12.5.

**Stage 2. Ranking of a Single Potent ER-Alpha Antagonist in a 10 000-Decoy Compound Set.** This set of 10 000 compounds was created to more stringently test the docking and scoring procedure utilizing the optimum levels of preprocessing identified from the above protocols as determined by their respective enrichment values in the 1000-ligand validation set. The efficacy of each protocol was determined according to the ability of each to prioritize the single active ligand contained in the data set. In each case, the protocols that achieved the highest enrichment rates for each level in experiments using the database of 1000 structures were employed and designated as levels 9–17. For example, if, in level 1 processing, SMILES generated using MOL2SMI and subsequent 3D conformers produced the optimum enrichment post docking, this protocol would then be used for single-conformer generation (denoted level 9) for the set of 10 000 compounds. A level of conformer generation producing 100 conformers of each compound in the data set (level 11) was also added to assess the effect of increased conformer generation. The classification of each of the database preprocessing protocols is outlined in Table 2.

**Stage 3. Ranking of a Diverse Set of ER-Alpha Antagonists in a 10 000-Decoy Compound Set.** To ensure that the results obtained from stage 2 reveal the full potential for variation of *E* rates, a decoy dataset of 9960 compounds “spiked” with the diverse set of 40 estrogen antagonists was utilized. This would account for any discrepancies that may be observed through diversity, as some of the antagonists

**Table 3.** Classification of 10 000 Compound Database Preprocessing Protocols Applied

level	protocol
18	single conformer
19	10 conformers
20	100 conformers
21	protonation + single conformer
22	protonation + 10 conformers
23	stereoisomers + single conformer
24	stereoisomers + 10 conformers
25	tautomers + single conformer
26	tautomers + 10 conformers

may not intrinsically have different protonation, tautomeric, and stereochemical states. The classification of each of the database preprocessing protocols is outlined as per Table 3. The efficacy of each protocol was measured by assessing false positive (FP) rates for 50% of the true positives.

## RESULTS AND DISCUSSION

**Computational Overheads.** The dependence of 2D to 3D conversion rates on the nature of input SMILES strings passed to the conversion tools is illustrated in parts a and b of Figure 3. An overview of the CPU time used and the conversion rate achieved using each conformer generation program where a single conformer is constructed from the SMILES strings in the validation set is provided. All programs were run on 32-bit Linux (Fedora) architecture with a 3.00 GHz Intel Xeon CPU with 2 GB of RAM.

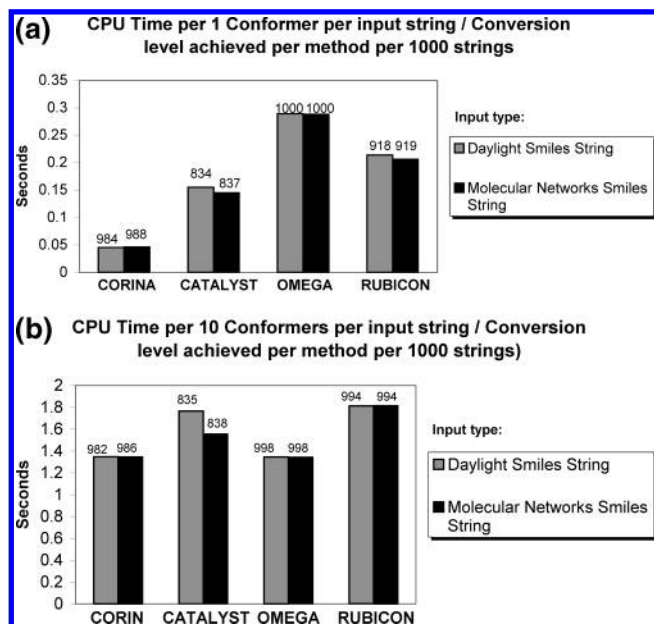
Upon examination of the data presented in Figure 3a, we find that CORINA converts the data set with a wall-clock time of <0.05 s/molecule, expending the least CPU time when processing MOL2SMI SMILES representations. However, ~1.6% of the data set remained unprocessed through conversion errors. The significance of this is apparent when

the data set is scaled to a more representative 1 million compounds—a database size employed for typical discovery screening. In such a set, up to 16 000 molecules could remain unprocessed, leading to potential “hits” being overlooked through this initial attrition. Similarly, CATALYST and RUBICON exhibit string-parsing errors, which could cumulatively impact on the quality of the database produced. OMEGA converts 100% of the data set in <0.29 s/molecule when input SMILES were generated using either MOL2SMI or CONVERT strings.

Good and Cheney have previously demonstrated that increasing sampling increases the chance of producing conformers closer to the bioactive conformer in a crystal structure.<sup>53</sup> Figure 3b shows conformational ensemble generation where 10 conformers of each molecule are produced. For CORINA and OMEGA, the conversion rates remain relatively unchanged, scaling as expected. With RUBICON, increased conformer sampling increases the performance of both conversion time and rate. CATALYST is seen to produce a single conformer per molecule in ~0.15 s from a SMILES string, but when the production of 10 conformers per molecule is undertaken, we observe an increase in processing time to 1.75 s per molecule. This is in line with all of the conversion tool rates. However, the low conversion rate observed when using CATALYST could mitigate against its incorporation as a large-scale preprocessing tool for database construction in SBVS.

**Stage 1. Effect of Preprocessing Levels on Enrichment Rate.** Table 4 depicts LEVEL1–8 of preprocessing using MOL2SMI and CONVERT SMILES string representations of the validation set of compounds. These data illustrate clearly the impact on the enrichment rate when protonated, tautomeric, stereochemical, and multiple conformations of the validation set are docked and scored, and they also illustrate the clear significance of using alternate SMILES strings as input for such SBVS protocols.

LEVEL1 and LEVEL2 depict “entry level” preprocessing, where consideration of either one (LEVEL1) or 10 (LEVEL2) conformers per ligand in the SBVS protocol is applied, in this instance, using both MOL2SMI- (X\_A) and CONVERT-generated (X\_B) SMILES input. In all cases, the enrichment is calculated using the number of molecules converted by each 2D–3D conversion program as input. We observe LEVEL1\_OMEGA\_A and LEVEL1\_OMEGA\_B outperforming other levels in LEVEL1. A large difference in enrichment rates can also be seen between the two alternate representations of SMILES strings. Screening a database prepared with LEVEL1\_OMEGA\_A, for instance, achieves the maximum possible enrichment until 1.5% and then reduces. LEVEL1\_OMEGA\_B, however, exhibits poorer enrichment rates in the same section of the database. The same trend is equally observed using RUBICON. Both RUBICON and CATALYST failed to convert ~17% and ~8.2% of the database, respectively, with lowered enrichment rates resulting. RUBICON conformer generation results in the lowest enrichment rate in the first 1% of the data set when only a single conformer is presented in the SBVS, but increasing the sampling rate, as with OMEGA and CORINA, considerably improves the outcome. Also, the conversion rates increase dramatically with RUBICON, converting 99.4% of the database. LEVEL2\_CORINA\_A and LEVEL2\_OMEGA\_A perform almost equally well



**Figure 3.** Graphical representation of 2-D SMILES string conversion from MOL2SMI (Daylight) SMILES string and CONVERT (Molecular Networks GmbH) SMILES string to 3-D molecules. Data labels over each column are equivalent to the number of molecules converted in the validation set. (a) When four conformer generation methods were used, one conformer was generated per molecule. (b) Conformational ensembles of 10 conformers were generated per molecule, for each of the four methods.



**Table 4.** Enrichment Results Obtained for Each Level (LEVEL1–8) of Preprocessing<sup>a</sup>

level	subset size %					database size
	0.5	1	1.5	2	4	
LEVEL1_CATALYST_A	20.85	20.85	19.46	16.68	9.38	834
LEVEL1_CATALYST_B	20.93	20.93	19.46	16.68	9.94	837
LEVEL1_CORINA_A	19.68	22.14	21.32	19.68	17.22	984
LEVEL1_CORINA_B	19.76	22.23	21.40	19.76	16.67	988
<b>LEVEL1_OMEGA_A</b>	<b>25.00</b>	<b>25.00</b>	<b>25.00</b>	<b>21.25</b>	<b>15.63</b>	<b>1000</b>
LEVEL1_OMEGA_B	25.00	22.50	21.66	18.75	11.88	1000
LEVEL1_RUBICON_A	18.36	18.36	19.89	17.21	11.48	918
LEVEL1_RUBICON_B	13.79	18.38	15.31	16.08	13.21	919
LEVEL2_CATALYST_A	20.88	20.88	19.48	18.78	14.09	835
LEVEL2_CATALYST_B	20.95	20.95	19.55	17.80	14.14	838
LEVEL2_CORINA_A	24.55	24.55	24.55	24.55	17.19	982
LEVEL2_CORINA_B	24.65	22.09	23.00	23.42	16.64	986
<b>LEVEL2_OMEGA_A</b>	<b>24.95</b>	<b>24.95</b>	<b>24.95</b>	<b>24.95</b>	<b>19.34</b>	<b>998</b>
LEVEL2_OMEGA_B	24.95	24.95	24.95	23.70	17.47	998
LEVEL2_RUBICON_A	24.85	22.37	21.54	21.12	16.77	994
LEVEL2_RUBICON_B	24.85	22.37	21.54	21.13	16.77	994
LEVEL3_CATALYST_A	20.65	18.59	15.14	14.46	9.80	826
LEVEL3_CATALYST_B	19.66	22.12	18.02	14.75	11.06	983
LEVEL3_CORINA_A	24.60	22.14	22.96	22.14	18.45	984
LEVEL3_CORINA_B	24.48	22.03	22.84	20.80	17.74	979
<b>LEVEL3_OMEGA_A</b>	<b>25.00</b>	<b>25.00</b>	<b>25.00</b>	<b>22.50</b>	<b>18.13</b>	<b>1000</b>
LEVEL3_OMEGA_B	25.00	25.00	23.33	23.75	14.38	1000
LEVEL3_RUBICON_A	19.40	19.40	21.02	20.61	15.76	970
LEVEL3_RUBICON_B	23.85	23.85	22.26	20.27	16.10	954
LEVEL4_CATALYST_A	20.65	20.65	19.27	17.55	13.42	826
LEVEL4_CATALYST_B	24.33	24.33	22.70	19.46	16.42	973
LEVEL4_CORINA_A	24.53	24.53	24.53	24.53	14.10	981
LEVEL4_CORINA_B	24.40	24.40	24.40	24.40	16.47	976
LEVEL4_OMEGA_A	24.93	24.93	24.93	23.68	15.58	997
<b>LEVEL4_OMEGA_B</b>	<b>24.93</b>	<b>24.93</b>	<b>24.93</b>	<b>23.68</b>	<b>16.20</b>	<b>997</b>
LEVEL4_RUBICON_A	24.85	22.37	23.19	22.37	16.77	994
LEVEL4_RUBICON_B	24.85	24.85	24.85	22.37	17.40	994
LEVEL5_CATALYST_A	20.65	20.65	19.27	17.55	11.36	826
LEVEL5_CATALYST_B	24.33	21.89	21.08	18.24	13.99	973
LEVEL5_CORINA_A	24.45	22.00	21.19	20.78	16.50	978
LEVEL5_CORINA_B	24.43	24.43	21.17	20.76	15.27	977
<b>LEVEL5_OMEGA_A</b>	<b>25.00</b>	<b>22.50</b>	<b>21.66</b>	<b>16.25</b>	<b>13.75</b>	<b>1000</b>
LEVEL5_OMEGA_B	20.00	22.50	21.66	16.25	13.75	1000
LEVEL5_RUBICON_A	18.44	18.44	18.44	16.14	13.25	922
LEVEL5_RUBICON_B	23.00	23.00	23.00	19.55	13.80	920
LEVEL6_CATALYST_A	20.65	20.65	20.65	19.62	12.90	826
LEVEL6_CATALYST_B	24.33	24.33	24.33	23.10	13.98	973
LEVEL6_CORINA_A	24.38	24.38	24.38	21.94	17.06	975
LEVEL6_CORINA_B	24.33	21.89	21.08	20.67	15.20	973
LEVEL6_OMEGA_A	24.95	24.95	23.29	23.70	17.47	998
<b>LEVEL6_OMEGA_B</b>	<b>24.95</b>	<b>24.95</b>	<b>24.95</b>	<b>23.70</b>	<b>16.84</b>	<b>998</b>
LEVEL6A_OMEGA_B	24.93	24.93	23.26	23.68	14.33	997
LEVEL6_RUBICON_A	24.60	22.14	22.96	20.91	15.38	984
LEVEL6_RUBICON_B	24.58	24.58	24.58	20.88	15.36	983
LEVEL7_CATALYST_A	22.08	22.08	19.13	16.56	10.49	883
LEVEL7_CATALYST_B	19.78	22.25	21.43	19.78	12.36	989
LEVEL7_CORINA_A	24.73	22.25	23.07	22.25	16.07	989
LEVEL7_CORINA_B	24.68	22.20	21.39	20.97	14.81	987
<b>LEVEL7_OMEGA_A</b>	<b>25.00</b>	<b>25.00</b>	<b>25.00</b>	<b>22.50</b>	<b>15.00</b>	<b>1000</b>
LEVEL7A_OMEGA_A	25.00	25.00	25.00	23.75	17.50	1000
LEVEL7_OMEGA_B	25.00	25.00	25.00	21.25	14.38	1000
LEVEL7_RUBICON_A	19.00	20.58	20.58	17.81	12.47	950
LEVEL7_RUBICON_B	23.40	21.06	18.72	16.38	10.53	936
LEVEL8_CATALYST_A	22.10	19.89	20.67	18.79	13.26	884
LEVEL8_CATALYST_B	24.75	24.75	24.75	21.04	14.23	990
LEVEL8_CORINA_A	24.95	24.95	24.95	23.70	15.59	998
LEVEL8_CORINA_B	24.95	24.95	24.95	22.46	14.35	998
LEVEL8_OMEGA_A	24.95	24.95	24.95	23.70	15.59	998
<b>LEVEL8_OMEGA_B</b>	<b>24.98</b>	<b>24.98</b>	<b>24.98</b>	<b>22.48</b>	<b>14.36</b>	<b>999</b>
LEVEL8_RUBICON_A	24.90	22.41	21.58	19.92	14.32	996
LEVEL8_RUBICON_B	24.93	24.93	23.26	21.19	13.71	997
<b>theoretical optimal value</b>	<b>25.00</b>	<b>25.00</b>	<b>25.00</b>	<b>25.00</b>	<b>25.00</b>	<b>1000</b>

<sup>a</sup> Optimum enrichments of each level obtained are highlighted in bold.

with enrichment rates up to 2% of 24.55 and 24.95, respectively.

LEVEL3 illustrates the impact on virtual screening arising from the introduction of a step to incorporate protonation states in the processing of the screening database. Single conformers generated by OMEGA contribute to achieving higher enrichment values and are enhanced by the addition of protonation, as seen in Table 4. Initial observation clearly shows that LEVEL3\_OMEGA\_A achieves a superior enrichment rate over the 4% of the database when compared with the others. LEVEL3\_RUBICON\_A produces the lowest *E*. Both CATALYST and RUBICON exhibit a large deviation in *E* between using two alternate SMILES representations as input. RUBICON generates higher *E* values using CONVERT SMILES strings.

A remarkable increase is observed using RUBICON where multiple conformers with a treatment of protonation are considered. LEVEL4\_RUBICON\_B accomplishes an *E* of 24.85, just slightly lower than that of LEVEL4\_OMEGA\_B. CORINA performs well also with MOL2SMI input giving an *E* of 24.53 across the first 2% of the ranked database. However, the opposite is seen with CATALYST at this stage in the process. CATALYST converts 97.3% of the database from CONVERT SMILES strings; however, only 82.6% is converted using MOL2SMI. Again, the effect of alternating SMILES string representations is highlighted here.

In general, protonation has a considerable influence on the orientation of a docked ligand, as the conformers produced by each 2D–3D tool will vary according to the positions of the hydrogens on a molecule. OMEGA has been previously shown to achieve significantly better results when the input structure is supplied from CORINA.<sup>54</sup> Additional refinement and minimization of the CORINA input structure under the MMFF force field before ensemble generation using OMEGA also enhances the performance. These concepts were integrated in the current version of OMEGA used in this study, and so, in general, the performance using OMEGA appears to be superior.

In the absence of specification of the chemical structure, the addition of arbitrary stereochemical information may introduce stereoisomers and regioisomers that may not actually exist in reality and impact on the enrichment rate achieved. STERGEN was used in the context of preserving SMILES strings with assigned stereochemical information and to assign multiple stereochemical representations to those ligands with partial or ambiguous information. The value of this is immediately apparent as typical commercial compound libraries often contain incomplete or no specific stereochemical information for a percentage of compound entries. LEVEL5 denotes how the consideration of stereochemical information assists Chemgauss and PLP in prioritizing the actives in the validation set only in the case of LEVEL5\_RUBICON\_B. Interestingly, this effect was more pronounced using FLIPPER (LEVEL6A\_OMEGA\_B), which we used to enumerate *all* possible stereoisomers for *all* ligands—without preserving the information of those with defined chirality. Accordingly, a slightly lower *E* is achieved as decoys were introduced to the docking and scoring procedure. There is clearly a fine balance which needs to be achieved in generating realistic stereochemical information for “ambiguous” structures in a data set and preserving the

known structural information down through the processing stages.

The introduction of a tautomer treatment to database preprocessing demonstrated that, in combination with a single conformer consideration, processed using LEVEL7\_OMEGA\_A/B, a higher enrichment could be achieved than when processing multiple conformers (LEVEL8\_OMEGA\_A/B). However, for all other tools, a benefit is observed on addition of multiple conformers to the database treatment. All possible tautomeric forms were enumerated, not solely the one considered to be most prevalent in solution at physiological pHs. The addition of tautomeric representations, for consideration in the SBVS protocol, makes the scoring function work harder but helps in more finely discriminating actives from inactives, as is demonstrated in LEVEL7\_OMEGA\_A/B. To facilitate a direct comparison of tools, LEVEL7A\_OMEGA\_A made use of an alternate processing utility, TAUTOMER (Openeye Scientific Software). A slight enrichment increase is observed over the 4% of the ranked data set using this tool. Tautomeric variation of a ligand also impacts on conformer orientation and, thus, the poses generated in a binding site during docking and prior to scoring. Although an additive effect on enrichment is observed by the addition of tautomers as with protonation, it is difficult to immediately establish if the highest scoring tautomer ranked is actually representative of the most prevalent species *in vivo*.

Finally, to ensure that the values in Table 4 for each level were statistically significant, we conducted an ANOVA two-way analysis of variance using Minitab 14.20. At a confidence level of 99%, a statistical difference was observed between programs within each level ( $p < 0.0001$ ). A Friedman two-way analysis of variance was also used to decide the optimum levels of preprocessing using the sum of ranks from enrichments observed in each level. We chose LEVEL5\_OMEGA\_A rather than LEVEL5\_OMEGA\_B as being superior because enrichment in the first 0.5% was 25 compared with 20, respectively. This is more important as when searching a larger database it is preferable to search the smallest ranked hit list possible (i.e., 0.5–1%).

**The Impact of Alternate SMILES Representations.** If alternate representations of SMILES strings are used as input 2D structures, such as those produced by MOL2SMI and CONVERT, radically different effects on the enrichment rates are achieved when the same preprocessing protocols are utilized in advance of rigid docking experiments. For example, LEVEL1\_RUBICON\_A and LEVEL1\_RUBICON\_B use the same preprocessing protocol of generating a single conformer, yet enrichment rates in the first 0.5% of the database were 18.36 and 13.79, respectively. Major enrichment differences can be seen throughout Table 4 because of variation in initial SMILES depictions. This effect is not restricted to any individual 2D–3D conformer generator studied, as is evident from LEVEL5\_OMEGA\_A/B, where enrichments are 25 and 20, respectively. These differences are possibly caused by the way in which each conformer generator parses a SMILES string. For example, programs that use a library of SMARTS strings to generate segments of a compound from a SMILES string would produce different conformers depending on the initial representation.

To emphasize the significance of this observation, we have endeavored to test six alternate SMILES string representations of the known ER-active modulator hydroxytamoxifen and compare the generated conformers with those of a set of six conformers produced from a *single* SMILES string representation of the ligand. The root mean square difference (RMSD) of each conformer generated was compared with the cocrystal structure of bound hydroxytamoxifen in the ER-active site and each conformer was also docked using FRED in the binding site of 3ERT to determine the best ranking conformer score. Table 5 shows the results.


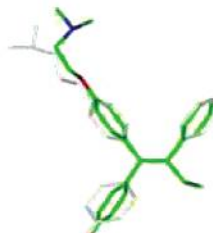
It is clear that different SMILES representations of the same molecule produce different conformers, and as we have shown in Table 3, this has a clear impact on enrichment values obtained also. Alternating the initial SMILES string produces a range of conformers, the best of which exhibits an RMSD of 0.75 Å when compared to the crystal structure of hydroxytamoxifen, and results in the highest scoring docked pose (–55.90) using FRED 2.01. The immediate conclusion to draw from these data is that, nominally, while each individual molecule is constant, not all SMILES representations are equal, and a hitherto unexplored link exists between this simplest cheminformatic treatment of molecular representations and the results obtained in 3D molecular recognition studies involving SBVS of this nature.

**Computational Overheads—How Big Is Too Big?** As previously discussed, each level of “multimeric” ligand treatment increases the physical size of the database under consideration. This size limit begins to impact on the feasibility and practicality of the SBVS when the base number of compounds is large. Figure 4 illustrates how the size of the data set using each protocol differs. This poses a significant quandary with respect to virtual screening. Database size must be reduced as much as possible to keep disk space at an affordable minimum, and also to prune the overall number of molecules to be screened from a performance cost/benefit perspective. If we consider the relative scaling for LEVEL8, 1000 compounds (a 312k SMILES file) expands to a 3D SDF file containing 27 212 multimeric forms (73 Mb). Scaling this up to a database of 1 000 000 screening compounds will accordingly generate 25–30 million ligands for consideration in both the preprocessing steps and the actual docking procedure. While the ability of newer software to utilize and read compressed data files may alleviate this difficulty, the issues of how large a data set one can afford to use, and correspondingly which preprocessing method one will employ, will ultimately vary according to available resources.

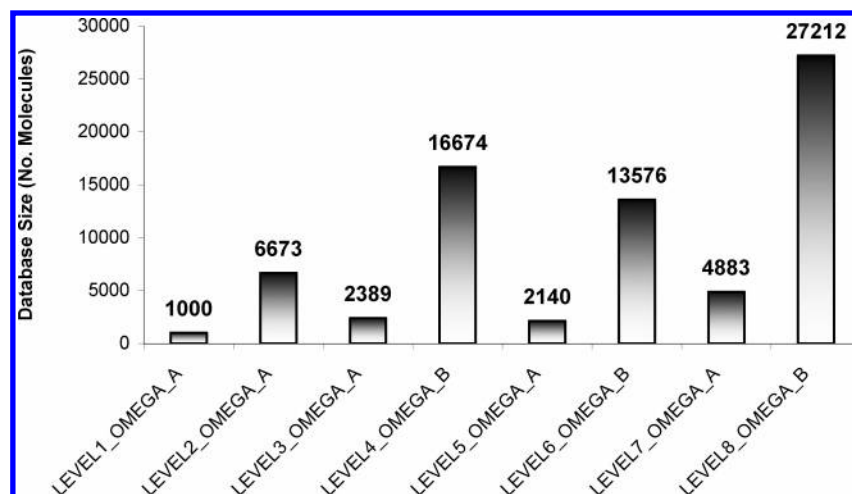
**Stage 2. Ranking of a Single Potent ER-Alpha Antagonist in a 10 000-Molecule Compound Set.** Tautomerism and protonation states have little synthetic consequence, although their consideration may be crucial in enhancing the enrichment level of a virtual screening protocol. A successful virtual screening protocol should not only prioritize compounds for biological testing but also provide useful information about the stereochemistry of the compound, should synthesis be required. We therefore sought to utilize from the previous steps protocols that provide the highest possible enrichment but also give the most information about the exact 3D or stereochemical nature of the compounds identified. The importance of this with respect to the estrogen receptor is clear when one considers that the *E* isomer of the ligand



**Table 5.** Comparison of RMSD of Alternate SMILES-Generated Conformers versus Conformer Generation Taken from a Single SMILES String<sup>a</sup>

6 SMILES permutations	RMSD	TOP SCORE PLP = -55.90
<chem>CC\C(c1ccccc1)=C(/c2ccc(O)cc2)c3ccc(OCCN(C)C)cc3</chem>	0.75	
<chem>CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc3</chem>	0.76	
<chem>CC\C(=C(/c1ccc(cc1)O)c1ccc(cc1)OCCN(C)C)c1ccccc1</chem>	1.07	
<chem>CN(C)CCO(c1ccc(cc1)C(/c2ccc(O)cc2)=C(c3ccccc3)\CC)</chem>	1.25	
<chem>CN(C)CCO(c1ccc(cc1)C(=C(c2ccccc2)\CC)\c3ccc(O)cc3)</chem>	0.77	
<chem>CN(C)CCOc(ccc(c1)C(/c(ccc(c2)O)c2)=C(c(cccc2)c2)\CC)c1</chem>	1.23	
6 SMILES of equivalent permutation	RMSD	TOP SCORE PLP = -54.91
<chem>CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc3</chem>	0.6	
<chem>CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc4</chem>	0.76	
<chem>CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc5</chem>	0.99	
<chem>CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc6</chem>	1.25	
<chem>CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc7</chem>	1.13	
<chem>CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc8</chem>	0.89	

<sup>a</sup> RMSD values were calculated using the OpenEye OEChem RMSD calculator, which fully accounts for automorphisms (self-symmetry of the molecules being compared). The crystal structure of hydroxytamoxifen (white) is superimposed with the conformer of hydroxytamoxifen (colored by atom) generated from a SMILES string with the lowest RMSD value. The PLP values shown correspond to the lowest RMSD structure docked in the active site of 3ERT.

**Figure 4.** Column data labels show the number of molecules produced from 1000 input SMILES strings using the optimum preprocessing techniques as previously evaluated.

tamoxifen exhibits estrogenic activity, while the Z isomer exhibits antiestrogenic potency. From our initial individual screening runs, it appears that the addition of stereochemical information to ensemble conformer generation provides the optimum benefit in pre- and post-screening, where the compound collection lacks defined stereochemical information.

To assess the utility of the various screening levels in a “real-world” application, our dataset of 10 000 ligands containing only a single known active was used. The impact of the various protocols on the ability of SBVS to identify the active is given in Table 6. Optimal ranking is achieved

using only 10 conformers, with no additional benefit seen when expanded to a treatment of 100 conformers per ligand. In these cases, all levels involved generation of conformers using OMEGA. Ligand conformer sampling is highly important in SBVS when the active or binding site is deemed to be flexible. The estrogen receptor exhibits a relatively rigid binding site upon examination. Nonetheless, slight variations in binding site residue positions occur, and a greater treatment of ligand flexibility is expected to improve the enrichment rate. Bostrom et al.<sup>54</sup> deemed 1000 conformations per molecule to be adequate sampling in the context of a virtual screen; while we agree that such a treatment is highly

**Table 6.** Ranking of a Single Active by FRED 2.01 from a Screening Database Totalling 10 000 Druglike Compounds Using Preprocessing Protocol Levels 9–17.

level	rank	protocol details
9	146	single conformer
10	<b>1</b>	10 conformers
11	<b>1</b>	100 conformers
12	399	protonation + single conformer
13	<b>6</b>	protonation + 10 conformers
14	254	stereoisomers + single conformer
15	<b>6</b>	stereoisomers + 10 conformers
16	163	tautomers + single conformer
17	<b>3</b>	tautomers + 10 conformers

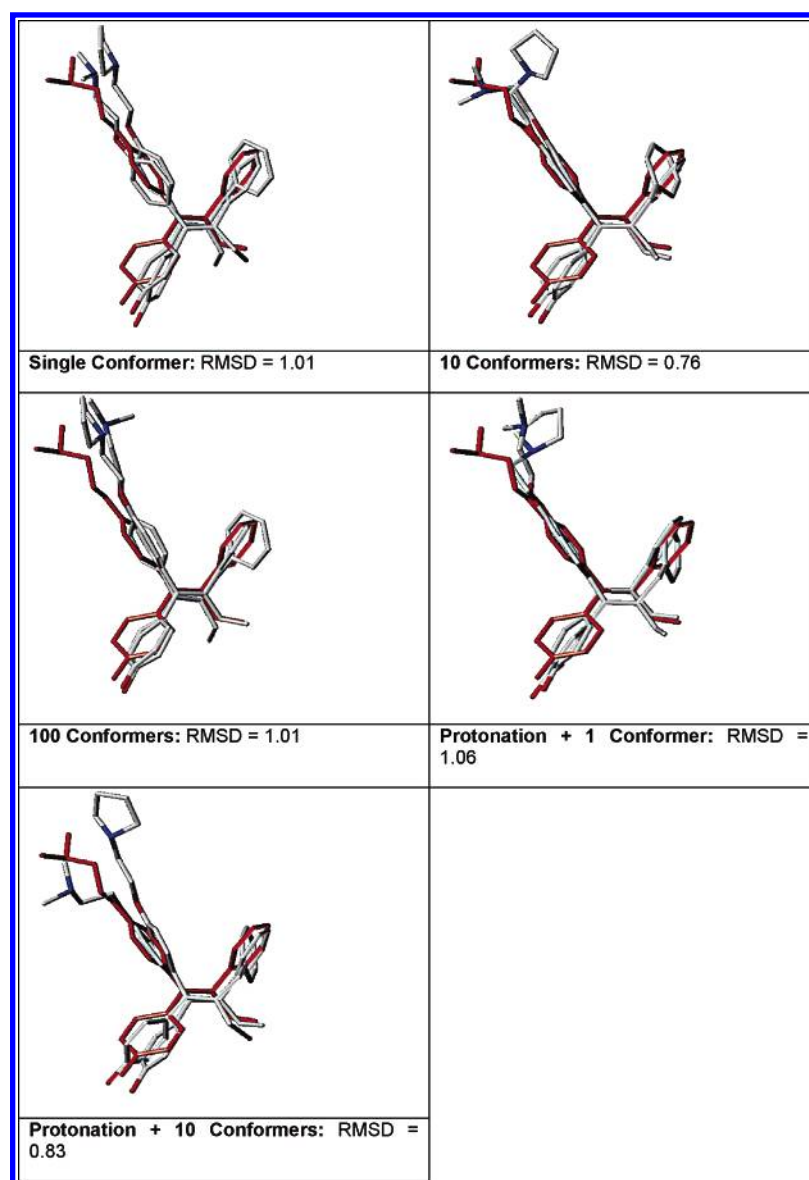
important when the target is flexible, no apparent benefit is seen for the estrogen receptor above 10 conformers. This emphasizes the requirement for SBVS protocol optimization on a target-by-target basis.

This comparative study highlights the wide impact on enrichment and variance possible when moving from single- to multiple-conformer treatments (level 9→10) and where protonation (level 9→12, level 10→13, and level 12→13),

chirality (level 9→14, level 10→15, and level 14→15), and tautomer treatments (level 9→16, level 10→17, and level 16→17) are explored in SBVS.

As a final comparison of the impact of preprocessing, we sought to elucidate the RMSD between docked solutions for hydroxytamoxifen-generated (colored by atom) conformers generated using each of the above preprocessing protocols and the ligand cocrystal structure pose (white) taken from 3ERT. All of the docked structures are shown to be close to the crystal structure of hydroxytamoxifen (Figure 5). The docked structure of the single potent antiestrogen used to seed the database of 10 000 structures is also overlaid (colored by atom) to show its equivalent docked solutions for each processing level. Interestingly, although a minimal difference is observed in the RMSD values, a large difference is seen in the relative rankings of the seed antiestrogen.

**Stage 3. FP Rate of 40 ER-Alpha Antagonists in a 10 000-Molecule Compound Set.** False positive rates for the ER-alpha antagonist set are outlined in Table 7. It is evident that the use of 10 conformers again gives the lowest

**Figure 5.** RMSD difference between docked conformers generated using each of the above preprocessing protocols and the ligand crystal structure taken from 3ERT.

**Table 7.** False Positive Rates for Recovery of 50% of the True Positives

level	FP 50%	protocol details
LEVEL18	0.44	single conformer
LEVEL19	0.35	10 conformers
LEVEL20	1.10	100 conformers
LEVEL21	0.58	protonation + single conformer
LEVEL22	2.30	protonation + 10 conformers
LEVEL23	0.65	stereoisomers + single conformer
LEVEL24	1.44	stereoisomers + 10 conformers
LEVEL25	1.87	tautomers + single conformers
LEVEL26	1.50	tautomers + 10 conformers

FP rate at 0.35% for a true positive rate of 50%, corroborating results from the previous section. The use of a more broad and diverse set of antagonists spiked within a large decoy set of 9960 enables us to definitively show the influence of each level of preprocessing on the database. We observe that the database conversion rate is also optimal using multiple conformers only. The addition of multiple conformers with protonation or stereochemical generation results in a higher FP rate and, thus, an increase in the number of random molecules ranked among the actives. The addition of tautomers also increases the FP rate.

## CONCLUSIONS

We have endeavored to elucidate the optimal preprocessing protocols and determine a method generic for the optimizing of enrichments in SBVS. Establishing a successful virtual screening process also requires that the techniques used must be CPU and resource “friendly”, while keeping database physical size to a minimum. This study is a snapshot of some of the most popular available techniques used in database preprocessing. All of the programs used were utilized with their default settings, and the adjustment of certain parameters may further enhance the enrichment rates achieved. The results presented here have important implications for those embarking on structure-based virtual screening experimentation. With increasing amounts of commercial and academic code available to researchers, the choice of preprocessing technique used to expand and represent a screening compound collection can and will have a significant impact on the performance of the virtual screen. Compound libraries are often represented in 2D format as SMILES strings and are converted to 3D format for the purpose of pharmacophore searching or structure-based virtual screening. SMILES strings can be constructed in a number of ways, and we have demonstrated clearly in this study that different representations have markedly different effects, not only on virtual screening enrichment rates achieved but also on the quality of the docked structures. If speed is a concern, then different tools are available with associated performance benefits, but not all methods will convert all input. A clear pattern is observed when using Daylight’s MOL2SMI SMILES strings, where optimum enrichment rates are achieved over those found using CONVERT string representations when compounds are enumerated as single conformations. Protonation, tautomerization, and the assignment of correct stereochemistry appears to have little benefit in this test case, but with respect to single conformers, an optimal enrichment can be achieved starting from a MOL2SMI string. SMILES strings generated from CONVERT require enumeration of multiple conformations to produce a high enrichment regardless of

whether they are protonated, tautomerized, or the stereochemistry is assigned.

Importantly, enrichment rates observed when using a smaller data set of 1000 compounds seeded with 40 actives show a marked difference to the ranking of a single active in 10 000. The best enrichment is achieved using OMEGA in combination with the propagation of 10 conformers per compound. No additional benefit is observed when using 100 conformers per compound with this particular receptor (ER), but this may not be the case when working with more flexible target systems. To allow a sufficient amount of synthetic information to be retrieved about each compound, a slight decrease in the ranked position of the active must be accepted where stereochemical information is added. However, a balance between the introduction of false positive results and the exact structural information needs to be achieved, and a cost/benefit assessment must be made for each target studied. A marked difference in the ranking ability of SBVS imbued by alternate preprocessing protocols is observed when using the larger test set. We therefore suggest the adoption of larger (“real-scale”) validation and training data sets as being more beneficial to those involved in training a docking procedure for the identification of active species in virtual screening. We are currently applying these findings to other virtual screening protocols optimized for the identification of novel modulators of the human estrogen receptors.

## ACKNOWLEDGMENT

We acknowledge the funding support from the Irish Health Research Board (HRB), the Higher Education Authority’s Program for Research in Third Level (PRTL) Institute for Information Technology and Advanced Computation (IITAC), and Science Foundation Ireland (SFI). We thank the software vendors for their continuing support of such academic research efforts, in particular, the contributions of Daylight Corporation and OpenEye Scientific, which significantly underpinned the work performed in this study.

## REFERENCES AND NOTES

- (1) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **2003**, 22 (2), 151–85.
- (2) Smith, A. Screening for drug discovery: The leading question. *Nature* **2002**, 418, 453–459.
- (3) Lahana, R. How many leads from HTS? *Drug Discovery Today* **1999**, 4 (10), 447–448.
- (4) Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z.; Gilliland, G.; Weissig, H.; Westbrook, J. The Protein Data Bank and the challenge of structural genomics. *Nat. Struct. Biol.* **2000**, 7 (Suppl.), 957–9.
- (5) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, 303 (5665), 1813–8.
- (6) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, 432 (7019), 862–5.
- (7) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, 7 (20), 1047–55.
- (8) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, 43 (25), 4759–67.
- (9) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, 44 (7), 1035–42.
- (10) 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 (11), 2213–21.
- (11) Schapira, M.; Abagyan, R.; Totrov, M. Nuclear hormone receptor targeted virtual screening. *J. Med. Chem.* **2003**, 46 (14), 3045–59.



- (12) Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* **2000**, *43* (3), 401–8.
- (13) Lloyd, D. G.; Hughes, R. B.; Zisterer, D. M.; Williams, D. C.; Fattorusso, C.; Catalanotti, B.; Campiani, G.; Meegan, M. J. Benzoxepin-derived estrogen receptor modulators: a novel molecular scaffold for the estrogen receptor. *J. Med. Chem.* **2004**, *47* (23), 5612–5.
- (14) Lloyd, D. G.; Buenemann, C. L.; Todorov, N. P.; Manallack, D. T.; Dean, P. M. Scaffold hopping in de novo design. Ligand generation in the absence of receptor information. *J. Med. Chem.* **2004**, *47* (3), 493–6.
- (15) Meegan, M. J.; Lloyd, D. G. Advances in the science of estrogen receptor modulation. *Curr. Med. Chem.* **2003**, *10* (3), 181–210.
- (16) Meegan, M. J.; Hughes, R. B.; Lloyd, D. G.; Williams, D. C.; Zisterer, D. M. Ethyl side-chain modifications in novel flexible antiestrogens—design, synthesis and biological efficacy in assay against the MCF-7 breast tumor cell line. *Anti-Cancer Drug Des.* **2001**, *16* (1), 57–69.
- (17) Meegan, M. J.; Hughes, R. B.; Lloyd, D. G.; Williams, D. C.; Zisterer, D. M. Flexible estrogen receptor modulators: design, synthesis, and antagonistic effects in human MCF-7 breast cancer cells. *J. Med. Chem.* **2001**, *44* (7), 1072–84.
- (18) MacGregor, J. I.; Jordan, V. C. Basic guide to the mechanisms of antiestrogen action. *Pharmacol. Rev.* **1998**, *50* (2), 151–96.
- (19) Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Ohman, L.; Greene, G. L.; Gustafsson, J. A.; Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **1997**, *389* (6652), 753–8.
- (20) Scott, J.; McGuire, W. L. New Molecular Markers of Prognosis in Breast Cancer. In *Endocrine-Dependent Tumors*; Voigt, K. D., Knabe, C., Eds.; Raven Press: New York, 1991; pp 179–196.
- (21) Lloyd, D.; Smith, H.; O'Sullivan, T. P.; Zisterer, D. M.; Meegan, M. *J. Med. Chem.* **2005**, in press.
- (22) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998**, *95* (7), 927–37.
- (23) Miller, M. A. Chemical database techniques in drug discovery. *Nat. Rev. Drug Discovery* **2002**, *1* (3), 220–7.
- (24) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47* (10), 2499–510.
- (25) Bostrom, J. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, *15* (12), 1137–52.
- (26) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470–89.
- (27) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57* (2), 225–42.
- (28) *FRED*, version 2.0.1; Openeye Scientific Software. <http://www.eyesopen.com>.
- (29) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* **2003**, *9* (1), 47–57.
- (30) *MACROMODEL*, version 6.5; Schrodinger Inc. <http://www.schrodinger.com>.
- (31) *Molecular Operating Environment*; Chemical Computing Group. <http://www.chemcomp.com>.
- (32) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein–ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 793–806.
- (33) Derwent World Drug Index. <http://thomsonderwent.com/products/lr/wdi>.
- (34) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46* (1–3), 3–26.
- (35) Daylight Chemical Information Systems Inc. <http://www.daylight.com>.
- (36) *FILTER*; Openeye Scientific Software. <http://www.eyesopen.com>.
- (37) Strausberg, R. L.; Schreiber, S. L. From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science* **2003**, *300* (5617), 294–5.
- (38) *CONVERT*; Molecular Networks GmbH. <http://www.mol-net.de>.
- (39) *TAUTOMER*; Molecular Networks GmbH. <http://www.mol-net.de>.
- (40) *UNITY*; Tripos Inc. <http://www.tripos.com>.
- (41) *TAUTOMER*; Openeye Scientific Software. <http://www.eyesopen.com>.
- (42) *QUACPAC*, version 1.1; Openeye Scientific Software. <http://www.eyesopen.com>.
- (43) *STERGEN*; Molecular Networks GmbH. <http://www.mol-net.de>.
- (44) *FLIPPER*; Openeye Scientific Software. <http://www.eyesopen.com>.
- (45) *CORINA*, version 3.6; Molecular Networks GmbH. <http://www.mol-net.de>.
- (46) *OMEGA*, version 1.8.1; Openeye Scientific Software. <http://www.eyesopen.com>.
- (47) *RUBICON*; Daylight Chemical Information Systems Inc. <http://www.daylight.com>.
- (48) *CATALYST*; Accelrys: San Diego, CA. <http://www.accelrys.com>.
- (49) Sadowski, J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000.
- (50) *SMARTS*; Daylight Chemical Information Systems Inc. <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>.
- (51) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.* **2002**, *20* (4), 281–95.
- (52) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14* (8), 731–51.
- (53) Good, A. C.; Cheney, D. L. Analysis and optimization of structure-based virtual screening protocols (1): exploration of ligand conformational sampling techniques. *J. Mol. Graph. Model.* **2003**, *22* (1), 23–30.
- (54) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graph. Model.* **2003**, *21* (5), 449–62.

CI050185Z