Ensemble Docking into Multiple Crystallographically Derived Protein Structures: An Evaluation Based on the Statistical Analysis of Enrichments

Ian R. Craig,*,† Jonathan W. Essex,‡ and Katrin Spiegel†

Novartis Institutes for Biomedical Research, Wimblehurst Road, Horsham, West Sussex, RH12 5AB, U.K., and School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K.

Received October 21, 2009

Docking into multiple receptor conformations ("ensemble docking") has been proposed, and employed, in the hope that it may account for receptor flexibility in virtual screening and thus provide higher enrichments than docking into single rigid receptor structures. The statistical analyses presented in this paper provide quantitative evidence that in some cases docking into a crystallographically derived conformational ensemble does indeed yield better enrichment than docking into any of the individual members of the ensemble. However, these "successful" ensembles account for only a minority of those examined and it would not have been possible to prospectively predict their identity using only protein structural information. A more frequently observed outcome is that the ensemble enrichment is higher than the mean of the enrichments provided by its individual members. An additional and promising finding is that, if a set of known active compounds is available, an approach based on induced-fit docking appears to be a reliable way to construct ensembles which provide relatively high enrichments.

INTRODUCTION

The incorporation of receptor flexibility in automated docking algorithms may enable more accurate binding pose prediction and better virtual screening enrichments, in addition to providing a more realistic description of the physics of the protein−ligand binding interaction. As a result, this area is attracting considerable current interest, resulting in diverse methodological developments that have been summarized in recent reviews.[1−4] Many of these approaches can be catagorised as "ensemble docking" (ED) strategies (also known as multiple protein structures).[5−8] In its simplest implementation, ED involves docking compounds into multiple conformations of the target receptor rather than just the single rigid receptor structure used in standard docking methods.[9] This ensemble of receptor conformations mimics the conformational equilibrium which characterizes the native state of the target protein[10] and provides a structural degree of freedom by which the conformation of the protein model may be matched to fit any particular ligand. Therefore, unlike standard rigid receptor docking (SD), there is no requirement that every ligand binds to the same conformation of the protein.

Ensemble docking has certain advantages over other flexible-receptor docking methods, such as induced-fit docking (IFD)[11,12] or deformation along low-frequency normal modes.[13,14] One is that, in principle, ED is capable of accounting for any type of protein motion, whatever the length-scale and however complex. In practice, whether the ensemble's coverage of protein conformational space is sufficient or not depends on the ligands to be docked and, crucially, on the conformational diversity of the source of protein structures. The established methods of generating protein structural information, which include crystallographic (X-ray), spectroscopic (NMR), and computational techniques, each have pros and cons in this latter regard.[15]

A disadvantage of the simple version of ED sketched above is that the required computational effort scales linearly with the number of receptor conformations in the ensemble. Ensemble docking can therefore take signficantly longer than standard docking protocols, although recent advanced ED variants reduce runtimes by simultaneously optimizing both ligand coordinates and a discrete variable indexing the different protein conformations in the ensemble.[16,17] However, perhaps worse than needing more computer time than SD is the possibility that ED may provide less accurate results. The conformational flexibility of the ensemble protein model can couple with imperfections in scoring functions to generate inaccurate predictions of some protein−ligand interactions. For example, an incorrect binding pose may be assigned a favorable score if an insigificant interaction with a particular protein conformation is erroneously overrewarded by the scoring function. Such problems might be exacerbated because common scoring functions (such as that used below) do not properly account for the relative energies of different receptor conformations. In the context of virtual screening, these issues mean that receptor flexibility can "optimize" protein−ligand interactions for inactive, as well as for active, compounds. This raises the possibility of increased false positive rates and decreased enrichments.[18,19]

In the past, the intuitive notion that multiple receptor conformations should improve the accuracy of binding predictions has meant that ED has been deployed merely on the *assumption* of superiority over SD. However, to better inform decisions about when ED is worthwhile, this work aims to investigate whether docking into crystallographically

* To whom correspondence should be addressed E-mail: ian.craig@novartis.com.
† Novartis Institutes for Biomedical Research.
‡ University of Southampton.

10.1021/ci900407c
Published on Web 03/11/2010

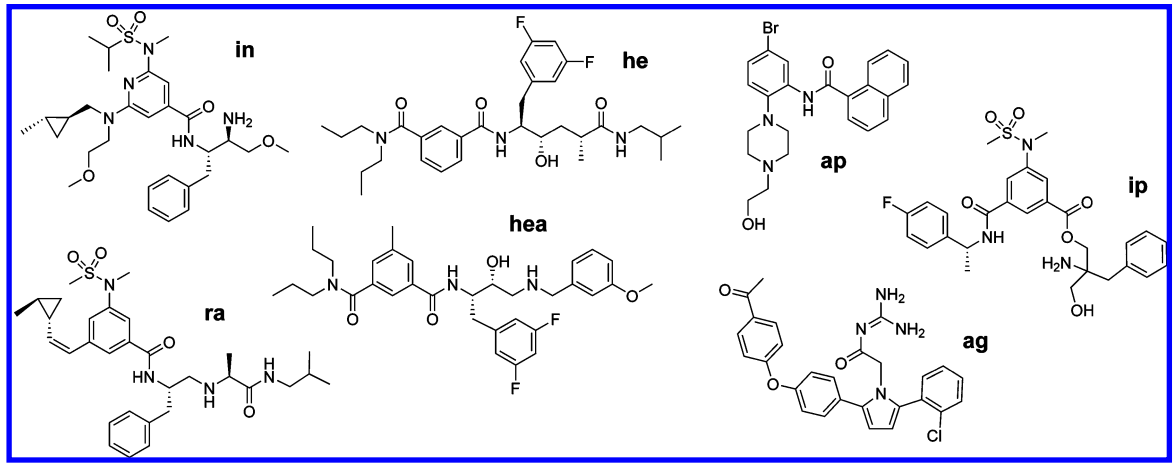

Figure 1. Representative molecules from each of the seven inhibitor chemotypes in the BACE test set (in = isonicotinamide,[31] ra = reduced amide,[32] hea = hydroxyethylamine,[33] he = hydroxyethyl,[34] ap = arylpiperazine,[35] ag = acylguanidine,[36] ip = isophthalamide[37]).

derived conformational ensembles provides better enrichment than docking to single rigid receptors. In particular, the relative ability of ED and SD to rank a diverse set of known inhibitors above property-matched (mimetic[20]) decoy molecules is examined. The comparison employs currently and widely available docking tools and is based on virtual screening against the aspartic protease β-secretase (BACE) and the cAbl kinase domain (cAbl). The binding sites of both proteins exhibit significant and functionally important flexibility, making them highly relevant model systems for the current evaluation.

It is worth noting that there are alternative ways of exploiting multiple protein structures in virtual screening and pose prediction which have also been termed "ensemble docking". These include docking to the average protein structure[21] and methods similar to FlexE, which generate novel composite protein conformations during the docking by combining fragments from more than one of the experimentally observed structures in the ensemble.[22−24] None of these alternative approaches are considered in this work.

Of the few previous attempts to evaluate ED methodologies, that of Barril and Morley[25] provides the closest predecent for this work. Among a number of topics addressed in a wide-ranging yet thorough study they compared the virtual screening performance of ED to that of SD for cyclin dependent kinase 2 (CDK2) and heat shock protein 90 (HSP90). Ensembles constructed by random selection from large numbers of X-ray structures (49 for CDK2 and 149 for HSP90) were found on average to offer no notable improvement over randomly selected single rigid receptors and in some cases were worse. The optimal ensemble did tend to improve upon the optimal single receptor, but only modestly, and the optimal single receptor *always* outperformed the average performance of random ensembles.

Other evaluations of ensemble docking were performed by Cavasotto and Abagyan,[26] Ferrari et al,[27] and Rao et al.[28] However, this work differs from these predecents in two significant respects: (1) Statistical tests are used to establish the significance of differences in virtual screening performance between ED and SD. This is facilitated by measuring enrichment using the area under curve (AUC) values of receiver operating characteristic (ROC) plots[29] because analytic estimates of the AUC variance are available. (2) Multiple rational ensemble construction strategies are systematically evaluated. Some are structure-based approaches to compiling highly diverse (or highly similar) sets of protein structures, such as principal component analysis. Others are ligand-based, and rely on the availability of a set of known actives. Previous evaluations constructed ensembles by random selection[25] or exhaustive enumeration of the possible two- or three-membered ensembles.[26,28]

The outline of the paper is as follows. The subsequent section describes various aspects of the evaluation methodology. The virtual screening performance of ensemble docking is then analyzed and compared to that of single receptor docking before conclusions are drawn and directions for future work are identified.

EVALUATION METHODOLOGY

A more extensive version of this section is included in the Supporting Information.

Ligand Test-Set Construction. A structurally diverse set of known BACE inhibitors was compiled using compounds extracted from the BindingDB,[30] a publicly available database of experimentally measured protein−ligand binding affinities. As of September 2008, the BindingDB contained 67 compounds with BACE-1 IC_{50} less than 1 μM and molecular weight less than 650 Da. This set was reduced by first identifying seven diverse chemotypes present within it (see Figure 1). Manual selection of eight or nine of the more structurally diverse representatives of each chemotype then produced a total of 59 BACE-active compounds.

A similar approach was used to assemble a diverse set of cAbl inhibitors from an in-house compound collection. About 3500 compounds with a molecular weight between 200 and 600 Da were found to have an IC_{50} of less than 1 μM in an enzymatic cAbl inhibition assay. A selection of 10 representatives from each of 14 diverse chemotypes produced a final set of 140 cAbl-active compounds.

For both BACE and cAbl, the test sets were completed by augmenting the active compounds with decoy molecules extracted from the publicly accessible ZINC database of commercially available compounds.[38] A key design principle was that decoy compounds were selected such that the overall decoy set possessed similar molecular weight and logP distributions to the active set,[20] as well as being structurally diverse. Compounds that were obvious multimers of smaller

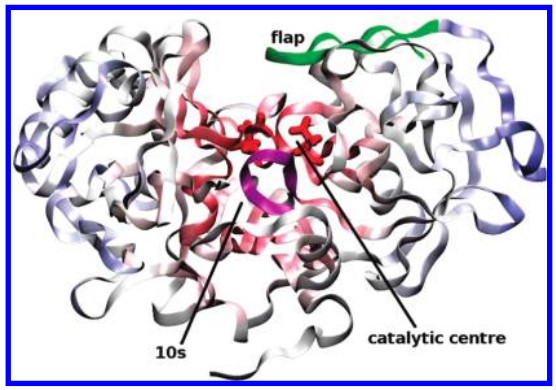

Figure 2. Diagram highlighting some important structural components of β-secretase. The flexible flap (green) is in the closed position, covering the two catalytic aspartate residues at the center of the figure (red). The 10s loop (purple) adopts a closed conformation. This diagram is derived from the crystal structure with PDB code 1FKN; the ligand has been removed to better reveal the active site.

molecules, or were highly peptidic in character, or had long aliphatic side-chains, or had more than 4 fused rings were rejected. The overall result was a set of 408 structurally diverse property-matched decoys for the BACE test set and similarly 397 decoys for cAbl. Previous work has demonstrated large changes in enrichment when the same set of actives is screened alongside decoy sets with different characteristics, for example, decoys derived using different approaches or from different sources.[39] To keep the cAbl test set as consistent as possible with the BACE test set, it was therefore decided not to derive the cAbl decoys from compounds confirmed as inactive in the same assay from which the actives were drawn.

Crystal-Structure Selection. Initial sets of crystal structures for BACE and cAbl were downloaded from the RCSB Protein Data Bank.[40] For both targets, principal component analysis (PCA)[41] was used to map the major conformational differences between these structures and to thereby identify small but conformationally diverse sets of receptor conformations. For BACE and cAbl in turn, the following first summarizes what is generally known about active-site flexibility and then sketches the results of the PCA for these particular sets of structures.

β-Secretase. The protein backbone is known to be mobile in at least two different parts of the BACE active site.[42] Most important is the motion of the "flap", a β-sheet loop which partially covers the catalytic center in its "closed" conformation but which can lift by ~5 Å to expose the binding site in the "open" conformation. The "10s" loop, which borders the S3 subpocket at one end of the binding site, is also seen to adopt at least two different conformations: when "closed" (or "down"[43]) the loop is positioned ~2−3 Å closer to the catalytic center than when it is "open" (or "up"[43]). These structural elements are illustrated in Figure 2.

A PCA was performed on the set of BACE crystal structures using the Cartesian C_α coordinates of the following active-site residues (all BACE residue numberings are those of the PDB structure 1FKN): Lys9 to Gln12 inclusive (10s) and Val69 to Gly74 inclusive (flap). A plot of the projections along the first two principal components is shown in Figure 3. This shows significant structural variation both of the flap (first principal component) and, independently, of the 10s

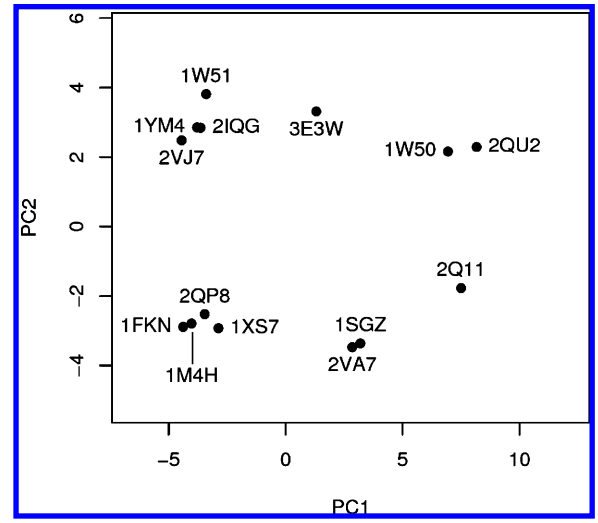

Figure 3. Principal component analysis of BACE active-site C_α coordinates: Lys9 to Gln12 inclusive (10s) and Val69 to Gly74 inclusive (flap). Residue numberings are those of the PDB structure 1FKN. More positive values of the first principal component (PC1) correspond to flap-open structures, and more positive values of the second principal component (PC2) to 10s-open structures.

loop (second principal component). For the same set of active-site residues, the mean heavy atom rmsd from the other flap-closed, 10s-closed structures (1M4H, 1XS7, 2QP8) to 1FKN is 0.91 Å. The equivalent quantity from the flap-open, 10s-closed structures (2Q11, 1SGZ, 2VA7) is 3.35 Å; from the flap-closed, 10s-open structures (1YM4, 1W51, 2IQG, 2VJ7), it is 2.40 Å, and from the flap-open, 10s-open structures (3E3W, 2QU2, 1W50), it is 4.00 Å. The nominal resolution of the BACE crystal structures ranges from 1.5 to 2.8 Å.

cAbl. The cAbl kinase domain exhibits significant structural variation in and around the ATP binding site. The "activation loop" undergoes a conformational reorganization of considerable amplitude, thereby performing the crucial functional role of switching the enzyme between its active and inactive states. Other discernible motions, albeit at smaller scales, are that of the DFG segment at the N-terminal end of the activation loop and of the glycine-rich loop which forms one side of the binding site. In particular, the latter adopts at least 4 different conformations ranging from an extended "tongue-like" form (active), through to more compact configurations, such as a "U" shape (inactive) and a "W" shape (of which there are both active and inactive subtypes). Figure 4 highlights the relevant parts of the protein.

The downloaded cAbl crystal structures included kinase-active (6), inactive (5), and intermediate (2) conformations. A PCA was performed on this set using the Cartesian C_α coordinates of the following restricted set of active-site residues (all cAbl residue numberings are those of the PDB structure 2HZI): Lys247 to Tyr257 inclusive (glycine-rich loop), Ile314 to Gly321 (hinge), and Lys378 to Pro402 (DFG/activation loop). Unsurprisingly, the first principal component corresponded to the switch in position of the activation loop between the active and inactive conformations. The only other significant structural variation was that of the conformation of the glycine-rich loop.

A second PCA was therefore performed using only the C_α atoms of the glycine rich loop, and the resulting

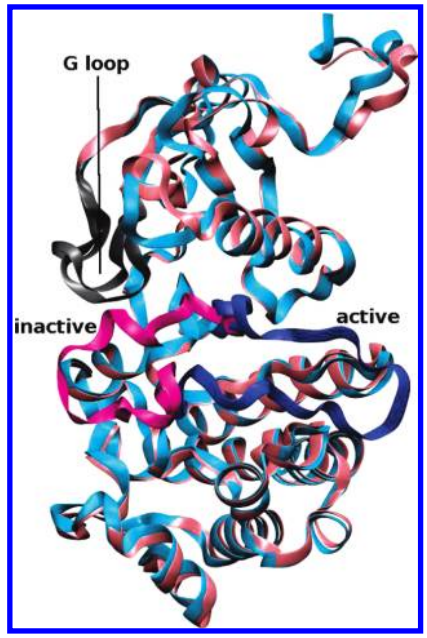

Figure 4. Alignment of an active (cyan/blue) and inactive (pink/magenta) structure of the cAbl kinase domain to illustrate the flexibility discussed in the text. From this perspective the ATP binding site is at the center of the figure (and toward the rear). The activation loop is shown in both its active (blue) and inactive (magenta) conformations. For both structures the glycine-rich loop (G-loop) is shown in gray. This diagram is derived from the crystal structures with PDB codes 1IEP (inactive) and 2GQG (active); the ligands have been removed to better reveal the active site.

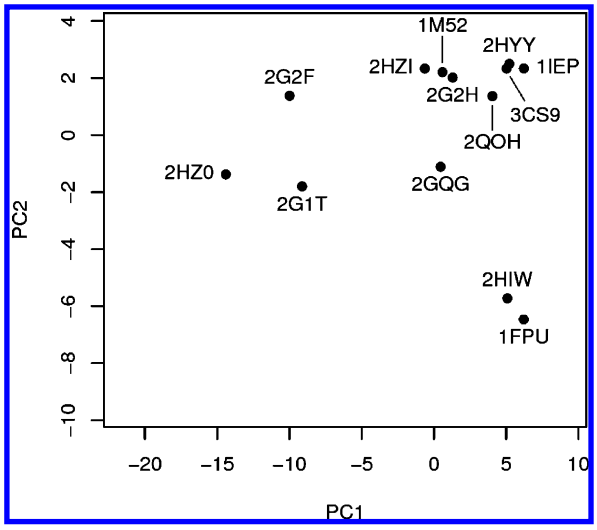

Figure 5. Principal component analysis of cAbl glycine-rich loop C_α coordinates: Lys247 to Tyr257 inclusive (glycine-rich loop). Residue numberings are those of the PDB structure 2HZI. More negative values of the first principal component (PC1) correspond to more extended (tongue-like) conformations of the loop. More positive values of the second principal component (PC2) correspond to W-shape rather than U-shape conformations of the loop.

projections along the first two principal components are plotted in Figure 5. In broad terms, more negative values of the first principal component correspond to more extended (tongue-like) conformations of the glycine-rich loop. More positive values of the second principal component correspond to "W" rather than "U" conformations of the loop. For the same set of residues, the mean heavy atom rmsd from the other active W-shape conformations (1M52, 2G2H, 2QOH) to 2HZI is 2.72 Å. The equivalent quantity from the inactive W-shape conformations (1IEP, 2HYY, 3CS9) is 3.29 Å; from the U-shape conformations (2HIW, 1FPU), it is 4.03 Å, and from the tongue-like structures (2HZ0, 2G2F, 2G1T, 2GQG), it is 5.04 Å. The nominal resolution of the cAbl crystal structures ranges from 1.7 to 2.7 Å.

Docking Protocol. *β-Secretase.* All test-set ligands were prepared for docking using the same preparation procedure. First, all hydrogen atoms were removed and then added back in Maestro.[44] Each compound was then submitted to LigPrep[45] to generate plausible ionization (pH 7) and tautomerization states. Up to eight stereoisomers were also produced by varying the configuration at undefined chiral centers. In this way, default LigPrep settings produced 1847 prepared ligand "states" (i.e., stereoisomers, tautomers, and ionization states of the original 467 test-set compounds).

After removal of any cocrystal ligands, Maestro's Protein Preparation Wizard was used to prepare the chosen set of crystal structures for conversion to docking receptor grids. Default settings were used, except that disulfide bonds were detected, and all crystallographic water molecules were removed. One of the catalytic aspartic acid residues was then protonated (Asp32[46,47]) before the structures were energy minimized. Glide[48] receptor grids centered at the mean position of residues Asp32 (protonated), Thr72, and Asp228 were generated from the resulting protein structures. In all other respects, the default settings were used, and no constraints were defined.

Docking of the test-set into each receptor grid was performed using Glide in SP mode. Default settings were used, except that nonplanar amide bonds were allowed (this resulted in improved enrichment performance in preliminary tests). For each receptor, only the best scoring pose for each test set compound was retained. Sorting these poses in order of increasing GlideScore (decreasingly favorable interactions) then created a ranked list. At this point, ensemble docking simply corresponds to merging the ranked lists of the constituent receptors. Various rules can be used to perform this operation.[8,26] In this work, of the multiple poses for each test-set compound (one in each of the constituents' ranked list), that with the most favorable docking score is selected.[25] In the case of Glide, this is the pose assigned the most negative GlideScore. The result is a ranked list for the ensemble.

cAbl. The docking protocol used for cAbl was very similar to that described for BACE, and only the differences are noted here. Ligand preparation produced 912 ligand states from the original 537 test-set compounds. Glide receptor grids were centered at the mean position of residues Gly251, Asp381, and Thr315 and, to comfortably cover the entire cAbl active site, had sides of length 22 Å rather than the default of 20 Å, which was used for BACE.

Enrichment Metrics. The virtual screening performance (i.e., enrichment) of each receptor, and each ensemble was quantified by the area under curve (AUC) of its receiver operating characteristic (ROC) plot.[29] The ROC plot is constructed by stepping sequentially through the ranked list of test-set compounds arranged in order of increasing GlideScore. At each position, the true positive rate (TPR) and false positive rate (FPR) are obtained by predicting that all compounds at higher ranks (lower GlideScore) are active. The ROC plot is then simply the plot of TPR versus FPR for all positions in the ranked list (see Figure 6).

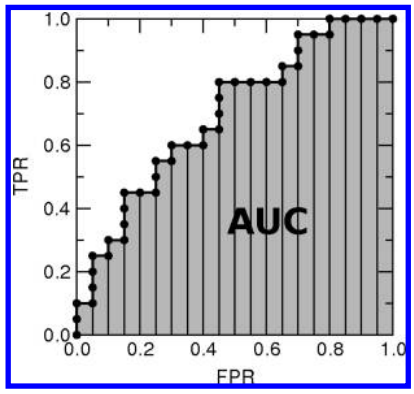

Figure 6. Sketch of a receiver operating characteristic (ROC) plot (TPR = true positive rate and FPR = false positive rate, see section on evaluation metrics). The area under curve (AUC) is illustrated.

Although there has been much debate about the relative merits of different enrichment metrics (see ref 49 and its citations), the AUC has a number of desirable properties[50] that made it an attractive choice for this study. Principal among these is the availability of analytical expressions for AUC variance.[20] This made estimation of error bounds very simple and also meant that statistical significance tests could be applied to differences between the AUC values of different receptors/ensembles (see below).

Pipeline Pilot[51] was used to determine AUC values from SD files containing the ranked lists output from the Glide dockings. Considering the sketch of a typical ROC plot in Figure 6, it is clear that trapezium rule integration along the FPR abscissa gives the AUC as

$$\mathrm{AUC} = \sum_{\mathrm{i}}^{\mathrm{decoys}} \Delta\mathrm{FPR}\,\mathrm{TPR}_i \tag{1}$$

where TPR_i is the true positive rate at decoy i (the number of actives ranked higher than decoy i divided by the total number of actives) and ΔFPR is the constant increment in the false positive rate. Since $\Delta\mathrm{FPR} = 1/N_{\mathrm{decoys}}$,

$$\mathrm{AUC} = \frac{1}{N_{\mathrm{decoys}}} \sum_{i}^{\mathrm{decoys}} \mathrm{TPR}_i = \langle \mathrm{TPR} \rangle_{\mathrm{decoys}} \tag{2}$$

and so the AUC is just the mean TPR of the decoys. This is the justification for interpreting the AUC as the probability of correctly ranking one randomly selected active above one randomly selected decoy, and provides the connection to the Wilcoxon statistic.[52] One can alternatively integrate along the TPR ordinate of the ROC plot to give

$$\mathrm{AUC} = 1 - \langle \mathrm{FPR} \rangle_{\mathrm{actives}} \tag{3}$$

The preceding equations cast the AUC in terms of the mean of the decoys' TPR distribution or, equivalently, the mean of the actives' FPR distribution. The variances of these two distributions,

$$\mathrm{Var_d} = \frac{1}{N_{\mathrm{decoys}}} \sum_{i}^{\mathrm{decoys}} (\mathrm{TPR}_i - \langle \mathrm{TPR} \rangle_{\mathrm{decoys}})^2 \tag{4}$$

$$\mathrm{Var_a} = \frac{1}{N_{\mathrm{actives}}} \sum_{i}^{\mathrm{actives}} (\mathrm{FPR}_i - \langle \mathrm{FPR} \rangle_{\mathrm{actives}})^2 \tag{5}$$

tend to be large when the decoys and actives are widely distributed throughout the ranked list. They are combined to provide the central-limit theory (CLT) estimate of the AUC standard error

$$\mathrm{SE_{AUC}} = \sqrt{\frac{\mathrm{Var_a}}{N_{\mathrm{actives}}} + \frac{\mathrm{Var_d}}{N_{\mathrm{decoys}}}} \tag{6}$$

A discussion of issues surrounding AUC error analysis was given recently by Nicholls.[20]

The standard error, $\mathrm{SE_{AUC}}$, in eq 6 can be used to obtain confidence limits (i.e., error bounds) for the AUC value of a single receptor or ensemble. However, answers to the questions posed in this evaluation required assessment of whether the difference between the AUC metrics of two receptors (or ensembles) was statistically significant. Comparing the confidence limits of individual receptors' AUC values would have been insufficient, not least because this neglects correlation between the docking scores of the same compound at the different receptors.[20]

What is needed instead is an explicit error analysis of the difference between two AUC values. Using eq 2, one can write

$$\Delta\mathrm{AUC} = \mathrm{AUC_A} - \mathrm{AUC_B} = \langle \mathrm{TPR} \rangle_{\mathrm{decoys,A}} - \langle \mathrm{TPR} \rangle_{\mathrm{decoys,B}} \tag{7}$$

However, in this study the same compounds were docked to each receptor/ensemble - in statistical terminology the samples (in eq 7, the decoys) are paired. As a result,

$$\Delta\mathrm{AUC} = \frac{1}{N_{\mathrm{decoys}}} \sum_{i}^{\mathrm{decoys}} (\mathrm{TPR}_{i,\mathrm{A}} - \mathrm{TPR}_{i,\mathrm{B}}) = \langle \mathrm{TPR_A} - \mathrm{TPR_B} \rangle_{\mathrm{decoys}} \tag{8}$$

or, alternatively, starting from eq 3,

$$\Delta\mathrm{AUC} = \frac{1}{N_{\mathrm{actives}}} \sum_{i}^{\mathrm{actives}} (\mathrm{FPR}_{i,\mathrm{B}} - \mathrm{FPR}_{i,\mathrm{A}}) = \langle \mathrm{FPR_B} - \mathrm{FPR_A} \rangle_{\mathrm{actives}} \tag{9}$$

Following the path laid out above, the variances associated with the means in these last two equations are

$$\mathrm{Var}_{\Delta,\mathrm{d}} = \frac{1}{N_{\mathrm{decoys}}} \sum_{i}^{\mathrm{decoys}} ((\mathrm{TPR}_{i,\mathrm{A}} - \mathrm{TPR}_{i,\mathrm{B}}) - \langle \mathrm{TPR_A} - \mathrm{TPR_B} \rangle_{\mathrm{decoys}})^2 \tag{10}$$

$$\mathrm{Var}_{\Delta,\mathrm{a}} = \frac{1}{N_{\mathrm{actives}}} \sum_{i}^{\mathrm{actives}} ((\mathrm{FPR}_{i,\mathrm{B}} - \mathrm{FPR}_{i,A}) - \langle \mathrm{FPR}_B - \mathrm{FPR}_A \rangle_{\mathrm{actives}})^2 \tag{11}$$

giving the standard error in ΔAUC, the difference in AUC values, as

$$\mathrm{SE}_\Delta = \sqrt{\frac{\mathrm{Var}_{\Delta,\mathrm{a}}}{N_{\mathrm{actives}}} + \frac{\mathrm{Var}_{\Delta,\mathrm{d}}}{N_{\mathrm{decoys}}}} \tag{12}$$

Table 1. AUC Metrics for the Single Rigid BACE Receptor Structures

PDB	AUC
2Q11	0.778
1XS7	0.751
1SGZ	0.743
1M4H	0.743
2VA7	0.732
2IQG	0.732
2VJ7	0.728
2QU2	0.726
3E3W	0.719
1YM4	0.718
1FKN	0.714
2QP8	0.701
1W50	0.688
1W51	0.679

To quantify the significance of any observed difference in the AUC values of receptor/ensemble A and receptor/ensemble B, the null hypothesis that the AUC values were the same was tested against the alternative hypothesis that AUC_A was different to AUC_B. Given the numbers of actives and decoys in the test set, this was done by calculating a two-sided p-value for the observed ΔAUC using the CLT Gaussian distribution with standard deviation equal to SE_Δ:[53]

$$p = \mathrm{erfc}\left(\frac{|\Delta\mathrm{AUC}|}{\sqrt{2}\mathrm{SE}_\Delta}\right) \tag{13}$$

where erfc(x) is the complementary error function. Note that eqs 7−9 may each give a slightly different value for ΔAUC if some small number of test-set compounds fails to dock to either of the compared receptors A and B. This occurs because the summations in eqs 8 and 9 can only cover the subsets of actives and decoys that successfully docked to both receptors. The p-values presented in the following section have been determined using ΔAUC from eq 9 because it is most consistent with standard error estimate in eq 12. However, there is no qualitative change in the results if eq 7 is used instead (see Supporting Information (Tables S7−S9)).

RESULTS

Single Rigid Receptor Docking. *β-Secretase.* Table 1 presents the AUC metrics of the 14 crystallographically derived single rigid receptors. Despite the structural differences, the AUC values of the receptors span a narrow range corresponding to a moderate ability to discriminate between actives and decoys. Aside from a couple of receptors at either end of the range, they are quite tightly grouped around an AUC of about 0.73. The statistical significance of these apparently small differences in the AUC values was assessed using the methodology described above. In particular, for each pair of receptors, a two-sided p-value (eq 13) was calculated to test the null hypothesis that their AUC values are the same against the alternative hypothesis that they differ. The results of this analysis are displayed in Table 2, where p-values less than 0.05 are boldfaced to emphasize significance at the 95% level.

Given the previous observation that there is limited variation among the AUC values, it might be expected that the differences between them are mostly insignificant. This is confirmed by Table 2. The difference in AUC values is statistically significant for only 11 pairs of receptors, whereas at the 95% level, five significant differences (0.05 × 91 pairs) would be expected by chance (type I error). Even the highest AUC value (for receptor 2Q11) is only significantly better than three of the others. Similarly, the lowest value (1W50) is significantly worse than only four of the others. Overall, only four of the receptors have signficantly higher AUC values than one or more of the others. In the {flap position}//{10s position} notation used throughout the rest of this paper, one of them is open//closed, two are closed//closed, and one is closed//open. It is not therefore apparent that one particular backbone conformation of the BACE active-site consistently provides the best virtual screening performance. Likewise, there appears to be no obvious connection between the side-chain conformations of active-site residues (such as Tyr71, Asp32, and Asp228) and the occurrence of statistically significant differences in the AUC value.

However, by analyzing the average ranks of the seven active chemotypes in the different receptors, some rather more subtle connections did emerge. For this purpose, the fractional rank of a compound was defined as its position in the ranked list (see section on evaluation metrics above) divided by the length of the ranked list. It was then observed that the median fractional rank of the 9 representatives of the acylguanidine chemotype was more than 0.5 in all receptors for which the hydroxy group of the phenolic Tyr71 side-chain is "tucked in" to the base of the flap. This is consistent with the available acylguanidine-BACE crystal structures, in which the Tyr71 side-chain is orientated at right angles to the main axis of the flap. A similar observation is that arylpiperazines only achieved a median fractional rank of less than 0.5 with open//closed structures. In contrast, the hydroxyethylamine, reduced-amide, and isophthalamide chemotypes achieved median fractional ranks of less the 0.15 against *all* receptors. Fractional ranks are summarized by chemotype for three receptors in Figure 7. Receptor 2Q11 is included as it has the highest observed AUC value, while the 1SGZ and 2QU2 receptors are notable as they assign particularly high ranks to certain chemotypes that are poorly ranked by other receptors (arylpiperazine and acylguanidine, respectively). Below, this chemotype analysis is shown to be the basis of one way to construct successful ensembles.

cAbl. The AUC metrics of the 13 cAbl crystallographically derived single rigid receptors are reported in Table 3. All except one of these cAbl AUC results are higher than any of those found for the BACE structures (Table 1), and even after discarding outliers they also cover a slightly wider range of values. The table also shows that structures with an inactive kinase conformation tend to produce higher AUC values for this test-set. The exceptions to this are active structures 2QOH and 2GQG, which have AUC values at the top and near the top of the range, respectively.

As above, the statistical significance of the difference in AUC values for each pair of receptors in Table 3 can be assessed using the p-values reported in Table 4. In comparison to the equivalent BACE results in Table 2, statistically significant differences in AUC metrics occur more frequently in this cAbl test case (see below). The difference in AUC values is statistically significant for 37 pairs of receptors, whereas at the 95% level, only four significant differences (0.05 × 78 pairs) would be expected by chance. The best

Table 2. *p*-Values for Differences in the AUC Values of Single Rigid BACE Receptor Structures[a]

	2Q11	1XS7	1SGZ	1M4H	2VA7	2IQG	2VJ7	2QU2	3E3W	1YM4	1FKN	2QP8	1W50	1W51
2Q11		0.469	0.375	0.346	0.230	0.165	0.097	0.168	0.162	0.085	0.082	**0.020**	**0.016**	**0.020**
1XS7			0.834	0.693	0.497	0.295	0.349	0.588	0.259	0.195	0.161	**0.033**	**0.020**	**0.011**
1SGZ				1.000	0.784	0.774	0.721	0.578	0.525	0.571	0.522	0.321	0.183	0.175
1M4H					0.670	0.544	0.602	0.705	0.317	0.215	0.128	0.055	**0.006**	**0.001**
2VA7						0.996	0.899	0.889	0.595	0.623	0.513	0.280	0.058	0.076
2IQG							0.849	0.880	0.630	0.544	0.354	**0.031**	**0.031**	**0.031**
2VJ7								0.969	0.803	0.722	0.607	0.279	0.221	0.091
2QU2									0.436	0.435	0.400	0.280	0.193	0.352
3E3W										0.975	0.880	0.584	0.172	0.169
1YM4											0.879	0.488	0.216	0.096
1FKN												0.561	0.284	0.168
2QP8													0.591	0.419
1W50														0.662
1W51														

[a] A two-sided *p*-value is shown for each pair of structures,and those below 0.05 are in bold to emphasize signficance at the 95% level.

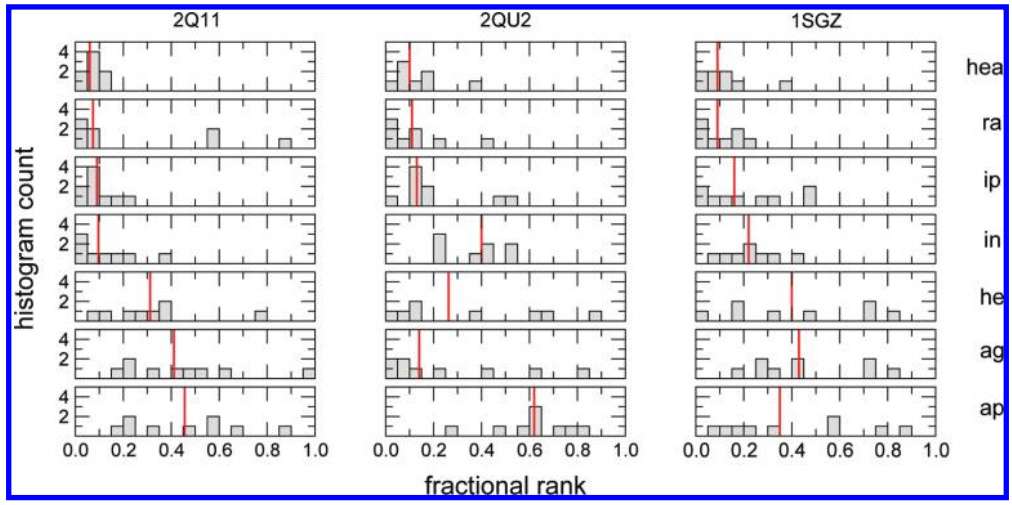

Figure 7. Breakdown of docking performance by chemotype for three rigid BACE receptor structures of interest (2Q11, 2QU2, and 1SGZ). Each cell shows a histogram of fractional ranks achieved by the actives of a particular chemotype (labeled on the right) in a particular receptor (labeled above), where a compound's fractional rank is its position in the ranked list (see section on evaluation metrics in text) divided by the length of the list. In each cell the median fractional rank is indicated by a vertical red line. Chemotype labels are as follows: hea = hydroxyethylamine, ra = reduced amide, ip = isophthalamide, in = isonicotinamide, he = hydroxyethyl, ag = acylguanidine, and ap = arylpiperazine.

Table 3. AUC Metrics and Activation Loop Conformation for the Single Rigid cAbl Receptor Structures

PDB	AUC	conformation
2QOH	0.910	active
3CS9	0.897	inactive
2HYY	0.888	inactive
2GQG	0.882	active
2HIW	0.869	inactive
1IEP	0.867	inactive
1FPU	0.866	inactive
1M52	0.858	active
2HZI	0.856	active
2HZ0	0.835	intermediate
2G2H	0.828	active
2G2F	0.822	active
2G1T	0.671	intermediate

performing single receptor (2QOH) has a significantly higher AUC than all except three of the other receptors. More generally, the top four receptors in Table 3 each have a significantly higher AUC value than at least four of the other receptors. Also of note is the fact that the AUC of 2G1T is significantly lower than all of the other receptors. This may be related to the somewhat unusual position of the kinase activation loop in this structure, which approaches neither of the familiar active and inactive conformations. The only other receptor with an activation loop conformation that falls outside the active/inactive classification is 2HZ0. It is therefore interesting that this receptor also has an AUC that is significantly worse than most of the other structures.

Before moving on to examine the ensemble docking results, it is worth considering how the nature, in particular the size, of the ligand test-set affects the observed frequency of statistical significance among differences in AUC values. According to eq 13, the *p*-value and, therefore, statistical significance is determined by two variables: ΔAUC, the magnitude of the difference between AUC scores, and SE_Δ, the standard error of that difference (eq 12). Since the ΔAUC values are on average lower for the BACE data set than for the cAbl data set (compare Table 1 and Table 3), statistical significance within the former requires smaller standard errors. Looking at eq 12, $Var_{\Delta,a} \simeq Var_{\Delta,d}$, while $N_{actives} > N_{decoys}$, so the principal constraint on reducing the β-secretase SE_Δ values is the limited number of BACE inhibitor chemotypes available in the BindingDB.[30] For the cAbl data set the differences among AUC values are larger and,

Table 4. *p*-Values for Differences in the AUC Values of Single Rigid cAbl Receptor Structures[a]

	2Q0H	3CS9	2HYY	2GQG	2HIW	1IEP	1FPU	1M52	2HZI	2HZ0	2G2H	2G2F	2G1T
2QOH		0.420	0.263	0.142	**0.033**	**0.022**	**0.021**	**0.005**	**0.002**	**0.000**	**0.000**	**0.001**	**0.000**
3CS9			0.789	0.488	0.158	0.066	**0.047**	**0.040**	**0.007**	**0.000**	**0.000**	**0.005**	**0.000**
2HYY				0.760	0.211	0.063	0.073	**0.047**	**0.024**	**0.001**	**0.005**	**0.005**	**0.000**
2GQG					0.517	0.378	0.374	0.176	0.122	**0.020**	**0.014**	**0.008**	**0.000**
2HIW						0.872	0.840	0.406	0.390	**0.010**	0.091	0.060	**0.000**
1IEP							0.902	0.502	0.447	**0.034**	0.082	0.101	**0.000**
1FPU								0.412	0.391	**0.040**	0.062	0.069	**0.000**
1M52									0.771	0.130	0.188	0.301	**0.000**
2HZI										0.240	0.162	0.295	**0.000**
2HZ0											0.961	0.903	**0.000**
2G2H												0.946	**0.000**
2G2F													**0.000**
2G1T													

[a] A two-sided *p*-value is shown for each pair of structures, and those below 0.05 are in bold to emphasize signficance at the 95% level.

Table 5. Strategies Used to Construct Ensembles from a Pool of Crystallographically-Derived Receptors

construction strategy	description
A	combine receptors with similar conformations
B	combine receptors with dissimilar conformations
C	add receptors to that with the lowest AUC
D	add receptors to that with the highest AUC
E	combine receptors which perform well for different chemotypes

furthermore, the standard errors in ΔAUC (eq 12) tend to be smaller because of the larger number of actives in the test-set (see raw data in Supporting Information (Tables S1 and S2)). In consequence, as observed above, more statistically significant differences are observed within the cAbl data set (Table 4) than within the BACE data set (Table 2).

Ensemble Docking. Table 5 summarizes a number of strategies that were used to assemble receptor ensembles from the crystallographically derived single-receptor structures examined in the previous subsection. The aim was to establish whether certain strategies are more likely to produce useful ensembles than others. Strategy A, combining receptors with similar protein backbone conformations, was motivated by observations made in a previous attempt at evaluating ensemble docking.[28] There, a retrospective examination found that the ensembles judged to be more successful tended to combine receptors with only subtly different conformations of key side-chains. The opposite strategy, combining receptors with dissimilar protein backbone structures (strategy B), is intuitively more compatible with the hypothesis that motivates ensemble docking. Namely, that ED should permit different chemotypes to bind to different protein conformations. Both strategies were implemented using the principal component analysis of active-site C_α coordinates discussed above. In particular, the shorter the distance between two crystal structures in the space of the first two principal components the more similar their backbone conformations were judged to be. In assembling sets of (dis)similar protein conformations, distances in principal component space (i.e., similarities) were judged "by eye" from plots such as Figure 3 and Figure 5.

Strategies C and D provide an interesting perspective on how ED compares to SD in "best-case" and "worst-case" scenarios, i.e. respectively improving on the single receptor with lowest AUC (strategy C) and on that with the highest AUC (strategy D). However, both would require a set of known active compounds and a considerable amount of time to apply prospectively, given that docking into a large set of structurally diverse single receptors is a prerequisite. Implementation of strategy E used the fractional rank analysis discussed above (Figure 7) to identify and combine single receptors that tended to assign high ranks to different (i.e., complementary) chemotypes. One specific example discussed below combines 2QU2 (assigns high ranks to the acylguanidine chemotype relative to other receptors), 1SGZ (assigns high ranks to the arylpiperazines), and 1XS7 (median fractional rank below 0.15 for all other chemotypes). Again, strategy E requires a set of known active compounds and, even then, its success presumably depends on how well the set of actives covers the relevant chemical space of a given virtual screen.

Of principal interest here is whether the crystallographically derived receptor ensembles achieve better enrichment than docking into the single, rigid receptor structures. Two rather different criteria were used to address this issue. For the first, the AUC value of each ensemble was compared to those of its constituent receptors by calculating two-sided *p*-values (see above). As for the analysis of the single receptor structures in the previous subsection, if a *p*-value for any of these comparisons was less than 0.05 then the difference in AUC between the ensemble and that single receptor was assumed to be significant. Using this "better than all" criterion, an ensemble is judged successful if its AUC value is significantly higher than those of all its constituent receptors. Owing to the multiplication of probabilities, under this criterion ensembles are quite unlikely to be successful by chance (type I error) at the 95% level.

For the second criterion, a *p*-value is calculated for the difference between the AUC of the ensemble and the mean AUC of its constituent receptors (see Supporting Information). If this *p*-value is less than 0.05, then the ensemble is judged successful under the "better than average" criterion. This comparison with the constituents' mean AUC may be relevant in the typical operational setting where one does not know a priori which single receptors yield high (or low) enrichment. Any ensemble docking protocol that provides an enrichment that is consistently better than this mean value would therefore be rather valuable. However, the "better than average" criterion does not discriminate between an ensemble that yields high enrichment because it effectively utilizes its

Table 6. Numbers of Successful Ensembles for Each Construction Strategy[a]

protein	construction strategy	total	better than all	better than average
BACE	A	13	1	5
	B	6	1	5
	C	13	0	7
	D	13	0	9
	E	7	3	6
cAbl	A	8	3	7
	B	13	6	13
	C	11	8	11
	D	12	5	12
	E	7	4	7

[a] An ensemble is judged "better than all" if its AUC value is significantly higher than that of each of its constituent receptors. Less stringently, an ensemble is counted as "better than average" if its AUC value is significantly higher than its constituents' mean AUC value.

multiple constituent receptors (to identify good poses for compounds of different chemotypes with different binding modes) and an ensemble that yields high enrichment simply because just one of its constituents yields high enrichment. In effect, the latter scenario is a masked rigid receptor docking, and a more efficient approach may be to try to identify the single high-enrichment receptor and to subsequently proceed with standard (single receptor) docking techniques. In contrast, the "better than all" criterion distinguishes between these two scenarios and may be more selective for ensembles that are able to retrieve multiple chemotypes.

Table 6 summarizes the numbers of ensembles that were successful under the two different criteria for each ensemble construction strategy (more detailed results are found in the Supporting Information (Tables S3−S6)). The rest of this section now discusses the performance of each ensemble construction strategy in turn, for both β-secretase and cAbl.

β-Secretase. To construct ensembles according to strategy A (combine similar structures), the PCA plot in Figure 3 was used to compose pairs of open//closed, closed//closed, closed//open, and open//open receptors. Of the 13 ensembles constructed in this way, Table 6 reports that only one (2QU2+3E3W) was successful according to the better-than-all criterion. This pair of open//open receptors individually have AUC values in the middle of the range observed in Table 1. A total of five ensembles produced an AUC that was significantly higher than the mean AUC of their constituents. Ensembles were also constructed using the opposite strategy B (combine dissimilar structures), including one example of each of the six possible combinations of closed//closed, open//closed, closed//open, and open//open conformations. Only 1XS7+2QU2, the closed//closed plus open//open ensemble, gave a significantly higher AUC than all constituents. However, all but one of the ensembles constructed using strategy B met the better-than-average criterion.

With regard to ensembles constructed using strategy C (improve worst single receptor), the results in Table 6 suggest that augmenting a low-AUC single receptor with another receptor does not produce ensembles that outperform both of these constituents. However, given the low AUC value of one of the constituents, it is perhaps not surprising that about half of these ensembles satisfy the better-than-average criterion. Table 6 also summarizes results for ensembles constructed using strategy D (improve best single receptor). This indicates that augmenting a high-AUC single receptor with another receptor does not significantly improve upon the AUC of both constituents. In comparison with strategy C, however, an even higher proportion of the strategy D ensembles yield an AUC that is significantly higher than the mean AUC of their two constituents. Thus, if one has reason to assume that a particular receptor will give a relatively good enrichment, then adding it to another receptor might be an effective way to improve the AUC of the latter. However, as discussed above, it may be most efficient to simply switch to using the single high-enrichment receptor in a standard docking protocol.

The final ensemble construction strategy considered in this subsection (strategy E) used the fractional rank analysis presented above (Figure 7) to combine receptors that complement each other by performing particularly well for different chemotypes. Although some obvious disadvantages to this approach were discussed above, it is clearly the most successful of the five ensemble construction strategies in Table 5. The combination of 2QU2 and 1SGZ with a third receptor is responsible for all three of the better-than-all ensembles constructed using this strategy (see Supporting Information (Table S4)). This combination even improved the AUC value of 1XS7, which has the second highest AUC of all the single crystallographically derived structures in Table 1. It seems likely that the success of this particular combination derives from the fact that on average 1SGZ and 2QU2 both assign particularly high ranks to certain chemotypes that are generally poorly ranked by the other receptors (arylpiperazine and acylguanidine, respectively, see Figure 7).

cAbl. The cAbl PCA plot in Figure 5 was used to implement strategy A by combining receptors with similar conformations of the activation and glycine-rich loops. As reported in Table 6, three out of the eight ensembles constructed in this way met the better-than-all criterion. Each of these three ensembles corresponds to a different conformation of the glycine-rich loop. Furthermore, two of them represent kinase-active structures and one is kinase-inactive. Thus, there does not appear to be any particular backbone conformation that leads to successful ensemble construction via strategy A. The same PCA analysis was used to construct 13 ensembles through strategy B. Of these, 6 yielded AUC values that were significantly higher than those of all of their constituents. Five of these six ensembles are composed of one kinase-active structure and one kinase-inactive structure. Such combinations presumably succeed because the active and inactive constituent receptors are respectively specialized to bind Type I (active) and Type II (inactive) kinase inhibitors. With regard to the glycine-rich loop, four of the five conformations observed in the PCA analysis (Figure 5) are found in at least one if these successful ensembles.

In marked contrast to the results for BACE, 8 out of the 11 cAbl ensembles constructed using strategy C (improve worst single receptor) and 5 of the 12 ensembles constructed using strategy D (improve best single receptor) met the better-than-all criterion. Thus, in the case of cAbl, taking the receptor with the highest AUC value and augmenting it

within an additional structure might be a productive way to construct ensembles, were it not for the problems with using this approach prospectively (noted above). A chemotype fractional rank analysis analogous to that presented for BACE in Figure 7 was used to prepare seven ensembles according to strategy E, that is, by combining cAbl receptors that perform well for different chemotypes. Of the four such ensembles that gave an AUC value that was significantly higher than those of all constituents, three involved more than one conformation of the activation loop.

Particularly noteworthy is that all except one of cAbl ensembles (constructed using strategy A) produced an AUC that significantly improved upon the mean AUC of their constituents. Overall, in comparison to the BACE ensembles, a greater percentage of the cAbl ensembles are successful. As for the AUC analysis of single receptor structures in the previous subsection, this has two principal causes (see eq 13). First, the differences between AUC values of the ensembles and their constituents receptors (ΔAUC) are larger on average for cAbl. Second, the standard errors in the differences in AUC values (SE_Δ in eq 12) are smaller. Both observations may in part be the result of differences in the composition and diversity of the two test-sets, fundamental differences in the chemical physics of binding to BACE and cAbl, and also differences in the accuracy with which the GlideScore scoring function represents these binding differences. Such factors may well effect differences in SE_Δ, $Var_{\Delta,a}$, and $Var_{\Delta,d}$. However, as discussed above, one reason for the lower average value of SE_Δ is certainly the increased number of actives in the cAbl test-set.

IFD-derived Ensembles. This subsection presents an approach to ensemble construction that uses an induced-fit docking (IFD) protocol[54] to prepare a set of conformationally diverse receptor structures. In IFD, after an initial rigid-receptor docking, induced-fit effects are simulated by allowing the protein structure to "relax" around the pose adopted by the ligand. In the particular IFD method used here, this relaxation is achieved with a force-field-based energy minimization and is then followed by a final (rigid-receptor) redocking. The result is a protein conformation optimized to the pose of the particular ligand. For further details of the IFD protocol, see ref 11. Direct use of IFD to dock an entire test set of hundreds of compounds would be computationally prohibitive. It is worth emphasizing that in this work IFD is used only to derive the receptor ensemble and *not* in the subsequent virtual screening. In particular, the strategy for ensemble construction involves performing IFD using one representative ligand from each of the active chemotypes in the test-set. For each chemotype representative, the protein conformation of the top-scoring protein−ligand pose is added to the ensemble.[55]

A related idea was sketched previously.[26] In that work, two-member ensembles were composed from a crystal-structure receptor plus another derived by employing a single active ligand in an analog of the IFD-based protocol described here. In contrast, the current work considers ensembles composed of several receptors derived using sets of several structurally diverse ligands. These larger ensembles should be more able to recognize multiple active chemotypes and thus yield better enrichments.

β-Secretase. For each of the seven active chemotypes in the BACE test set, the representative compound was chosen as that with the lowest IC_{50} value. The 1W51-derived receptor was chosen as the initial protein structure for the IFD protocol. This receptor produced the lowest AUC value of those shown in Table 1 and thus there is considerable scope for improving the enrichment.

Table 7. BACE Ensemble ensIFDa Constructed Using the Induced-Fit Docking (IFD) Approach with Receptor 1W51 as the Initial Structure: *p*-Values for Differences in the AUC Values of the Ensemble, the Crystallographically Derived Single Receptors, and the Constituent IFD-derived Single Receptors[a]

ensemble	AUC	PDB	AUC	ΔAUC	*p*-value
ensIFDa	0.824	2Q11	0.774	0.050	0.250
		1XS7	0.757	0.067	0.062
		1SGZ	0.740	0.084	0.093
		1M4H	0.748	0.076	0.014*
		2VA7	0.730	0.094	0.002*
		2IQG	0.732	0.092	0.014*
		2VJ7	0.742	0.082	0.054
		2QU2	0.717	0.107	0.058
		3E3W	0.714	0.110	0.000*
		1YM4	0.725	0.099	0.001*
		1FKN	0.723	0.101	0.004*
		2QP8	0.697	0.126	0.002*
		1W50	0.686	0.138	0.000*
		1W51	0.688	0.135	0.000*
		ensIFDaag	0.734	0.090	0.002*
		ensIFDaap	0.696	0.128	0.000*
		ensIFDahe	0.760	0.064	0.013*
		ensIFDahea	0.780	0.043	0.130
		ensIFDain	0.753	0.071	0.036*
		ensIFDaip	0.716	0.108	0.002*
		ensIFDara	0.755	0.069	0.004*

[a] The *p*-values are two-sided for H_1: ΔAUC not equal to 0, where ΔAUC is determined using eq 9 rather than eq 7. The *p*-values below 0.05 are marked with an asterisk to emphasize significance at the 95% level. The reference (i.e., training) compounds used in the IFD protocol were excluded from the analyses used to produce the values in this table. The constituent IFD-derived single receptors are labelled ensIFDa*x*, where *x* indicates the chemotype used to prepare the receptor (ag = acylguanidine, ap = arylpiperazine, he = hydroxyethyl, hea = hydroxyethylamine, in = isonicotinamide, ip = isophthalamide, ra = reduced amide).

According to the procedure described above, the induced-fit docking of the seven chemotype representatives produced a seven-membered ensemble ("ensIFDa"). This achieved an AUC of 0.824, which is significantly higher than that of the initial 1W51 receptor. As reported in Table 7, ensIFDa also yielded an AUC value higher than any other single receptor. More precisely, the AUC value of ensIFDa is significantly higher than that of 9 of the 14 crystallographically derived single receptors (including 1W51). The AUC of 0.824 also equals the highest AUC value obtained from the BACE ensembles of the previous section. Therefore, using this IFD-based ensemble construction approach, optimal enrichment performance has been achieved starting from the worst performing single receptor structure. Table 7 also shows that ensIFDa has a significantly higher AUC value than 6 out of 7 of its constituent receptors. This performance is in marked contrast to that of almost all the ensembles considering in the previous subsection.

The structural variation among the members of the ensemble was characterized by measuring side-chain torsion angles of important and highly flexible active-site residues (see Supporting Information (Table S11)). In broad terms,

the torsional angle distributions derived from the ensIFDa receptors tend to overlap with the distributions derived from the 14 crystallographic protein structures. Considering the root-mean-square deviation (rmsd) from the 1W51 reference structure in this torsion angle space, similar ranges of values are covered by the two sets of structures. However, during the course of the IFD procedure some chemotypes induce much larger modifications to the starting structure than others. The smallest rmsd is observed with the hydroxyethylamine chemotype; the cocrystal ligand in 1W51 is also a hydroxyethylamine.

A second application of the IFD-based ensemble construction approach used 2Q11 as the initial structure. This receptor produced the highest AUC value in Table 1 and therefore leaves less room for improved enrichment, and more room for deterioration, then 1W51. The AUC metrics and comparisons for the resulting ensemble (ensIFDb) are reported in the Supporting Information (Table S10), but the pattern which emerges matches that found for ensIFDa. In particular, the ensemble ensIFDb produces a significantly higher AUC value than (i) 11 out of the 14 crystallographically derived receptors and (ii) 5 out of 7 of the constituent IFD-derived receptors.

cAbl. To derive ensembles of the same size as those constructed for BACE, the seven most structurally diverse core representatives (see section on ligand test-set construction above) of the 14 cAbl chemotypes were employed in the IFD-based ensemble construction protocol described above. Several ensembles were constructed using a number of different cAbl receptors (Table 3) as the initial protein structure for the induced-fit docking. Table 8 reports analysis of ensemble derived from the 2QOH receptor (ensIFDc), which yielded an AUC of 0.937. This ensemble has a significantly higher AUC value than (i) the initial structure 2QOH, (ii) all the other single crystallographically derived cAbl receptors, and (iii) 6 out of 7 of the constituent IFD-derived receptors. Application to the 3CS9 structure produced a similar pattern, except that the resulting ensemble had an AUC value that was only significantly better than 11 out of 13 of the single receptors (including 3CS9) and than four out of seven of its constituents. However, application to 2G2F produced a qualitatively less successful ensemble with an AUC value significantly higher only than (i) the initial structure 2G2F, (ii) four of the single receptors, and (iii) five of seven constituents.

Just as for strategy E of the previous subsection, a clear drawback to this IFD-based approach is that it requires a set of known actives (chemotypes). Its success presumably also depends on how well the actives cover the chemical space spanned by the set of true (and unknown) actives in a prospective virtual screen. Table 9 presents some relevant analysis. First, it shows that each constituent receptor of ensIFDa has a higher AUC than the initial 1W51 receptor (four significantly so), even though in each case the protein conformation was only optimized to match one of the test-set chemotypes. Second, the table reports AUC metrics and comparisons for ensembles constructed from some but not all of the ensIFDa constituents. Four approaches to selecting three out of the seven original ensIFDa receptors are adopted: ensIFDa-x1 and ensIFDa-x2 include the receptors with the three highest and three lowest AUC values, respectively. In addition, ensIFD-x3 and ensIFD-x4 include the receptors derived from the three chemotypes with the highest and lowest structural similarity with the other chemotypes, respectively. In this case, structural similarity was quantified using Tanimoto coefficients between EPFP_6 fingerprints. All four three-member ensembles have a significantly higher

Table 8. cAbl Ensemble ensIFDc Constructed Using the Induced-Fit Docking (IFD) Approach with Receptor 2QOH as the Initial Structure: *p*-Values for Differences in the AUC Values of the Ensemble, the Crystallographically Derived Single Receptors, and the Constituent IFD-derived Single Receptors[a]

ensemble	AUC	PDB	AUC	ΔAUC	*p*-value
ensIFDc	0.937	2QOH	0.911	0.030	0.007*
		3CS9	0.895	0.041	0.002*
		2HYY	0.885	0.051	0.001*
		2GQG	0.884	0.052	0.001*
		2HIW	0.866	0.071	0.000*
		1IEP	0.868	0.072	0.000*
		1FPU	0.867	0.070	0.000*
		1M52	0.860	0.081	0.000*
		2HZI	0.854	0.086	0.000*
		2HZ0	0.834	0.114	0.000*
		2G2H	0.831	0.110	0.000*
		2G2F	0.823	0.114	0.000*
		2G1T	0.670	0.267	0.000*
		ensIFDc1	0.924	0.017	0.048*
		ensIFDc2	0.903	0.038	0.007*
		ensIFDc6	0.886	0.051	0.000*
		ensIFDc8	0.871	0.070	0.000*
		ensIFDc11	0.923	0.018	0.081
		ensIFDc17	0.900	0.036	0.000*
		ensIFDc18	0.871	0.066	0.000*

[a] The *p*-values are two-sided for H_1: ΔAUC not equal to 0, where ΔAUC is determined using eq 9 rather than eq 7. The *p*-values below 0.05 are marked with an asterisk to emphasize signficance at the 95% level. The reference (i.e., training) compounds used in the IFD protocol were excluded from the analyses used to produce the values in this table. The constituent IFD-derived single receptors are labelled ensIFDc*x*, where *x* indicates the number of the chemotype used to prepare the receptor.

Table 9. Individual Constituents of the IFD-derived BACE Ensemble ensIFDa and Reduced Three-Membered Ensembles Compared to the Initial 1W51 Structure

initial structure	AUC	constituent/ ensemble	AUC	ΔAUC	*p*-value
1W51	0.688	ensIFDaag	0.734	−0.045	0.228
		ensIFDaap	0.696	−0.007	0.895
		ensIFDahe	0.760	−0.072	0.013*
		ensIFDahea	0.780	−0.092	0.007*
		ensIFDain	0.753	−0.065	0.026*
		ensIFDaip	0.716	−0.028	0.406
		ensIFDara	0.755	−0.066	0.023*
		ensIFDa-x1	0.802	−0.113	0.000*
		ensIFDa-x2	0.781	−0.092	0.037*
		ensIFDa-x3	0.769	−0.081	0.001*
		ensIFDa-x4	0.787	−0.098	0.027*

p-Values for differences in the AUC values are two-sided for H_1: ΔAUC not equal to 0, where ΔAUC is determined using eq 9 rather than eq 7. The *p*-values below 0.05 are marked with an asterisk to emphasize signficance at the 95% level. The reference (i.e., training) compounds used in the IFD protocol were excluded from the analyses used to produce all the values in this table. Constituents and reduced ensembles are labelled as for Table 7 and as described in the text.

AUC value than the initial 1W51 structure. This suggests that the IFD-derived ensembles are to some extent able to assign high ranks to compounds from chemotypes that were not represented in the IFD ensemble construction phase. However, further work will be needed to confirm these initial results.

CONCLUSIONS

The virtual screening performance of ensemble docking has been compared with that of standard single rigid-receptor docking to establish if and when docking to multiple receptor conformations is worthwhile. The study considered receptor ensembles constructed from crystallographically derived protein structures of the aspartic protease β-secretase and the cAbl kinase domain. Virtual screening enrichment was quantified using the area under curve (AUC) metric of receiver operating characteristic plots.[29] Since the standard error of the AUC is straightforward to estimate via simple analytical expressions,[20] standard statistical tests could be used to quantitatively compare the enrichment provided by different docking methods and to identify statistically significant differences in virtual screening performance.

Another aspect of this evaluation was the consideration of several rational strategies for constructing the receptor ensembles (see Table 5). Those based purely on protein structural information, which therefore avoided using any information about known active compounds, did indeed produce some ensembles which yielded better enrichment (i.e., significantly higher AUC values) than docking into any individual member of the ensemble. However, for BACE, the protein-structure-derived ensembles rarely met this "better than all" success criterion. They were notably more prevalent in the case of cAbl, but still comprised less than 50% of those constructed. Furthermore, it would not have been possible to prospectively predict which of these ensembles would be successful in this way. This could only be rationalized retrospectively by considering the relative performance of each constituent receptor for the various active chemotypes in the test-set (e.g., Figure 7). A second, rather different success criterion judged an ensemble as successful if it yielded an AUC value that was significantly higher than the mean AUC value of its constituents. Many BACE ensembles and all except one of the cAbl ensembles, satisfied this "better than average" condition.

Previous work also noted the difficulty of constructing useful ensembles with methods based only on protein structural (i.e., conformational) information.[28] A proposed solution was to employ ensembles that optimized the mean docking score of the top 1% of the ranked list (regardless of whether these compounds were active or inactive). This approach was motivated by a retrospectively observed correlation between this mean docking score and the chosen enrichment metric (a modified enrichment factor). In this study, however, no correlation was observed between the mean docking score and the AUC enrichment metric (see Supporting Information (Tables S1−S3 and S5)).

By far the most promising ensemble construction strategy considered here is that based on an induced-fit docking protocol.[11] In particular, the IFD-derived ensembles tend to deliver an AUC value significantly higher than that of (i) most of the single crystallographically derived receptors and (ii) most of their constituent receptors. With one exception, this outcome was reached regardless of the initial receptor structure chosen for the IFD protocol. The second property here is particularly noteworthy since it suggests that IFD-derived ensembles frequently provide better virtual screening performance than their constituent receptors. In this sense and in contrast to most of the ensembles constructed using other strategies, docking to IFD-derived ensembles seems to perform as ensemble docking has often been assumed to perform. As discussed above, however, it does suffer from the need for a set of known active compounds with which to initiate the IFD phase.

One obvious hypothesis for the apparent success of this IFD-based ensemble construction approach is that it really does adapt the protein conformation toward the true experimentally observed binding conformation for each ligand or chemotype with which it is presented. If a sufficient diversity of ligands are included in the IFD protocol then the resulting ensemble would then contain at least one appropriate receptor conformation for every ligand subsequently screened. A subtly different and perhaps more likely explanation for its success may be that the protein structure is adjusted to optimize protein−ligand interactions according to the given (imperfect) scoring function, in this case GlideScore. In this scenario, the IFD-optimized protein−ligand poses might not be any closer to the experimentally observed binding modes, and may instead reflect the composition, bias, and idiosyncracies of the scoring function. In addition to comparing to experimental crystal structure poses, one way to investigate this issue would be to perform the virtual screening with a different scoring function to that used in the IFD phase. In any case, despite the initial promise shown here, more work is need to properly validate the IFD-based strategy for ensemble construction and establish its ability to identify novel chemotype hits.

It is worth considering the weaknesses of this current study to identify directions for future research. First, in a couple of respects the generality of the conclusions is somewhat limited. Clearly, since only two test systems have been used, caution must be exercised in extrapolating from these findings to make assumptions about the performance of ensemble docking against other proteins. Moreover, within the scope of this project it has not been possible to use more than one docking program. It is conceivable that alternative docking algorithms (or parameter settings) may lead to different conclusions. The same holds true for any forthcoming scoring functions that improve upon the current treatments of binding-site desolvation effects and the changes in protein and ligand conformational entropy upon binding.[56,57] One caveat common to all evaluations of docking methods also holds here: despite the efforts to maximize structural diversity, any finite set of ligands can only represent an incomplete coverage of chemical space. Alternative sets of actives and decoys may conceivably lead to different conclusions.

A second consideration is that, by adopting the AUC metric, comparisons between the "early enrichment" performance of different receptors/ensembles are largely absent from this work. It would therefore be interesting to apply one of the early enrichment metrics to the ranked lists derived here, preferably one for which analytical error estimates are available (e.g., BedROC[58]). However, the recently observed

correlation between the AUC and BedROC over a data set covering hundreds of virtual screens suggests that the conclusions derived from such an analysis may not differ qualitatively from those reached here.[20] Third, additional ensemble construction strategies can be envisaged. In particular, it would be worthwhile to develop an approach based only on protein structural information which nevertheless took account of the positions of active-site side-chains, in contrast to the unsuccessful approaches used here which only used information on backbone conformation.

A further weakness is that the scoring function used here does not include any estimate of the potential energy of the receptor conformation.[23,25] Including such terms may reduce the incidence of false positives because it would disfavor the high-energy protein conformations that may be responsible for binding inactive/decoy compounds via interactions that are erroneously over-rewarded by imperfect scoring functions. Additionally, this study has focused on receptor ensembles constructed from publicly available protein crystal structures in the RSCB PDB.[40] These structures may provide only a limited coverage of the conformational space accessible to the protein. For this reason, a follow-up study will consider molecular dynamics simulations as an alternative source of protein conformations for ensemble docking.[6,7]

While bearing these limitations in mind, if the results found here can be assumed to hold more widely, then some important implications follow. First, and least specifically, caution would appear advisible when considering ensemble docking for use in an operational virtual screening program. Naive construction of an ensemble via one of the simple approaches used above should not be expected to produce an enrichment that is significantly better than those of *all* the constituent receptors (the better-than-all criterion). In particular, those ensemble construction strategies based purely on protein structural information do not seem capable of reliably producing ensembles that are successful in this way. A more likely, but not assured, outcome is that the ensemble will yield an enrichment which is higher than the *mean* enrichment of its constituents (the better-than-average criterion). As discussed at more length above, this improvement on the average enrichment may indeed be rather valuable. However, the more stringent better-than-all criterion is more selective for ensembles that are genuinely able to retrieve multiple chemotypes by utilizing multiple receptor structures representing alternate conformational states of the protein. The fact that far fewer ensembles succeed under this measure suggests that it may be incorrect to assume that ensemble docking is reliably able to account for protein flexibility and deliver multiple chemotypes or multiple binding modes from a virtual screen. Finally, if a set of known active compounds is available and if it can be assumed that they provide at least a reasonable coverage of the relevant chemical space, then the IFD-based approach to ensemble construction does appear to hold some promise.

ACKNOWLEDGMENT

The authors thank Peter Gedeck and Peter Hunt for useful discussions regarding the design, analysis, and interpretation of this study. Marcel Verdonk, Paul Mortenson, Jason Cole, John Liebeschuetz, and Oliver Korb also gave helpful suggestions for analysing the data. Financial support from the Education Office of the Novartis Institutes for Biomedical Research is gratefully acknowledged.

Supporting Information Available: Prepared structures of the BACE test-set ligands, prepared protein structures for all BACE and cAbl receptors and GlideScore ranked lists for each single receptor and each ensemble. This material is available free of charge via the Internet at www.soton.ac.uk/~chemphys/jessex/resources.html. Additional details of the evaluation methodology and more extensive tables of results. This material is available free of charge via the Internet at http://pubs.acs.org

REFERENCES AND NOTES

(1) Carlson, H. A. Protein flexibility and drug design: How to hit a moving target. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447–452.
(2) Teodoro, M. L.; Kavraki, L. E. Conformational flexibility models for the receptor in structure based drug design. *Curr. Pharm. Des.* **2003**, *9*, 1635–1648.
(3) Wong, C. F. Flexible ligand-flexible protein docking in protein kinase sysems. *Biochim. Biophys. Acta* **2008**, *1784*, 244–251.
(4) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. Target flexibility: An emerging consideration in drug discovery and design. *J. Med. Chem.* **2008**, *51*, 6237–6255.
(5) Pang, Y. P.; Kozikowski, A. P. Prediction of the binding-sites of huperzine-A in acetylcholinesterase by docking studies. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 669–681.
(6) Yoon, S.; Welsh, W. J. Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 88–96.
(7) Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenze neuraminidase. *J. Med. Chem.* **2008**, *51*, 3878–3894.
(8) Totrov, M.; Abagyan, R. Flexible ligand docking to multiple receptor conformations: A practical alternative. *Curr. Opin. Struct. Biol.* **2008**, *18*, 178–184.
(9) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein−ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.
(10) Shehu, A.; Kavraki, L. E.; Clementi, C. On the characterization of protein native state ensembles. *Biophys. J.* **2007**, *92*, 1503–1511.
(11) 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.
(12) Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. A new method for ligand docking to flexible receptors by dual analine scanning and refinement (SCARE). *J. Comput.-Aided Mol. Des.* **2008**, *22*, 311–325.
(13) Cavasotto, C. N.; Kovacs, J. A.; Abagyan, R. A. Representing receptor flexibility in ligand docking through relevant normal modes. *J. Am. Chem. Soc.* **2005**, *127*, 9632–9640.
(14) May, A.; Zacharias, M. Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins: Struct. Funct. Bioinf.* **2008**, *70*, 794–809.
(15) Damm, K. L.; Carlson, H. A. Exploring experimental sources of multiple protein conformations in structure-based drug design. *J. Am. Chem. Soc.* **2007**, *129*, 8225–8235.
(16) Huang, S.-Y.; Zou, X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins: Struct. Funct. Bioinf.* **2007**, *66*, 399–421.
(17) Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *J. Med. Chem.* **2009**, *52*, 397–406.
(18) Smith, G. R.; Sternberg, M. J. E.; Bates, P. A. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J. Mol. Biol.* **2005**, *347*, 1077–1101.
(19) Alonso, H.; Bliznyuk, A. A.; Gready, J. E. Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev.* **2006**, *26*, 531–568.
(20) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
(21) Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Bio.* **1997**, *266*, 424–440.

(22) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377–395.
(23) Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. Testing a flexible receptor docking algorithm in a model binding site. *J. Mol. Biol.* **2004**, *337*, 1161–1182.
(24) Corbeil, C. R.; Englebienne, P.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model.* **2007**, *47*, 435–449.
(25) Barril, X.; Morley, S. D. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J. Med. Chem.* **2005**, *48*, 4432–4443.
(26) Cavasotto, C.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337*, 209–225.
(27) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47*, 5076–5084.
(28) Rao, S.; Sanschagrin, P. C.; Greenwood, J. R.; Repasky, M. P.; Sherman, W.; Farid, R. Improving database enrichment through ensemble docking. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 621–627.
(29) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.
(30) Liu, T.; Lin, Y.; Wen, X.; Gilson, M. K. BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.
(31) Stauffer, S. R.; Stanton, M. G.; Gregro, A. R.; Steinbeiser, M. A.; Shaffer, J. R.; Nantermet, P. G.; Barrow, J. C.; Rittle, K. E.; Collusi, D.; Espeseth, A. S.; Lai, M.-T.; Pietrak, B. L.; Holloway, M. K.; McGaughey, G. B.; Munshi, S. K.; Hochman, J. H.; Simon, A. J.; Selnick, H. G.; Graham, S. L.; Vacca, J. P. Discovery and SAR of isonicotinamide BACE-1 inhibitors that bind β-secretase in a N-terminal 10s-loop down conformation. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1788–1792.
(32) Coburn, C. A.; Stachel, S. J.; Jones, K. G.; Steele, T. G.; Rush, D. M.; DiMuzio, J.; Pietrak, B. L.; Lai, M.-T.; Huang, Q.; Lineberger, J.; Jin, L.; Munshi, S.; Holloway, M. K.; Espeseth, A.; Simon, A.; Hazuda, D.; Graham, S. L.; Vacca, J. P. BACE-1 inhibition by a series of ψ[CH_2NH] reduced amide isosteres. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 3635–3638.
(33) Maillard, M. C.; Hom, R. K.; Benson, T. E.; Moon, J. B.; Mamo, S.; Bienkowski, M.; Tomasselli, A. G.; Woods, D. D.; Prince, D. B.; Paddock, D. J.; Emmons, T. L.; Tucker, J. A.; Dappen, M. S.; Brogley, L.; Thorsett, E. D.; Jewett, N.; Sinha, S.; John, V. Design, synthesis, and crystal structure of hydroxyethyl secondary amine-based peptidomimetic inhibitors of human β-secretase. *J. Med. Chem.* **2007**, *50*, 776–781.
(34) Hom, R. K.; Gailunas, A. F.; Mamo, S.; Fang, L. Y.; Tung, J. S.; Walker, D. E.; Davis, D.; Thorsett, E. D.; Jewett, N. E.; Moon, J. B.; John, V. Design and synthesis of hydroxyethylene-based peptidomimetic inhibitors of human β-secretase. *J. Med. Chem.* **2004**, *47*, 158–164.
(35) Garino, C.; Tomita, T.; Pietrancosta, N.; Laras, Y.; Rosas, R.; Herbette, G.; Maigret, B.; Qulver, G.; Iwatsubo, T.; Kraus, J.-L. Naphthyl and courmarinyl biarylpiperazine derivatives as highly potent human β-secretase inhibitors. Design, synthesis, and enzymatic BACE-1 and cell assays. *J. Med. Chem.* **2006**, *49*, 4275–4285.
(36) Cole, D. C.; Manas, E. S.; Stock, J. R.; Condon, J. S.; Jennings, L. D.; Aulabaugh, A.; Chopra, R.; Cowling, R.; Ellingboe, J. W.; Fan, K. Y.; Harrison, B. L.; Hu, Y.; Jacobsen, S.; Jin, G.; Lin, L.; Lovering, F. E.; Malamas, M. S.; Stahl, M. L.; Strand, J.; Sukhdeo, M. N.; Svenson, K.; Turner, M. J.; Wagner, E.; Wu, J.; Zhou, P.; Bard, J. Acylguanidines as small-molecule β-secretase inhibitors. *J. Med. Chem.* **2006**, *49*, 6158–616.
(37) Rajapakse, H. A.; Nantermet, P. G.; Selnick, H. G.; Munshi, S.; McGaughey, G. B.; Lindsley, S. R.; Young, M. B.; Lai, M.-T.; Espeseth, A. S.; Shi, X.-P.; Colussi, D.; Pietrak, B.; Crouthamel, M.-C.; Tugusheva, K.; Huang, Q.; Xu, M.; Simon, A. J.; Kuo, L.; Hazuda, D. J.; Graham, S.; Vacca, J. P. Discovery of oxadiazoyl tertiary carbinamine inhibitors of β-secretase (BACE-1). *J. Med. Chem.* **2006**, *49*, 7270–7273.
(38) Irwin, J. J.; Shoichet, B. K. ZINC—A free database of commerically available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(39) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(40) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(41) Jolliffe, I. T. Definition and derivation of principal components. In *Principal Component Analysis*, 2nd ed.; Springer: New York, 2002; pp 1—6.
(42) Gorfe, A. A.; Caflisch, A. Functional plasticity in the substrate binding site of β-secretase. *Structure* **2005**, *13*, 1487–1498.
(43) McGaughey, G. B.; Colussi, D.; Graham, S. L.; Lai, M.-T.; Munshi, S. K.; Nantermet, P. G.; Pietrak, B.; Rajapakse, H. A.; Selnick, H. G.; Stauffer, S. R.; Holloway, M. K. β-secretase (BACE-1) inhibitors: Accounting for 10s loop flexibility using rigid active sites. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1117–1121.
(44) *Maestro*, version 8.5; Schrödinger, LLC: New York, 2008.
(45) *LigPrep*, version 2.2; Schrödinger, LLC: New York, 2005.
(46) Park, H.; Lee, S. Determination of the active site protonation state of β-secretase from molecular dynamics simulation and docking experiment: Implications for structure-based inhibitor design. *J. Am. Chem. Soc.* **2003**, *125*, 16416–16422.
(47) Polgar, T.; Keseru, G. M. Virtual screening for β-secretase (BACE1) inhibitors reveals the importance of protonation states at Asp32 and Asp228. *J. Med. Chem.* **2005**, *48*, 3749–3755.
(48) *Glide*, version 5.0; Schrödinger, LLC: New York, 2008.
(49) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.
(50) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: Pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179–190.
(51) *Pipeline Pilot*, version 7.0; Accelrys Software Inc.: San Diego, CA, 2008.
(52) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36.
(53) Moore, P.; Cobby, J. Comparison of sample means. In *Introductory Statistics for Environmentalists*; Prentice Hall Europe: Hemel Hempstead, U.K., 1998; pp 107—112.
(54) *Schrödinger Suite 2008 Induced Fit Docking protocol; Glide*, version 5.0; Schrödinger, LLC: New York, 2005; *Prime*, version 1.7; Schrödinger, LLC: New York, 2005.
(55) The Schrödinger IFD score is a linear combination of the GlideScore docking score and the Prime (receptor optimization) score.[11]
(56) Huang, S.-Y.; Zou, X. Inclusion of solvation and entropy in the knowledge-based scoring function for protein—ligand interactions. *J. Chem. Inf. Model.* **2010**, *50* (2), 262–273, 10.1021/ci9002987.
(57) Trbovic, N.; Cho, J.-H.; Abel, R.; Friesner, R. A.; Rance, M.; Palmer III, A. G. Protein sidechain dynamics and residual conformational entropy. *J. Am. Chem. Soc.* **2009**, *131*, 615–622.
(58) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early enrichment" problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

CI900407C