Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4

Nicolas Triballeau,*,†,‡ Francine Acher,† Isabelle Brabet,§ Jean-Philippe Pin,§ and Hugues-Olivier Bertrand‡

Laboratoire de Chimie et Biochimie Pharmacologiques et Toxicologiques, UMR8601-CNRS, Université René Descartes—Paris V, 75270 Paris Cedex 06, France, Accelrys SARL, 91893 Orsay Cedex, France, and Laboratoire de Génomique Fonctionnelle, Département de Pharmacologie Moléculaire, CCIPE, 34094 Montpellier Cedex 5, France

Received November 10, 2004

The "receiver operating characteristic" (ROC) curve method is a well-recognized metric used as an objective way to evaluate the ability of a given test to discriminate between two populations. This facilitates decision-making in a plethora of fields in which a wrong judgment may have serious consequences including clinical diagnosis, public safety, travel security, and economic strategies. When virtual screening is used to speed-up the drug discovery process in pharmaceutical research, taking the right decision upon selecting or discarding a molecule prior to in vitro evaluation is of paramount importance. Characterizing both the ability of a virtual screening workflow to select active molecules and the ability to discard inactive ones, the ROC curve approach is well suited for this critical decision gate. As a case study, the first virtual screening workflow focused on metabotropic glutamate receptor subtype 4 (mGlu4R) agonists is reported here. Six compounds out of 38 selected and tested in vitro were shown to have agonist activity on this target of therapeutic interest.

Introduction

During World War II, a graphical technique developed from Neyman and Pearson's decision theory[1,2] found its first practical application. At that time, the so-called "receiver operating characteristic" (ROC) curve method helped British radio operators to distinguish between signals due to random interferences and those featuring the approach of warplanes targeting London. Since then, this test evaluation technique has been implemented in a plethora of other fields in which it soon became a gold standard. Starting with psychology[3,4] and radiology,[5,6] it is now used in various disciplines such as medicine,[7] acoustics,[8] meteorology,[9] and criminology[10] to assess the accuracy of a given detection device and subsequently to make better decisions from the provided measures[11] (see ref 12 for a review).

Given its widespread use in many fields, the ROC curve method is surprisingly underused or underexploited when evaluating in silico selection techniques in the drug design area. The vast majority of virtual screening papers employ "enrichment curves" to estimate the ability of the described workflow to retrieve active compounds out of a set of inactives. Others rely on some in-house metrics and the minority resort to the ROC approach. In other words, medicinal and computational chemists suffer from a lack of standard method to evaluate the accuracy of a newly designed in silico "assay". Therefore, it remains extremely difficult to compare a new selection technique with others published on the same therapeutic target.

The objective of this paper is to introduce the advantages and promote the usage of the ROC plots method in drug design and more particularly when evaluating a virtual screening workflow. With this approach, expert modelers and medicinal chemists can select the most promising compounds in a more objective way and concentrate their efforts on the synthesis of molecules that are more likely to be active against the investigated target. Indeed, the ROC curve analysis helps answer two paramount questions: (1) Considering current knowledge, how good is my model at selecting the known active molecules and discarding the inactive ones compared to another model? (2) Where should the score threshold be set between selected molecules that are worth being further tested and those that should be discarded as likely inactives?

Once the theoretical aspect has been depicted, the ROC curve method will be illustrated by the construction of a virtual screening workflow focused on metabotropic glutamate receptor subtype 4 (mGlu4R) agonists.[13] More precisely, we will show how the ROC curve method allowed us to tune some docking parameters to select the most appropriate scoring function for the retrieval of mGlu4R agonists and to set a selection threshold according to practical needs. The resulting workflow was finally used to select compounds from several commercial vendors. The "virtual hits" were purchased and tested in vitro. Some results of this screening campaign are reported here.

Methodology

The scope of virtual screening is to enrich a set of molecules extracted from a database with active compounds by weeding out those that are likely to be inactive, prior in vitro assays. Moreover, when inte-

* To whom correspondence should be addressed. Address: Accelrys SARL, 91893 Orsay Cedex, France. Phone: +33 1 69 35 32 32. Fax: +33 1 69 41 99 09. E-mail: nth@accelrys.com.

† Université René Descartes—Paris V.

‡ Accelrys SARL.

§ CCIPE.

10.1021/jm049092j CCC: $30.25
Published on Web 03/08/2005

Table 1. The Four Steps of the ROC Curve Method To Assess the Performance of Virtual Screening Workflows (Termed "Computer Tests")

step	objective
1. Choice of an activity cutoff according to the needs of biology	Include a pharmacology criterion to ensure relevance and usefulness of the computer test
2. Selection of a suitable sample with active and inactive molecules	Include knowledge of SAR on the target under investigation in the assessment
3. Virtual screening of the sample of molecules	Evaluating compounds of known activities with the designed computer test
4. ROC curve plotting and subsequent analysis	Evaluation of the performance of the test and defining a selection threshold

grated in a drug discovery process, it should be capable of achieving this within a reasonable time frame. Many approaches have been developed to fulfill these expectations.[14−16] However, although data availability may often reduce the choice, selecting the most appropriate method remains a real issue. The ROC curve method, described in this section, is one possible way to tackle this problem.

Before getting into details, it is worth noting that the ROC curve method is applicable only to quantitative approaches for which the test results are numerical. Fortunately this is the most common case in the drug design area[17] where test results are generally provided by a similarity metric[18] (e.g., Tanimoto indices), a fit on a pharmacophore,[19] a QSAR model prediction,[20] or a protein−ligand affinity score.[21] In the case of a quantitative test, different thresholds may be applied to class the compounds as potentially active (to be selected) or inactive (to be discarded). In the present paper, test results will be referred to as "scores" and, for consistency with the ROC curve theory, virtual screening workflows will be termed "computer tests".

Guidelines

The assessment of a computer test by the ROC curves method requires four steps as summarized in Table 1.

Step 1. Choice of a Pharmacological Activity Cutoff. Depending on the pharmacological method that will be further used to evaluate the selected molecules, the first stage consists of choosing an appropriate cutoff between active compounds and those considered as inactive for the target. For instance, if a high-throughput screening (HTS) campaign evaluates the activity at a 10 μM concentration, the activity cutoff should be set to this value to ensure the usefulness of the preceding selection process. This preliminary step complies with the concern about integration of virtual screening and HTS[22] in that the first is adapted to the needs of the second.

Step 2. Selection of a Sample of Molecules. Once this decision has been made, step 2 requires the selection of a sample of molecules (containing both actives and inactives) of relevance to the target under investigation. The sample used for a ROC curve analysis allows performance assessment for a given computer test by analyzing the results obtained with molecules of previously known activities. For obvious reasons, a sample is only a tiny part of the entire chemical space and it should therefore contain the appropriate information in terms of available structure−activity relationships (SAR).

As its first quality, a good sample reflects the pharmacology of the target under investigation. Ideally, it should contain structurally diverse compounds of known activity in order to cover the chemical space as "thoroughly" as possible. The more numerous and diverse the molecules are, the more objective is the analysis and the more it reflects current knowledge (i.e., SAR) regarding a particular target. The performance and accuracy of such a test will increase as the sample is iteratively enriched during the hit/lead finding process. Since a ROC curve analysis evaluates both abilities to select active molecules and to discard inactive ones (as defined in step 1), the second prerequisite for a good sample is that it contains both kinds of molecules, ideally in equal numbers in order not to favor one class over the other. Finally, inactive molecules should not be randomly picked but should rather have a chemical structure similar to the structures of the chosen actives. This last rule goes against usual practice, which consists of seeding some known active compounds in a set of randomly gathered "druglike" molecules that are treated as inactives. In a recent publication, Verdonk et al. pointed out the importance of choosing inactive compounds properly.[23] Indeed, it is more difficult for a computer test to detect an activity "signal" in a set of molecules when it is disrupted by inactive related structures emitting a strong interfering "noise". For instance, it is more challenging for a similarity metric to distinguish an active steroid on the estrogen receptor from another inactive steroid than, for instance, from acetylsalicylic acid! Abiding by this last rule should induce some diversity among inactive compounds that is similar to the diversity of the active molecules.

Step 3. Virtual Screening of the Sample. In step 3, the computer test is applied to the sample to be evaluated in order to calculate the computer score for each molecule. As stated before, the test may resort to many different approaches (similarity, pharmacophore, etc.). Even a combination of techniques may be considered.

Step 4. ROC Curve Analysis. Last, in step 4, the ROC curve method is applied as follows to assess the performance of the test performed in step 3. Figure 1 illustrates the overall technique for theoretical distributions of actives and inactive compounds. For a given selection scoring threshold (panel a), the classification of all compounds are reported in a "confusion matrix" (panel b).

Confusion Matrix. Manallack et al. have been using the concept of confusion matrices in the context of drug design.[24,25] Here, it is used as a tool to comprehend the ROC curve method better as it allows quick calculation of sensitivity and specificity from a comparison between in vitro (active/inactive) and in silico (selected/discarded) classifications (see Figure 1b).

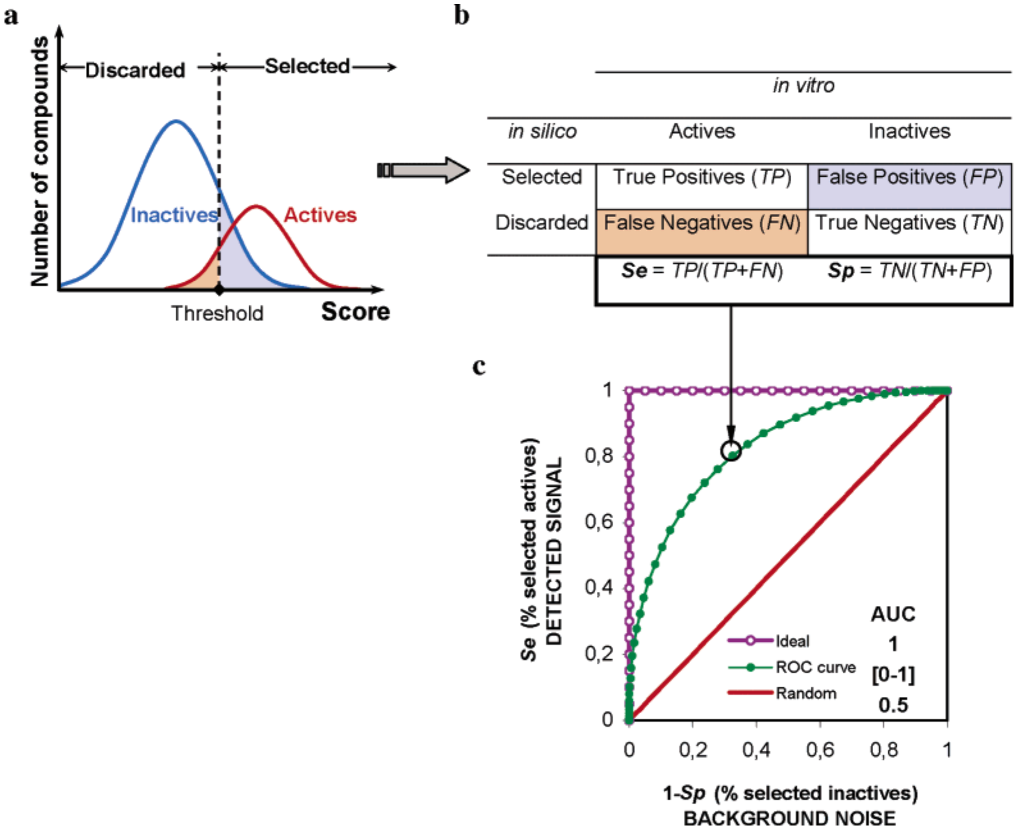


Figure 1. ROC curves in a nutshell. (a) Theoretical distributions of scores are obtained for both actives (red) and inactives (blue) after processing the sample by a suitable computer test. For intelligibility of the figure, it was hypothesized that the scores for both active and inactive compounds had normal (i.e., Gaussian) distributions, although they are unlikely to be so in a usual case. Generally, these distributions overlap, leading to false predictions (colored area). Upon threshold modification (dashed line), proportions of such erroneous classifications (reported in a confusion matrix (b)) change dramatically. (c) For all possible score thresholds, the evolution of the deduced sensitivity (Se) and specificity (Sp) is reported on a ROC graph, Se as a function of 1 − Sp. Calculating the area under the ROC curve is a practical way to quantify the overall performance of the computer test.

Sensitivity and Specificity. Sensitivity and specificity are the main characteristic features of any test. In the drug design context, sensitivity (Se) would be the percentage of truly active compounds being selected from the virtual screening workflow: the number of true positive (TP) results divided by the sum of true positives and false negatives (FN).

$$\mathrm{Se} = \frac{N_{\text{selected actives}}}{N_{\text{total actives}}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Obviously, this fraction can vary between 0 (all actives missed) and 1 (when all actives are selected). Therefore, sensitivity gives information about active molecules that will be missed: the false negatives. The higher is the sensitivity, the lower is this number and the better is the test in selecting actives.

Specificity (Sp), on the other hand, is the percentage of truly inactive compounds being correctly identified by the computer test and therefore being discarded, that is, the number of true negative results (TN) divided by the sum of true negatives and false positives (FP):

$$\mathrm{Sp} = \frac{N_{\text{discarded inactives}}}{N_{\text{total inactives}}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

Specificity can also vary between 0 (all inactives are selected) and 1 (all inactives are discarded), giving insight on inactive compounds that are wrongly classified: the false positives. The higher is the specificity, the lower is this number and the better is the test in discarding inactive compounds.

Upon modification of the selection threshold from the lowest to the highest score provided by the test, sensitivity and specificity will evolve in opposite ways and cover all possible values between 0 and 1. Indeed, when the threshold is set to the lowest score, all compounds are selected whether they are truly actives or inactives, leading to (Se = 1, Sp = 0). Contrastingly, when the threshold is set above the highest score, all molecules are discarded, leading to (Se = 0, Sp = 1). Consequently, it is not possible to optimize both sensitivity and specificity at the same time, and a tradeoff is to be found. The ROC curve analysis allows us to make such a decision by providing a comprehensive picture of the ability of a test to make the distinction over all selection thresholds.[7]

ROC Curves. Plotting a ROC curve consists of reporting the evolutions of sensitivity and specificity together, Se as a function of (1 − Sp). In other words, the activity "signal" (i.e., % actives) is plotted versus the detected "noise" (% inactives) at all possible detection thresholds. A theoretical illustration of such curves is in Figure 1c. On such a graph, a random classification of the compounds would be represented by a diagonal rising from the origin to the upper right corner, whereas a test capable of detecting the correct signal would have a ROC plot that curves above that diagonal. For ideal distributions, where active compounds are completely separated from the inactives, the curve skyrockets vertically to the upper-left corner (Se = Sp = 1) and then joins the upper-right corner horizontally. Hence, the more a ROC curve bends towards the upper left corner of the diagram, the more distinct the signal appears.

In practice, the ROC curve is not as smooth as displayed on the theoretical illustration (Figure 1c), but it is rather jagged and "bumpy". This jagged aspect is due to the discrete values that sensitivity and specificity can only take, reflecting the fact that the confusion matrices are filled with integers. As a matter of fact, as the threshold changes, inclusion of a true positive will lead to a vertical line whereas inclusion of a false positive will produce a horizontal displacement. Less serrated curves are obtained when the sample contains more compounds.

From ROC Curves to Absolute Accuracy of Computer Tests

The ROC curve allows a direct comparison of different computer tests (for instance, different virtual screening workflows). This is because the closer a curve comes to the upper-left corner of the graph, the better the test is at isolating signal from background noise. Thus, such a method provides not only a way to fine-tune the parameters of a given computer test but is also a means of comparing different virtual screening methods (e.g., pharmacophore-based versus docking-based) by plotting their respective curves.

Since the relative positions of ROC plots give an insight into the respective accuracies,[7] the area under the curve (AUC) is a practical way of measuring the overall performance of the tests. If the AUC is close to 0.5 (random test), the test is said to be poor; the highest possible AUC is 1, corresponding to an ideal case. In general, the greater the AUC, the more effective the virtual screening workflow is in discriminating active from inactive compounds. In terms of probabilities, an AUC of 0.9 means that a randomly selected active molecule has a higher score than a randomly selected inactive 9 times out of 10. However, it does not mean that a positive result occurs with a probability of 0.9, nor that a positive result is associated with activity 90% of the time. Indeed, ROC plots characterize the inherent quality of the defined test only and by no means are indicative of the quality of a particular compound. Accessing the activity probability of a given selected molecule (positivity predictive value, PPV) is possible in the rare cases for which the probability of activity (i.e., the yield of actives, Ya) is known for the screened database prior to selection. In those cases, the PPV increases according to both the quality of the computer test (Se, Sp) and the yield of actives in the given database. Similarly, the inactivity probability of a given discarded molecule prior to selection (negative predictive value, NPV) depends on the test being used and on the yield of actives. Mathematically, this is illustrated by Bayes' theorem:

$$\mathrm{PPV} = \frac{(\mathrm{Se})(\mathrm{Ya})}{(\mathrm{Se})(\mathrm{Ya}) + (1 - \mathrm{Sp})(1 - \mathrm{Ya})}$$

$$\mathrm{NPV} = \frac{\mathrm{Sp}(1 - \mathrm{Ya})}{(1 - \mathrm{Se})\mathrm{Ya} + \mathrm{Sp}(1 - \mathrm{Ya})}$$

where

$$\mathrm{Ya} = \frac{N_{\text{actives screened}}}{N_{\text{total screened}}}$$

However, as stated before, the aim of virtual screening being to enrich a set of molecules to be tested in vitro with active compounds, the number of actives in the whole database ($N_{\text{actives screened}}$) is not be known a priori. Consequently, although it pinpoints the importance of the quality of the computer test being used, Bayes' approach is of limited interest for performance assessment of virtual screening workflows. The computer test quality (measured, for instance, by its ROC curve AUC) is the only branch we can hang onto.

In practice, however, the AUC is not enough to certify the quality of a given test as a high AUC value may be obtained by chance if the sample used is too small. For example, if one considers the trivial case where the sample contains only one active and one inactive compound, there are only two possible classifications: the first estimates the active better than the inactive one, exhibiting an AUC of 1, and the second, oppositely, by ranking compounds in the wrong order would have an AUC of 0. Hence, in such circumstances, a good AUC can be easily obtained by sheer chance. In other words, a good AUC should be sustained by a sample of reasonable size. One possible route to statistically validate the computer test is to use Fisher's randomization test that compares the results provided by the test under evaluation to the results provided by multiple random distributions.

Another way to facilitate the choice of the most appropriate method from its ROC curve is to search for the test with the best specificity for a given sensitivity (or vice versa). For instance, if a prerequisite for the test is to have a sensitivity of at least 0.95, one would select the test of highest specificity given that Se > 0.95.

At first glance, ROC curves have a similar outlook as the most commonly used graphical method: the enrichment curves (sometimes termed "cumulative recall curves"). Instead of reporting Se as a function of (1 − Sp), enrichment curves report the yield of actives (Se in fact) as a function of the ranking (or alternatively the percentage of the screened database). Two theoretical enrichment curves are displayed in Figure 2. Like ROC curves, enrichment curves lift from the lowest left corner up to the upper right and above a diagonal line that represents a random ranking. The further away is the enrichment curve from the diagonal, the better is the computer test. Enrichment curves are generally easier to plot, but they suffer from two major drawbacks. First of all, the ideal enrichment curve directly depends on the ratio of actives in the screened set of molecules. Hence, when the ratio of actives increases, the ideal enrichment curve gets closer to the curve featuring the random distribution. As a consequence, enrichment curves are stuck in a narrower space limited by the ideal and the random curves, impairing proper comparisons. The ideal ROC curve, in contrast, is strictly independent

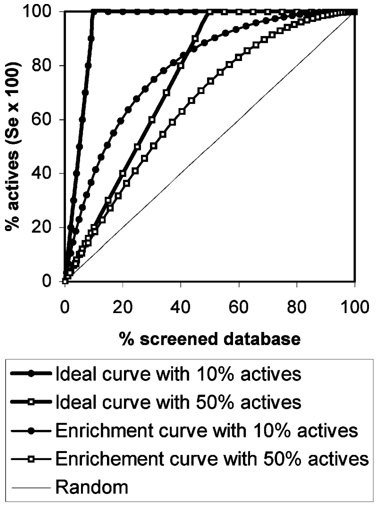


Figure 2. Theoretical enrichment curves of the same test for two different ratios of actives. This graph illustrates the difficulty of comparing test performances by relying on enrichment curves because they directly depend on the ratio of actives in the screened database.

of this ratio and always passes through the upper-left corner of the graph. That explains why ROC curves are said to evaluate the *absolute* accuracy of a test, whereas enrichment curves give only insight on its *relative* accuracy for a given activity ratio. Second, enrichment curves capture explicitly only one of the two aspects depicted by the ROC plots, that is, sensitivity. Because any test has a dual feature, i.e., the ability to retrieve actives (described by Se) and ability to discard inactives (Sp), one may argue that enrichment curves provide only half of the information that is necessary to make good decisions.

Where To Set the Threshold: Making the Right Decision

Current Methods. There is a plethora of metrics that have been developed to find an optimal threshold for quantitative tests.[26−31] Among the most commonly used, the enrichment compares the yield of actives (here defined as sensitivity) afforded by the in silico prescreen with the yield of actives obtained from a random cherry picking. A large majority of these metrics surrounds both concepts of sensitivity and specificity but without using them as such. However, like enrichment curves, many of them suffer from being dependent on the ratio of actives in the data set.

Last, Neyman and Pearson, who pioneered hypothesis testing,[1] asserted that there is no general rule for balancing errors; in any given case, the determination of "how the balance [between wrong and correct classifications] should be struck, must be left to the investigator". In summary, balancing false-positive and false-negative rates has "nothing to do with statistical theory but is based instead on context-dependent pragmatic considerations where informed personal judgment plays a vital role[32]". ROC curves were developed to allow the incorporation of practical considerations in order to make appropriate decisions (see below).

Deciding from the ROC Curve. If the computational test provides a quantitative result (such as a Tanimoto similarity index or a protein−ligand affinity score), the ROC curve offers the possibility of choosing a selection threshold in a very simple way. Indeed, as shown before, any given point on the curve, (1 − Sp, Se), corresponds to a given threshold and vice versa. Therefore, choosing a point on the ROC curve corresponds to choosing a threshold value.

When practical considerations are taken into account, two different attitudes may be adopted depending on the costs and benefits of an error by excess (selecting an inactive for in vitro assays) versus the cost of an error by default (losing an active compound). There are many advantages in adopting a conservative attitude (privileging specificity over sensitivity by requiring a high score) by choosing a point on the lowest left corner of the ROC curve. First of all, a high specificity allows the majority of inactives to be pushed aside, leading to a higher yield of actives (i.e., improves the enrichment). This particular point has been claimed to be the main advantage of in silico selection of compounds prior to HTS. Besides, the number of molecules to be synthesized and tested in vitro is reduced compared to the popular strategy adopted in the 1990s that used to advise screening any available compound. In the case where cost and time are the main issues, for example, in a small company or in highly competitive domains of research, privileging specificity may be advisable. In other words, a conservative attitude (high specificity) is faster, cheaper, motivating, and apparently, the most efficient way to accelerate drug discovery. However, attention should be drawn to the dangers of a drift toward excessive conservatism.

Indeed, the alternative (i.e., choosing a point in the upper-left part of the curve), and more liberal strategy preferring sensitivity to specificity also shows some significant advantages. First of all, such an attitude is a way to account for the uncertainty of models whereas tests of high specificity may lend too much credit to the adopted approximations. Second, when sensitivity is increased, fewer actives are lost, including compounds with more diverse structures. This is not surprising because a model is derived from the known SAR, explaining that related structures are more easily detected than novel chemotypes. This is particularly obvious if the model uses similarity-searching methods. Consequently, when innovation should be privileged, for instance, while seeking scaffold "hopping" during the hit/lead finding stage of the drug discovery process or when patent deposition is the main concern, sensitivity is to be favored over specificity. The drawback here is that excessive liberalism would lead to the acceptance of most (if not all) compounds in the selection, therefore effectively bypassing the virtual screening step in the drug discovery. Naturally, one wise strategy would be to find a compromise solution between these extremes.

Many strategies can be implemented with the ROC curve method, and this theoretical part constitutes only an introduction to this approach. In particular, combinations of test results (with AND or OR association rules) are common practice in clinical diagnosis in order to improve both sensitivity and specificity values.[7] Here, we will simply compare the results obtained with the same technique (namely, docking-scoring) with different combinations of parameters.

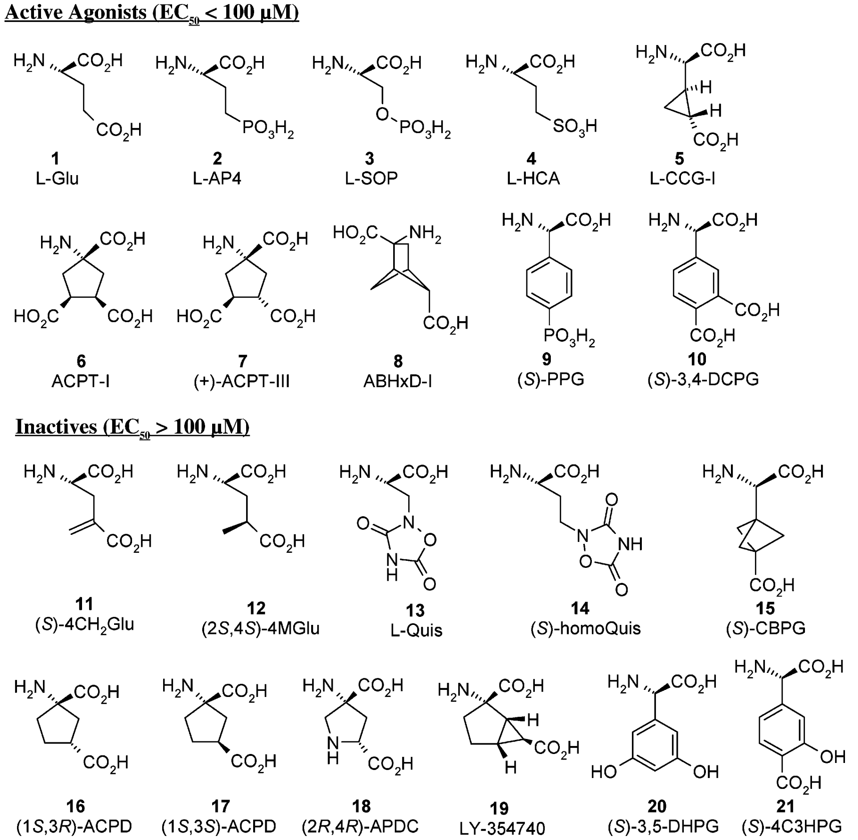


Figure 3. Sample of 21 molecules representing current knowledge about SAR data on mGlu4R. For the purpose of the present study (ROC curve analysis), agonists with EC_{50} below 100 μM are considered actives and those above 100 μM are taken as inactives.

Results and Discussion

Application to mGlu4R Agonists. Metabotropic glutamate receptors (mGluR) are of particular interest in medicinal chemistry because they are believed to be suitable targets for treating a large variety of brain disorders[33,34] such as convulsions,[35] pain,[36,37] drug addiction,[38] anxiety disorders,[39] and several neurodegenerative diseases.[40] Characterized by a large extracellular amino-terminal domain, mGluRs constitute the family 3 (or C) of G-protein-coupled receptors together with $GABA_B$, Ca^{2+} sensor, taste, olfactory, and pheromone receptors.[41] This extracellular domain constitutes the ligand binding module (LBM), the crystal structure of which has been recently resolved.[42] As earlier hypothesized,[43] the LBM adopts a bilobate fold separated by a flexible hinge region. This allows the agonist to be trapped and subsequently trigger the receptor's activation.[44]

On the basis of sequence similarity, transduction mechanism, and pharmacological profile, the eight known subtypes of mGluRs are classified into three groups. Group III contains subtypes 4 and 6−8. Mainly located presynaptically, where they act as autoreceptors,[45] group III mGluRs decrease adenylyl cyclase activity via a $G_{i/o}$ protein and are specifically activated by L-AP4 (see structure **2** in Figure 3). Among this group, mGlu4R is thought to be a possible new target for Parkinson's disease,[40] but the lack of a highly specific agonist has seriously impaired target validation studies. Furthermore, despite many chemical variations around the structure of glutamate (see Figure 3), L-AP4 **2** remains the strongest mGlu4R agonist with an EC_{50} of only 0.32 μM. New chemotypes of higher potency and specificity are to be found, but the poor diversity and the shortage of available SAR have made their research particularly tricky.

To identify new leads for subtype 4, the ROC curve method was applied to fine-tune parameters of a virtual high-throughput screening (vHTS) workflow so that an "activity signal" could be detected among the small number of molecules of known mGluR activity. The resulting vHTS workflow was then applied to select molecules available from five different vendors (more than 720 000 available compounds). The selected molecules were finally purchased and tested in vitro using a calcium imaging HTS approach at a concentration of 100 μM. Since the most potent mGlu4R agonists have EC_{50} values in the low micromolar range, this concentration was estimated to be a good tradeoff between poor agonist activities and activities of hits that could possibly be optimized. As a matter of fact, an EC_{50} of 100 μM barely represents 1 log of activity above L-glutamate **1**.

Sample Building for Computer Test Assessment. In the case of mGlu4R, the small number of agonists of known activity dictates the construction of the sample. In fact, the vast majority of these molecules, reported in different papers (see refs 13 and 46 and references therein), were incorporated in the sample. Since our HTS campaign was conducted at a 100 μm concentration, the limit between active and inactive compounds was set to this value. In other words, molecules with known EC_{50} below 100 μM are considered as actives whereas molecules with known EC_{50} above 100 μM are taken as inactives.

Table 2. Norms of the Three Principal Moments of Inertia (PMI) (in Å) for the Four Reference Agonists Calculated According to Their Docked Conformation by Molecular Dynamics

class	agonist	PMI (relative contribution to inertia)
1	L-Glu	1.89 (53%), 0.85 (24%), 0.80 (23%)
1	L-AP4	1.69 (52%), 0.87 (27%), 0.69 (21%)
2	ACPT-I	1.87 (43%), 1.37 (32%), 1.05 (25%)
3	(*S*)-PPG	2.73 (59%), 1.13 (24%), 0.73 (16%)

Hence, a sample of 21 molecules was derived on the basis of both activity cutoff and pharmacology (refer to Figure 3 for structures and activity classes). Therefore, out of 21 molecules, 10 are regarded as actives and the remaining molecules are regarded as inactives. It is worth noting that all molecules are α-amino acids with various side chains ended by an H-bond acceptor moiety (most often an acidic function). In particular, the inactive molecules were chosen for their structural similarity to active molecules as recommended in the above guidelines. That is, instead of choosing compounds at random to build-up the inactive class, we purposely chose compounds that are more likely to produce a strong interfering "noise" for the computer test we wish to assess. As a matter of fact, the inactive molecules on subtype 4 are known agonists of other mGluRs belonging to the other groups of mGluRs (i.e., group I or group II). Three structural groups were identified among both active and inactive compounds: small and rather flexible molecules such as L-glutamate **1**, L-AP4 **2**, or, among inactives, (*S*)-4-methyleneglutamate ((*S*)-4CH$_2$-Glu **11**); bulky and semirigid structures such as ACPT-I **6** and (1*S*,3*R*)-ACPD **16**; and phenylglycines (e.g., (*S*)-PPG **9**, (*S*)-3,4-DCPG **10**, 3,5-DHPG **20**), which are characterized by a slightly longer distance between the α-amino acid moiety and the distal hydrogen-acceptor function.

Computer Test Development. Structural Issues. To tackle the docking in a reasonable time frame, most high-throughput docking programs consider only the flexibility of the ligands, whereas the protein targets are maintained rigid throughout the calculation. In such a simplified approach, the possible induced fit between ligand and protein is therefore not considered, albeit many cases have been reported to support its importance.[47] The LigandFit program is not an exception to the rule. Here, to take protein flexibility implicitly into account, all compounds were docked in four different and relevant conformations of the receptor ("ensemble docking"). To do so, a comparative model of mGlu4R LBM was refined by molecular dynamic (MD) in the presence of four mGlu4 receptor agonists representing the three structural classes described above: L-glutamate **1** (endogenous ligand) and L-AP4 **2** (the most potent mGlu4R agonist) for the first class, ACPT-I **6** for the second, and (*S*)-PPG **9** for the third (phenylglycines) according to the method published by Bertrand et al.[48] One practical way to compare those agonists in their MD-refined conformation is to calculate their three principal moments of inertia (PMI; see Table 2). Without getting into many details, it is clear that L-glutamate **1** and L-AP4 **2** are similar in terms of both size (as reported by the norms of the PMIs) and shape (as exhibited by the relative parts of inertia). Although (*S*)-PPG **9** is characterized by a similar shape, it is longer than the class 1 agonists, with the largest PMI almost 1 Å larger than for L-glutamate **1** and L-AP4 **2**. In contrast, the largest PMI for ACPT-I **6** is similar to those of class 1 agonists but this compound is notably bulkier along the other axes of inertia. With regard to the receptor model, these ligands induced some conformational modifications mainly in the part of the binding pocket that interacts with the glutamate side chain. It comprises several basic and highly flexible residues (Lys74, Arg78, Lys317, and Lys405), which orient their side chains according to each agonist's structure. A figure representing the binding site of L-AP4 **2** in mGlu4R is available as Supporting Information.

For our vHTS workflow, the four MD-docked agonists were considered as structural references to estimate the quality of the LigandFit docked molecules. In particular, the position of their primary ammonium moiety seems to be highly conserved upon agonist binding to any mGluR LBM thanks to three H-bond interactions (Ala180, Thr182, Asp312 in subtype 4), one ionic bridge (Asp312), and a cation−π interaction (Tyr230). This assertion is strongly supported by site-directed mutagenesis studies that demonstrate the paramount importance of interaction with Thr182 (dramatic loss of activity in the T182A) versus the interaction with the distal binding pocket[49] (with the exception of the R78A mutant). In a more recent study,[44] it was shown that the ionic interaction with Asp309 in subtype 8 (corresponding to Asp312 in subtype 4) strongly stabilizes the closed-activated form of the LBM (EC$_{50}$ for L-AP4 at 0.2 μM for the wild-type versus 129 μM for the D309A mutant).

High-Throughput Docking. The 21 molecules of the sample were processed through the workflow summarized in Figure 4. LigandFit is a shape-based docking engine with a subsequent force-field-driven positioning improvement designed for vHTS.[50] The first step of the calculation defines a binding volume characterized by its three PMIs and therefore by its shape. Since we had four MD-refined conformations of the LBM (termed C$_{Glu}$, C$_{AP4}$, C$_{ACPT}$, and C$_{PPG}$), four site models were derived from LigandFit's cavity detection algorithm. For each ligand, LigandFit generates various conformations using a Monte Carlo search method restricted to the torsion space. Once aligned within the site according to the principal moments of inertia, conformations that match the shape of the site model are refined by a rigid-body minimization. Resulting poses are ranked according to either a force-field-based (DREIDING[51]) or an empirical-based (PLP1[52]) docking function termed DockScore. This docking process allows us to generate reasonable local minima for each ligand, assuming that the bioactive orientation will be among the top 20 poses ranked by the DockScore.

Docking Results Refinement. After the poses are docked with LigandFit in the ensemble of conformations (C$_{Glu}$, C$_{AP4}$, C$_{ACPT}$, and C$_{PPG}$), the poses are relaxed with an in situ minimization protocol (flexible ligand in a rigid protein model) with the DREIDING[51] force field. Since all molecules exhibit a primary ammonium substructure, poses having their ammonium moiety further than a given distance from the ammonium of the MD-docked reference compounds are rejected. Thanks to this structural filter, only high-quality poses are submitted

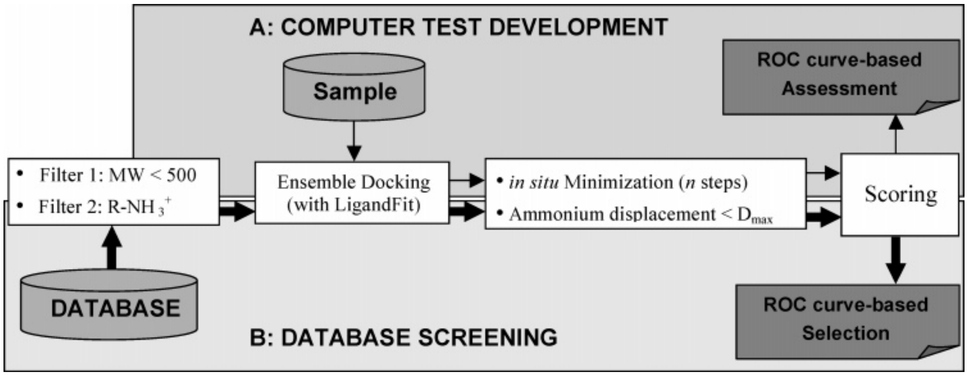

A: COMPUTER TEST DEVELOPMENT
Sample
ROC curve-based Assessment
• Filter 1: MW < 500
• Filter 2: R-NH$_3^+$
Ensemble Docking (with LigandFit)
• *in situ* Minimization (*n* steps)
• Ammonium displacement < D$_{max}$
Scoring
DATABASE
ROC curve-based Selection
B: DATABASE SCREENING

Figure 4. Diagram illustrating the overall workflow described in this paper. Pathway A (above, thin arrows) depicts the workflow implemented to tune the different parameters according to the ROC curve method. Pathway B (below, thick arrows) shows the way used by molecules from commercial databases once the parameters were set with the ROC curve approach. Hence, these molecules are preprocessed prior to docking. Only compounds having a molecular weight (MW) below 500 g/mol (filter 1) and exhibiting a ionizable primary ammonium (filter 2) are allowed to proceed through the next step. Given that the sample of 21 molecules used for ROC curve analysis already satisfies these prerequisites, it is directly input into the ensemble docking step.

Table 3. Envisaged Values for Each of the Five Parameters Considered for the Optimization of the Computer Test

parameter	envisaged values
docking strategy	one protein conformation vs ensemble
docking function (DockScore)	DREIDING or PLP1 derived
number of in situ minimization steps	0, 250, 500, 1000
ammonium displacement ($D_{\max}$), Å	0. 5, 1, 1.5, 2, above 2
scoring functions	LigScore1-2, PLP1-2, Jain, PMF, Ludi2

to the final scoring and prioritizing step. This is consistent with SAR data as discussed earlier.

Scoring. The remaining poses docked in the four receptor conformations are scored by all scoring functions available in the Ligand Scoring module of Cerius^2 (Accelrys Inc.). Indeed there is no clear evidence yet that one scoring function performs better than the other for this family of protein. We therefore envisage force-field-based, knowledge-based, and empirical-based scoring functions.[21] Finally, all poses from the four sites are collected and only the best pose (according to the considered scoring function) for each molecule is retained to plot the corresponding ROC curve, regardless of the receptor conformation.

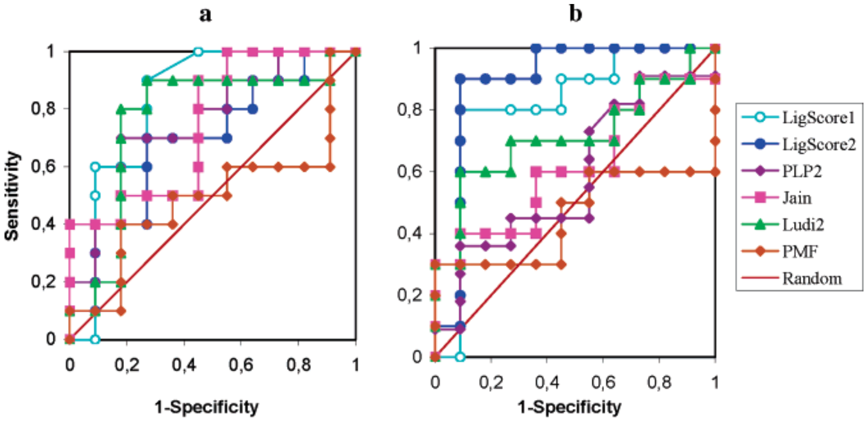
Our goal being to optimize the performance (as assessed by the ROC curve method) of our computer test, various combinations of parameters are envisaged (refer to Table 3).

ROC Curve Method. Application with Different Virtual Screening Parameters. To fine-tune (1) the number of receptor conformations, (2) the type of docking function, (3) the number of in situ minimization steps after docking, (4) the distance tolerance ($D_{\max}$) for the structural filter and (5) to choose the appropriate scoring function, the ROC curves method was applied for various combinations of screening parameters (see Table 3). The objective at this stage is to maximize the AUC of the computer test.

A systematic approach for evaluating each combination being unreasonable from a time perspective (560 combinations in total according to Table 3), we evaluated each parameter in turn, keeping the others constant, and operated in an "evolutionary" way. That is, if one parameter modification improves the AUC, it is kept for the following parameter to be assessed. Although this method cannot guarantee that the best combination of parameters is obtained, it allows a satisfactory combination to be reached for our screening purpose.

First of all, the "ensemble docking" approach proved to be appropriate for our case because none of the four LBM conformations were capable, on its own, of recognizing all known agonists (data not shown) and therefore of impairing the plot of ROC curves. For instance, LigandFit cannot fit (*S*)-3,4-DCPG **10** in C_{ACPT} and ACPT-I **6** cannot be docked in C_{PPG}. Simple geometric constraints can explain these observations: the site model of C_{ACPT} is too short for phenylglycines, and the site model of C_{PPG} is not bulky enough to accept cyclopentyl derivatives. We therefore kept the docking results for all four conformations. Such an improvement with the ensemble docking approach has been reported in other papers[53,54] and is consistent with the current paradigm that describes proteins in a preexisting ensemble of conformational states to which ligands can bind with different affinities.[55] This is line, as well, with a recent hypothesis regarding the activation process of mGluRs where the agonist displaces the dynamic equilibrium toward the closed-activated form[42,44] and, during that process, maximizes its interactions with the protein. This is the reason why we consider only closed forms of the LBD for our vHTS experiments.

Figure 5 reports two sets of ROC curves that allowed us to tune the subsequent parameters. The first observation to be made from this figure is that in most cases the ROC curves remain above the diagonal representing a random distribution. Although it is known that scoring functions are not always capable of identifying the best agonist for a given target, this result tends to support the fact that they are at least capable of discriminating active agonists from inactive compounds; this is actually what is needed for vHTS purposes. If we compare the case where 250 iterations of in situ minimization is added to the workflow (Figure 5), we may notice that the AUC results do not vary significantly except those for LigScore2, which, as a force-field-



Scoring function	AUC
LigScore1	0.82
LigScore2	0.69
PLP2	0.75
Jain	0.75
Ludi2	0.76
PMF	0.49

Scoring function	AUC
LigScore1	0.82
LigScore2	0.89
PLP2	0.59
Jain	0.62
Ludi2	0.72
PMF	0.45

Figure 5. ROC curves obtained with the ensemble docking approach using the PLP1-based docking function and plotted for various scoring functions: (a) results without in situ minimization or ammonium displacement restriction (above) and the calculated AUCs (table below); (b) same as for panel a but with 250 iterations of in situ minimization.

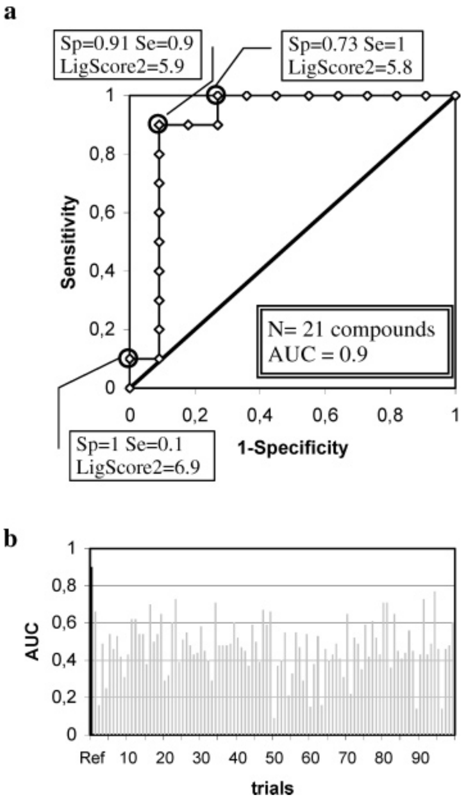


Figure 6. (a) ROC graph provided by LigScore2 (open diamonds) with the optimized set of parameters reported in the text. The random distribution is displayed as a bold diagonal. (b) Bar chart reporting the AUC values obtained for the 99 randomization trials (shown in gray). The reference AUC corresponding to the correct assignment of activities is displayed in black (AUC = 0.9).

based scoring function, is clearly favored by this force-field-driven step.

The combination of parameters that produced the largest AUC with our method is the following one: (1) PLP1 was used as a docking function, including the ligand's internal energies. (2) 250 steps of in situ minimization were applied to the 20 best poses. (3) A tolerance of $D_{\max} = 0.5$ Å between the poses' ammonium and the reference was applied. (4) LigScore2 was used as a scoring function. Regarding the docking function, PLP1, somewhat surprisingly, provided better AUC results than the DREIDING based DockScore. However, this may be explained by the fact that the first, being less restrictive, can authorize some imperfections that are then corrected by the in situ minimization protocol. A total of 250 and 1000 iterations of in situ minimization yielded strictly identical ROC curves, suggesting that 250 iterations were enough to maximize the detection of "activity signal". A tolerance of 0.5 Å for the ammonium position slightly improved the AUCs but in any case allows a dramatic reduction of the number of poses to be scored without losing the important information. Below 0.5 Å, the filter was too restrictive because all poses of some known actives were discarded. Interestingly, this value may be related to the positional precision of about 0.45 Å (according to Cruickshank's formula reported in ref 56) that was reached with the template (resolved at 2.2 Å, $R_{\text{free}} = 0.227$ [42]) we used to build the homology model.

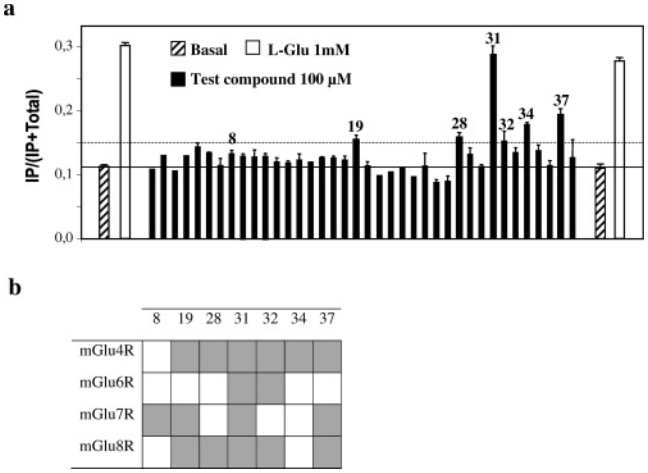
The best ROC curve over the tested combinations is displayed in Figure 6a. This curve evolves parallel to the ideal graph along the left side and even overlaps it for low specificity values. The lowest value of LigScore2 is 4.91 (for the inactive compound (*S*)-CBPG **15**, $EC_{50} > 1$ mM), and the highest is 6.95 for (*S*)-3,4-DCPG **10** ($EC_{50} = 8.8$ μM). According to the correlation between affinity and scores calculated for LigScore2, this range corresponds to about 2 orders of magnitude in terms of pK_i. Its area under the curve reaches 0.9. This means that the designed test is actually capable of giving a higher score to a randomly selected active compound than to a randomly selected inactive in 9 trials out of

10. Such a high value is often considered as "excellent". However, as stated above in the theoretical section, further validation is needed.

Statistical Validation of the Selected Set. To check if the ROC curve obtained with the optimized set of parameters was due to chance or, on the contrary, was reflecting a strong relationship between the structures (here described as a score after docking into a protein model, LigScore2) and the pharmacological activity, a statistical validation was performed. The 21 activities of the sample were scrambled 99 times; i.e., they were randomly redistributed over the 21 compounds in 99 trials. The 99 ROC curves of the trials were then plotted to calculate the 99 corresponding AUCs, which are reported in Figure 6b. Since all trials have a lower AUC than the reference curve (0.9), Fisher's test allows us to state that the reported vHTS method reflects the SAR of the compounds with a statistical significance of 99%.

Choice of a Selection Threshold. As our objective was to discover new chemotypes as mGlu4R agonists, a rather liberal approach privileging sensitivity over specificity was chosen (corresponding to points in the upper-right part of the validated graph). Besides its high AUC, the shape of the obtained ROC curve is a clear advantage when such a strategy is to be adopted because it overlaps the ideal graph on a large section of the upper-left corner, therefore allowing good specificity values to be reached at the maximum sensitivity. In fact, for a maximum sensitivity of 1, specificity can be maximized at 0.73, corresponding to a LigScore2 threshold of 5.8 (refer to Figure 6a). In other words, with a LigScore2 threshold set to 5.8, 100% of known actives are selected (Se = 1) and 73% of known inactives are discarded. Misclassified (overestimated) compounds are (*S*)-4C3HPG **21**, (*S*)-4CH$_2$Glu **11**, and (2*S*,4*S*)-4MGlu **12**, which highlight some imperfections of the computer test in discarding some inactive molecules. Once the models of (*S*)-4CH$_2$Glu **11** and (2*S*,4*S*)-4MGlu **12** are docked, it was observed that their glutamic substructure overlaps almost perfectly with the L-glutamate **1** itself, orienting their extra carbon in an empty space (data not shown). Therefore, their interaction pattern is analogous to the interaction pattern of endogenous agonist. A similar observation can be made when comparing (*S*)-4C3HPG **21** to the known agonist (*S*)-3,4-DCPG **10**.

Database Screening. Virtual Screening. The computer test described above was then applied to databases of compounds of unknown activities available from different providers. Prior to the docking calculations, two filters were applied to enrich the data set of molecules before docking. A first filter was applied to discard compounds with a molecular weight above 500 g/mol in order to reflect the rather small size of known ligands and, consequently, the rather small size of the four site models. A second filter was set to satisfy the interaction pattern evoked above. All known agonists exhibit a primary ammonium moiety that, according to both X-ray studies (for subtype 1) and MD studies (for other subtypes), strongly interacts with highly conserved residues (as described above). Since there is hardly any chemical group bioisosteric to $-NH_3^+$ (i.e., three polar hydrogens and a positive charge worn by a single atom) capable of satisfying the conserved interactions, only molecules exhibiting an ionizable primary ammonium were retained. This second filter dramatically reduced the number of compounds to be docked (see above). This simply characterizes the lack of primary amines in the studies' databases and may be explained by the fact that primary amines are often (wrongly) considered as chemical reagents instead of valuable HTS compounds. Hence, from originally ca. 720 000 molecules, only 1069 molecules were retained for docking. Setting the docking-scoring parameters at their optimized value as described above, the 1069 molecules were docked in the ensemble of receptor conformations. Compounds that exhibited a LigScore2 above 5.8 were visually analyzed in the context of the LBM. Those that formed several H-bonds (apart from their ammonium moiety) or interacted with hydrophobic regions of the binding pocket were selected for purchase. We also analyzed the top 10% of molecules that were scored below 5.8 in the same way and short-listed compounds with novel ("non-glutamate-like") structures. Among the 41 selected compounds, 38 were available from the vendors' stock and purchased for in vitro assays.



	8	19	28	31	32	34	37
mGlu4R		■	■	■	■	■	■
mGlu6R				■	■		
mGlu7R	■	■		■			■
mGlu8R		■	■	■	■		■

Figure 7. Pharmacology of the 38 selected compounds at 100 μM: (a) agonist activity against mGlu4R; (b) selectivity between subtypes of group III mGluRs. Compounds with agonist activity are marked with a gray box.

Functional Screening. Agonist activity of the 38 selected compounds was tested on HEK 293 cells expressing the mGlu4 receptor together with the chimeric Gqi9 protein. The presence of this G-protein allows this Gi-coupled receptor to activate phospholipase C and therefore to stimulate inositol phosphate formation, rather than inhibiting cAMP formation. Compared to basal activity, six compounds significantly increased IP formation when tested at 100 μM (refer to Figure 7a). This corresponds to a hit rate on subtype 4 of ca. 16% for this HTS campaign, improving to several orders of magnitude hit rates usually observed by random screening (less than 0.01% or so according to ref 57). For example, Doman et al. have reported an HTS campaign with a hit rate of 0.02% while investigating protein tyrosine phosphatase-1B, compared to 35% achieved by a prior virtual screening-based selection.[58] These numbers, however, are by no mean indicative of the very quality of a given virtual screening campaign because hit rates depend, for a start, on the screened

Table 4. Statistical Analysis of Calculated log *P* Values (AlogP98)

	total number	min	max	av	standard deviation
sample actives	10	−1.7	0	−1.06	0.52
sample inactives	11	−2	0	−0.74	0.60
total sample	21	−2	0	−0.89	0.57
database	1069	−7	7.6	1.16	1.53
selection	38	−2.8	4	0.36	1.31
hits	6	−1.7	4	−0.95	1.84

database itself. Indeed, current drug design paradigm hypothesizes that, upon lead optimization, for instance, chemical derivatives of a given lead structure would constitute a library with a higher probability of activity than commercial databases with diverse compound structures. Hit rates may even depend on the biological target being investigated. For instance, Luu et al. have recently reported[59] that the gold fish odorant receptor OR5.24, which happens to be in the same GPCR family as mGluRs, could be accurately modeled by comparative modeling with the LBD of mGlu1R. Despite these similarities, OR5.24 is activated by the 20 proteinogenic amino acids at a 100 μM concentration[60] whereas mGluRs specifically recognize L-glutamate only (data not shown). This simple observation shows that the same database screened on homologous receptors can yield radically different hit rates.

When tested on other group III subtypes (i.e., mGlu(6−8)R), a rather poor selectivity was observed among the six hits (Figure 7b). The most potent compound at 100 μM (compound **31**, Figure 7a) is capable of activating all group III subtypes. Among the five other positives, only one (compound **34**) was selective for subtype 4. More surprising, when tested on other group III subtypes, one negative compound on mGlu4R was shown to activate mGlu7R. These observations regarding selectivity suggest those even virtual screening workflows relying on a docking approach are not specific enough to induce receptor selectivity in the selection process. Having said this, 6 out of 7 hits have an agonist activity on subtype 4, which is acceptable.

A Posteriori Analysis Regarding the Calculated log *P* Values. After the pioneering work of Lipinski et al. suggesting the use of computational filter to improve solubility and permeability of screened compounds,[61] it is now common to include criteria such as molecular weight (MW) or hydrophobicity (quantified by the *n*-octanol/water partition coefficient, log *P*) in virtual screening workflows in order to favor druglike compounds during the selection process (see refs 58 and 62 as examples). Here, only molecules with a MW below 500 g/mol were retained but not so much to abide by Lipinski's rule than to discard large molecules that obviously cannot fit in the small binding site of our model. To evaluate the importance of filtering according to the predicted log *P* (AlogP98) in the case of mGlu4R, a straightforward a posteriori study was carried out (Table 4). First of all, it appears that when considering the sample used for the ROC curve analysis, AlogP98 does not discriminate the active population from the inactives. This implies that the ROC curve profiles would not be modified if a supplementary filter focused on log *P* values was implemented. When comparing the results obtained with the database screened, it is noteworthy that commercially available compounds were rather more hydrophobic than the sample of molecules of known activities (1.16 on average compared to −0.89). Since all known agonists exhibit a negative AlogP98, we could have decided to discard any molecule with positive AlogP98. As a consequence, only 211 molecules would have been docked, out of which 11 compounds would have been selected for HTS. Among these 11 compounds, 4 exhibit an agonist activity on mGlu4R corresponding to an improved hit rate of 36%. These results tend to back up the use of log *P* filters because they can yield better hit rates with less effort. However, this is not in the line of the liberal strategy adopted here for mGlu4R agonists. Indeed, two active compounds, that is, one-third of the total hits, would never have been found.

Conclusion

Despite the plethora of scientific fields that rely on the ROC curve approach for important decision-making, the drug discovery area is late in adopting it. This is the case even though making decisions on whether to continue to invest in the evaluation of a molecule or to discard another is considered as a central issue in many steps of drug discovery and development.[63] Even the rare papers that report ROC curves in this field underutilize them and base their conclusions on other methods such as enrichment curves. And yet, the ROC curve method features several key advantages compared to other existing methods.[7] First, this graphical method provides a comprehensive representation of the pure accuracy of a given vHTS workflow ("computer test"). Indeed, only the entire spectrum of sensitivity/specificity pairs provides a complete picture of test accuracy reporting the dual aspect of any test, namely, the ability to select active compounds and discard inactive ones. Second, the ROC curve method is strictly independent of the rate of active molecules in the sample set, allowing inclusion of any information related to both activity and inactivity. Third, ROC curves allow a direct visual comparison between tests on a common graph.[7] Their relative simplicity facilitates transdisciplinary communication among research groups and results in publications that permit comparison to others (provided that the used samples represent the same SAR data). Several vHTS approaches can even be compared on the same graph, facilitating the choice of method that suits the objectives the best (e.g., scaffold hoping or lead optimization). The absence of theoretical background to the ROC curve method applied to drug discovery may explain the observed reluctance in using it. Filling this lack was the first objective of this paper. The second objective of this paper was to introduce a real prospective application of the ROC curve approach. The chosen system (agonists of mGlu4 receptor) was rather tricky owing to the poor pharmacological data availability. However, despite the small size of the derived sample for mGlu4R, the ROC method still managed to extract a clear activity signal (area under the curve of 0.9) and subsequently to find 6 agonists among 38 selected compounds out a database of more than 720 000 molecules.

Experimental Section

Computer Test Elaboration. To use the ROC curve approach, the computer test was developed by following the guidelines described in the theoretical section.

1. Choice of Activity Cutoff. The activity cutoff was set to 100 μM, corresponding to the concentration at which candidates are tested in our HTS assay.

2. Sample Selection and Building. Twenty-one molecules representing the pharmacological profile of mGlu4R were selected and distributed into two classes: "actives" having an EC_{50} below 100 μM and "inactives" above (see Figure 3 for structures and activity classes). All molecules have a common substructure: a primary ammonium.

A comparative model of mGlu4R LBM was refined by molecular dynamics in the presence of four structurally diverse mGlu4 receptor agonists (L-AP4 **2**, ACPT-I **6**, L-Glu **1**, and (*S*)-PPG **9**) according to the method described by Bertrand et al.[48] Each molecule in the sample was manually built with the sketcher of the InsightII software package (version 2000.1, Accelrys Inc., San Diego, CA) and protonated at physiological pH.

3. High-Throughput Docking-Scoring. The resulting molecules were then processed into a high-throughput docking-scoring workflow. Each of them was docked with LigandFit (version 4.9, Accelrys Inc.) in each of the four models using either PLP1 or DREIDING to derive the docking function (DockScore). Twenty poses for each complex were saved and minimized (steepest descent) in the context of the protein (held rigid) with the DREIDING force field.[51] Results were then processed in order to remove all the poses orienting the ligands' ammonium fragment further than a certain distance ($D_{\max}$) away from the ammonium moiety of the reference molecule. The remaining poses were finally scored with the entire panel of scoring functions available within the Cerius^2 software package (version 4.9, Accelrys Inc.): LigScore1,[50] LigScore2 (both using the DREIDING force field[51]), PLP1,[52] PLP2,[64] Jain,[65] Ludi2,[66] and PMF.[67] Poses from the four conformations of the receptor were pooled, and for a given scoring function, only the highest scoring pose was retained to plot the corresponding ROC curve.

4. ROC Curves Plotting. For a given combination of parameters, the ROC curves corresponding to each of the above scoring functions were plotted. To do so, for a given scoring function, the 21 score values of the sample were used as selection thresholds. By this approach, the number of calculations and the number of approximations due to binning effects are reduced to their minima. For each threshold, selected and discarded compounds are counted for both actives and inactives (see the confusion matrix on Figure 1b). The pairs of sensitivity and specificity deduced from the previous count were then used to plot the ROC curves. Finally, the AUCs are calculated as the sum of the areas of the rectangles below the ROC curve: $AUC = \sum_i[(Se_{i+1})(Sp_{i+1} - Sp_i)/2]$.

Database Virtual Screening. Databases Preparation and Filtering. Commercial databases were prepared as follows. First, molecules were converted into 3D coordinates with CORINA (available in TSAR, version 3.3, Accelrys Inc.) using the option to strip away counterions and solvent molecules. The obtained SD file was processed by the set_charge executable available from Accelrys Inc. to set formal charges according to physiological pH for most known ionizable groups (aliphatic amines, amidines, guanidines, carboxylic acids, etc.). The two filters (MW < 500 g/mol and primary ammonium substructure) were implemented by means of Catalyst queries (Catalyst, version 4.9, Accelrys Inc.). The remaining compounds were exported in SD file format and concatenated with the 21 compounds of known activities (internal references) prior to docking.

High-Throughput Docking, Postprocessing, and Scoring. Remaining molecules followed the vHTS workflow designed with the ROC curve method. Hence, they were docked in the four conformations of the receptor's LMB with LigandFit using PLP1 as a docking function. After docking, poses were refined by in situ minimization (250 steps). The resulting poses were filtered according to the displacement of their ammonium (maximum 0.5 Å) and scored with LigScore2 (using the DREIDING force field). Only the best pose obtained on the ensemble of four conformations was retained for classification.

Compound Selection. All compounds exhibiting a LigScore2 above 5.8 were visually inspected as well as the top 10% of compounds displaying a lower LigScore2. Novel structures were selected according to their interaction pattern with the LBM model and purchased when available.

Functional Screening. Agonist activity of the 38 selected compounds was tested on HEK293 cells transiently transfected with the rat mGlu4 expressing plasmid pRKG4 and the chimeric G-protein Gqi9 by electroporation, as previously described.[68] Cells were plated in 96-well culture plates and labeled overnight with [^{3}H]myoinositol. The day after, cells were washed three times with Krebs buffer, incubated for 10 min with LiCl 5 mM, and then incubated for 30 min in the absence (basal) or in the presence of the indicated compounds at 100 μM. The total amount of [^{3}H]phosphatidylinositol accumulated in the cells was determined after Dowex purification as previously described.[69] All experiments were carried out seven times. Compounds were considered as hits if the IP formation was significantly higher than the one for basal activity in at least 3 experiments out of 7.

The log *P* Calculations. Estimated *n*-octanol/water partition coefficient (log *P*) were calculated within Cerius^2 (version 4.9, Accelrys Inc.) according to Ghose and Crippen's atomic approach[70] (AlogP98).

Acknowledgment. This work was supported by grants from the CNRS (J.-P.P., F.A.), the Action Concertée Incitative (Molecules et Cibles Thérapeutiques of the French Ministry of Research and Technology) (J.-P.P., F.A.), the Comité Parkinson of the Fondation de France (F.A., J.-P.P.), the Fondation Paul Hamel (J.-P.P., F.A.), and the European Community (Grant LSHB-CT-200-503337 to J.-P.P.). N.T. is the recipient of a CIFRE Grant (No. 738/2002) provided by the Association Nationale pour la Recherche Technique and Accelrys. We thank R. Hoffmann and S. O'Brien for critical reading of the manuscript and Anne Goupil-Lamy and Sonja Meddeb for their help throughout the course of this work.

Supporting Information Available: Figure representing the binding site of L-AP4 **2** in mGlu4R. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Neyman, J.; Pearson, E. S. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc, London, Ser. A* **1933**, *231*, 289−337.

(2) Neyman, J.; Pearson, E. S. The testing of statistical hypotheses in relation to probabilities a priori. *Proc. Cambridge Philos. Soc.* **1933**, *20*, 492−510.

(3) Green, D. M.; Swets, J. A. *Signal Detection Theory and Psychophysics*; John Wiley & Sons: New York, 1966.

(4) Swets, J. A. The relative operating characteristic in psychology. *Science* **1973**, *182*, 990−1000.

(5) Lusted, L. B. Decision making studies in patient management. *N. Engl. J. Med.* **1971**, *284*, 416−424.

(6) Lusted, L. B. Signal detectability and medical decision-making. *Science* **1971**, *171*, 1217−1219.

(7) Zweig, M. H.; Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **1993**, *4*, 561−577.

(8) Torre, P.; Cruickshanks, K. J.; Nondahl, D. M.; Wiley, T. L. Distortion product otoacoustic emission response characteristics in older adults. *Ear Hear.* **2003**, *24*, 20−29.

(9) Kharin, V. V.; Zwiers, F. W. On the ROC score of probability forecasts. *J. Climate* **2003**, *16*, 4145−4150.

(10) Barbaree, H. E.; Seto, M. C.; Langton, C. M.; Peacock, E. J. Evaluating the predictive accuracy of six risk assessment instruments for adult sex offenders. *Crim. Justice Behav.* **2001**, *28*, 490−521.

(11) Swets, J. A. Measuring the accuracy of diagnostic systems. *Science* **1988**, *240*, 1285−1293.

(12) Swets, J. A.; Dawes, R. M.; Monahan, J. Better decisions through science. *Sci. Am.* **2000**, *283*, 82–87.

(13) Pin, J.-P.; Acher, F. The metabotropic glutamate receptors: structure, activation mechanism and pharmacology. *Curr. Drug Targets* **2002**, *1*, 297–317.

(14) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening—an overview. *Drug Discovery Today* **1998**, *3*, 160–178.

(15) Böhm, H.-J.; Schneider, G. *Virtual Screening for Bioactive Molecules*; Wiley-VCH: Weinheim, Germany, 2000.

(16) Bajorath, J. Integration of virtual screening and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.

(17) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903–911.

(18) Willett, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(19) Langer, T.; Hoffmann, R. D. Virtual screening: An effective tool for lead structure discovery? *Curr. Pharm. Des.* **2001**, *7*, 509–527.

(20) Debnath, A. K. Quantitative structure–activity relationship (QSAR) paradigm—Hansch era to the new millennium. *Mini-Rev. Med. Chem.* **2001**, *1*, 187–195.

(21) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.

(22) Good, A. C.; Krystek, S. R.; Mason, J. S. High-throughput and virtual screening: core lead discovery technologies move towards integration. *Drug Discovery Today* **2000**, *5*, S61–S69.

(23) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; et al. Virtual screening using protein–ligand docking: Avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–806.

(24) Manallack, D. T.; Pitt, W. R.; Gancia, E.; Montana, J. G.; Livingstone, D. J.; et al. Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1256–1262.

(25) Manallack, D. T.; Tehan, B. G.; Gancia, E.; Hudson, B. D.; Ford, M. G.; et al. A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 674–679.

(26) Güner, O. F.; Henry, D. R. Metric for analyzing hit lists and pharmacophores. *Pharmacophore Perception, Development and Use in Drug Design*; International University Line: La Jolla CA, 2000; pp 193–212.

(27) Bradley, E. K.; Miller, J. L.; Saiah, E.; Grootenhuis, P. D. J. Informative library design as an efficient strategy to identify and optimize leads: Application to cyclin-dependent kinase 2 antagonists. *J. Med. Chem.* **2003**, *46*, 4360–4364.

(28) Jacobsson, M.; Lidén, P.; Stjernschantz, E.; Boström, H.; Norinder, U. Improving structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.* **2003**, *46*, 5781–5789.

(29) Shepherd, A. J.; Gorse, D.; Thornton, J. M. Prediction of the location and type of β-turns in proteins using neural networks. *Protein Sci.* **1999**, *8*, 1045–1055.

(30) Diller, D. J.; Li, R. Kinases, homology models, and high throughput docking. *J. Med. Chem.* **2003**, *46*, 4638–4647.

(31) Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.

(32) Hubbard, R.; Bayarri, M. J. Confusion over measures of evidence (p's) versus errors (alpha's) in classical statistical testing. *Am. Stat.* **2003**, *57*, 171–178.

(33) Bräuner-Osborne, H.; Egebjerg, J.; Nielsen, E. Ø.; Madsen, U.; Krogsgaard-Larsen, P. Ligands for Glutamate Receptors: Design and Therapeutic Prospects. *J. Med. Chem.* **2000**, *43*, 2609–2645.

(34) Holden, C. Excited by glutamate. *Science* **2003**, *300*, 1866–1868.

(35) Moldrich, R. X.; Chapman, A. G.; Sarro, G. D.; Meldrum, B. S. Glutamate metabotropic receptors as targets for drug therapy in epilepsy. *Eur. J. Pharmacol.* **2003**, *476*, 3–16.

(36) Bhave, G.; Karim, F.; Carlton, S. M.; Gereau, R. W., IV. Peripheral group I metabotropic glutamate receptors modulate nociception in mice. *Nat. Neurosci.* **2001**, *4*, 417–423.

(37) Neugebauer, V. Metabotropic glutamate receptors—important modulators of nociception and pain behavior. *Pain* **2002**, *98*, 1–8.

(38) Kenny, P. J.; Markou, A. The ups and downs of addiction: role of metabotropic glutamate receptors. *Trends Pharmacol. Sci.* **2004**, *25*, 265–272.

(39) Bergink, V.; Megen, H. J. G. M. v.; Westenberg, H. G. M. Glutamate and anxiety. *Eur. Neuropsychopharmacol.* **2004**, *14*, 175–183.

(40) Marino, M. J.; Valenti, O.; Conn, P. J. Glutamate receptors and Parkinson's disease. *Drugs Aging* **2003**, *20*, 377–397.

(41) Pin, J.-P.; Galvez, T.; Prézeau, L. Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacol. Ther.* **2003**, *98*, 325–354.

(42) Kunishima, N.; Shimada, Y.; Tsuji, Y.; Sato, T.; Yamamoto, M.; et al. Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor. *Nature* **2000**, *407*, 971.

(43) O'Hara, P. J.; Sheppard, P. O.; Thogersen, H.; Venezia, D.; Haldeman, B.; et al. The ligand-binding domain in metabotropic glutamate receptors is related to bacterial periplasmic binding proteins. *Neuron* **1993**, *11*, 41–52.

(44) Bessis, A.-S.; Rondard, P.; Gaven, F.; Brabet, I.; Triballeau, N.; et al. Closure of the Venus Flytrap module of mGlu8 receptor and the activation process: insights from mutations converting antagonists into agonists. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11097–11102.

(45) Cartmell, J.; Schoepp, D. D. Regulation of neurotransmitter release by metabotropic glutamate receptors. *J. Neurochem.* **2000**, *75*, 889–907.

(46) Schoepp, D. D.; Jane, D. E.; Monn, J. A. Pharmacological agents acting at subtypes of metabotropic glutamate receptors. *Neuropharmacology* **1999**, *38*, 1431–1476.

(47) Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discovery* **2003**, *2*, 527–541.

(48) Bertrand, H.-O.; Bessis, A.-S.; Pin, J.-P.; Acher, F. C. Common and selective molecular determinants involved in metabotopic glutamate receptor agonist activity. *J. Med. Chem.* **2002**, *45*, 3171–3183.

(49) Hampson, D. R.; Huang, X.-P.; Pekhletski, R.; Peltekova, V.; Hornby, G.; et al. Probing the ligand-binding domain of the mGluR4 subtype of metabotropic glutamate receptor. *J. Biol. Chem.* **1999**, *274*, 33488–33495.

(50) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289–307.

(51) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. G., III. DREIDING: A generic force field for molecular simulation. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

(52) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; et al. Molecular recognition of inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317–324.

(53) Mangoni, A.; Roccatano, D.; DiNola, A. Docking of flexible ligands to flexible receptors in solutions by molecular dynamics similation. *Proteins* **1999**, *35*, 153–162.

(54) Keserû, G. M.; Kolossvary, I. Fully flexible low-mode docking: Application to induced fit in HIV integrase. *J. Am. Chem. Soc.* **2001**, *123*, 12708–12709.

(55) Carlson, H. A. Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447–452.

(56) Blow, D. *Outline of Crystallography for Biologists*; Oxford University Press: New York, 2002; p 236.

(57) Borman, S. Proteomics: taking off where genomics leaves off. *Chem. Eng. News* **2000**, *78*, 31–37.

(58) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; et al. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213–2221.

(59) Luu, P.; Acher, F.; Bertrand, H.-O.; Fan, J.; Ngai, J. Molecular determinants of ligand selectivity in a vertebrate odorant receptor. *J. Neurosci.*, in press.

(60) Speca, D. J.; Lin, D. M.; Sorensen, P. W.; Isakoff, E. Y.; Ngai, J.; et al. Functional identification of a goldfish odorant receptor. *Neuron* **1999**, *23*, 487–498.

(61) 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.* **1997**, *23*, 3–25.

(62) Grüneberg, S.; Stubbs, M. T.; Klebe, G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation. *J. Med. Chem.* **2002**, *45*, 3588–3602.

(63) Pritchard, J. F.; Jurima-Romet, M.; Reimer, M. L. J.; Mortimer, E.; Rolfe, B.; et al. Making better drugs: Decision gates in nonclinical drug development. *Nat. Rev. Drug Discovery* **2003**, *2*, 542–553.

(64) Gehlhaar, D. K.; Bouzida, D.; Rejto, P. A. Rational Drug Design: Novel Methodology and Practical Applications; *American Chemical Society*: Washington, DC, 1999; pp 292–311.

(65) Jain, A. N. Scoring noncovalent protein–ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427–440.

(66) Böhm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.

(67) Muegge, I.; Martin, Y. C. A general and fast function for protein–ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791–804.

(68) Gomeza, J.; Mary, S.; Brabet, I.; Parmentier, M. L.; Restituito, S.; et al. Coupling of metabotropic glutamate receptors 2 and 4 to G alpha 15, G alpha 16, and chimeric G alpha q/i proteins:

characterization of new antagonists. *Mol. Pharmacol.* **1996***, 50*, 923−930.

(69) Goudet, C.; Gaven, F.; Kniazeff, J.; Vol, C.; Liu, J.; et al. Heptahelical domain of metabotropic glutamate receptor 5 behaves like rhodopsin-like receptors. *Proc. Natl. Acad. Sci. U.S.A.* **2004***, 101*, 378−383.

(70) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **1998***, 102*, 3762−3772.

JM049092J