Measuring CAMD Technique Performance. 2. How "Druglike" Are Drugs? Implications of Random Test Set Selection Exemplified Using Druglikeness Classification Models

Andrew C. Good*,† and Mark A. Hermsmeier‡

Bristol-Myers Squibb, 5 Research Parkway, Wallingford, Connecticut 06492, and Bristol-Myers Squibb, P.O. Box 4000, Princeton, New Jersey 08543

Received August 10, 2006

Research into the advancement of computer-aided molecular design (CAMD) has a tendency to focus on the discipline of algorithm development. Such efforts are often wrought to the detriment of the data set selection and analysis used in said algorithm validation. Here we highlight the potential problems this can cause in the context of druglikeness classification. More rigorous efforts are applied to the selection of decoy (nondruglike) molecules from the ACD. Comparisons are made between model performance using the standard technique of random test set creation with test sets derived from explicit ontological separation by drug class. The dangers of viewing druglike space as sufficiently coherent to permit simple classification are highlighted. In addition the issues inherent in applying unfiltered data and random test set selection to (Q)SAR models utilizing large and supposedly heterogeneous databases are discussed.

INTRODUCTION

With some notable exceptions,[1−5] careful data set selection and analysis has often become a casualty in the development and validation of novel CAMD techniques. This article forms the second in a series of papers highlighting many of the inherent problems with and potential improvements to said validation efforts.[6]

In this study we have focused our attention on the subject of druglikeness. Since Lipinski et al. developed the rule of five as the first widely recognized molecular description of druglikeness,[7] the concept has become the focus of significant effort within the CAMD community. Many different analyses of drug databases have been undertaken, and these have been extensively reviewed elsewhere.[8−11] The rule of five was originally defined as a filter for early discovery to focus on more promising leads.[7] Its popularity meant, however, that the focus soon turned the spotlight on the technique as a description of drug (ADME) space.[12] Druglikeness classification techniques form a natural extrapolation of this concept and have received significant attention in the field.[13−17] Machine learning algorithms are typically applied to separate compounds from druglike databases (most commonly the WDI[18]and MDDR[19]) from molecules in reagent databases (generally the ACD[19]). A wide variety of machine learning methods (e.g., neural nets,[13,14] recursive partitioning,[15] and support vector machines[16,17]) and properties (e.g., MDL substructure keys,[14] Ghose descriptors,[13,17] simple 1d properties such as molecular weight, logP, rotatable bond count,[14] functional group counts, etc.[14,15]) have been applied, with seemingly impressive results achieved in test set prediction (as low as ∼10% misclassification[17]).

An implicit concept underlying druglikeness classification is that compounds with favorable ADME properties contain key structural similarities that define a druglike property space. Analysis of the underlying data used in model construction, however, highlights a number of pitfalls in model validation that render the results at best difficult to interpret. It has been shown that rigorous tests of QSAR/QSPR models are critical to test model prediction.[20] Of particular importance is the selection of as large a test set as possible. On initial inspection drug classification models seem blessed in this regard, since the size of the databases used permit the construction of tests sets which are larger than the training sets used in model derivation. A closer look, however, highlights that it is not just size but also content of the test set that is critical to allow proper testing. For classification models, random training sets are culled from the databases to construct a classification model, and these are validated using test sets also selected at random. Consideration of the nature of data present in the constituent databases illustrates the potential dangers of this approach. Pharmaceutical companies tend to run fast follow on projects against their own and their competitors most promising compounds. As a consequence druglike databases are heavily populated with closely related analogues. This in turn means that many closely related compounds end up in both the training and test sets when random selections are culled from said databases. When models are constructed and tested using such a strategy, there is thus a large chance that the test set result will overpredict model utility. The end result is that the model may well be less a reflection of druglike space and more a look up table of drug substructures. The issue of analogue bias pervades many of the data sets used in model construction, from scoring function design[21] to virtual screening enrichment studies.[6]

Another potential problem comes in the form of the decoy database used to define nondruglike compounds, most commonly the ACD database. The history of the ACD as a collection of reagents for compound elaboration leads to the database containing a systematic difference in compound size

* Corresponding author e-mail: andrew.good@bms.com.
† Bristol-Myers Squibb, Wallingford, CT.
‡ Bristol-Myers Squibb, Princeton, NJ.

10.1021/ci6003493 CCC: $37.00
Published on Web 12/06/2006

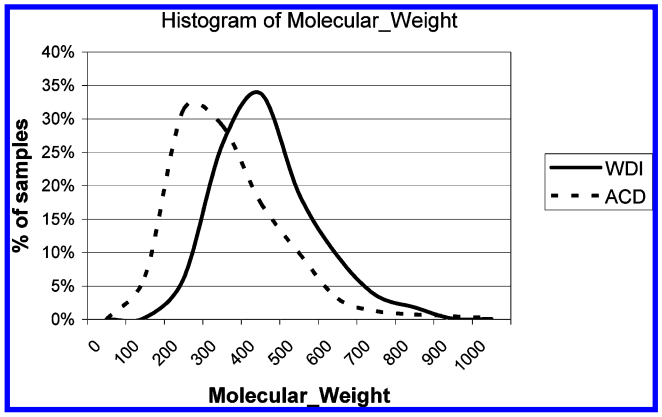

Figure 1. Molecular weight distribution of ACD compared with the WDI compound subset spanning ontology space—see the Methods section. Systematic bias toward lower molecular weight in the ACD compounds based on its heritage as a reagents database is clear, highlighting the danger of using MW as a descriptor when ACD data are not prefiltered to remove the systematic bias.

relative to the compounds in the WDI (see Figure 1). Such a difference could potentially lead to systematic differences in other properties, again increasing the chances that the classification models will end up performing as look up tables for differentiating databases based on their systematic differences (e.g. this molecule is smaller and hence more hydrophilic, therefore it probably belongs to the ACD. While likely true, this is not particularly informative in the context of druglikeness.). Issues inherent in the use of the ACD in this context have been highlighted[13,22] but largely disregarded.

In the studies detailed below a number of alterations have been made to data set construction in order to mitigate the issues described above, thus providing a more exacting test of classification models and the druglike space they purport to map.

METHODS

WDI Database Analysis. To diminish the problem of analogue look up between training and test sets, the WDI data used in model construction have been divided up based on relevant drug ontology classes as defined by Schuffenhaur et al.[23] The WDI database (version WDI2004.2) includes one or more mode(s) of action (MOA) for each compound. Using exact and approximate keyword searching as many as possible of the 953 unique MOAs have been linked to the targets defined in the Schuffenhaur ontology. In this way 25 059 WDI compounds have been divided into 17 target classes. Subsequent removal of redundant and near similar structures produced the resulting database division shown in Table 1.

ACD Database Subset Selection. As mentioned in the Introduction it is important to understand the nature of the data being utilized in model construction to ensure optimal training and test set selection. An excellent example of this comes in the context of leadlike versus druglike compound selection based on the analysis of drug discovery project compound histories.[24] In a similar vein, to lesson the risk of systematic bias in our decoys due to the history of ACD compound acquisition, the molecular weight distribution in our chosen WDI compound data set was used as a constraint in ACD compound subset selection. There were 200 MW bins (bin size is 10 au) with 1−303 molecules per bin based on the WDI distributions. Random compounds were chosen from the ACD (version ACD2004.2) to fill up the slots for each bin until all the bins reached the allotted number, producing a noise data set of equivalent size and distribution. The resultant effect on compound ALogP is shown in Figure 2, highlighting the potential knock on effects such a bias can introduce.

Table 1. WDI Data Division into 17 Drug Ontology Classes[a]

drug ontology class	no. of
amine binding class A GPCR	1810
peptide binding class A GPCR	973
prostanoid	242
nucleotide-like	111
class B secretin-like	24
class C metabotropic glutamate pheromone	23
nuclear receptor	529
thyroid hormone-like	193
estrogen-like	336
glutamate cationic receptor	285
nicotinicoid	362
oxidoreductase	978
transferase	1006
hydrolase	1731
lyase	103
isomerase	280
ligase	4

[a] Compound counts associated with each class are shown.

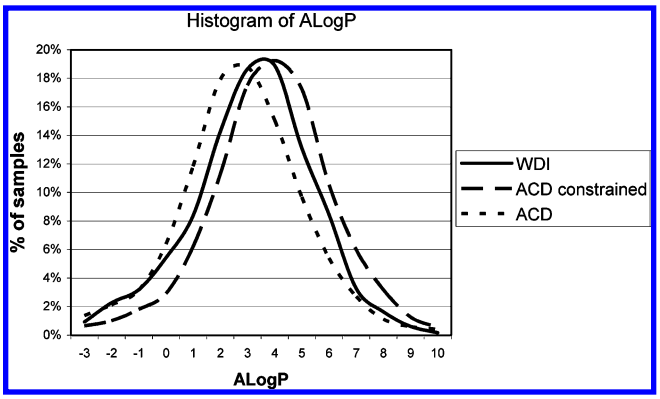

Figure 2. Effects on ALogP of the ACD molecular weight constraint. It is clear that the selected ACD molecules are significantly more lipophilic than the molecules contained in the full ACD set, as are the compounds in the WDI selection.

Druglike Classification Model Construction. The Scitegic Bayesian classifier was applied as the machine learning tool for these studies. The descriptors used to construct the Bayesian model include the following: ECFP_4 (Extended Class Finger Prints - descriptors that use the Morgan Algorithm to iteratively compute the atomic environment within N topological bond lengths from each atom[25]), ALogP, molecular weight, number of rotatable bonds, number of H-bond donors, number of H-bond acceptors, and the 166 MDL Public Keys.

Two descriptor sets were utilized to determine model stability at different levels of descriptor resolution. (a) For the first study, only MDL public key fingerprints were used to derive the model.[26] (b) For the second study the kitchen sink approach was applied, with ECFP_4 fingerprints, ALogP, molecular weight, donor, acceptor and rotatable bond count, and MDL Keys all used in model derivation. Molecular weight was kept as a descriptor since, while systematic

Table 2. (a) Drug Classification Results Obtained for MDL Public Key Fingerprints and (b) for ECFP Fingerprints, Property Counts, and MDL Key Descriptors Combined.

drug ontology class omitted	% random ACD test set classified druglike	% WDI random test set classified druglike	% omitted class classified druglike
	(a)		
amine binding class A GPCR	27	71	78
peptide binding class A GPCR	27	72	64
prostanoid	28	74	75
nucleotide-like	27	72	57
class B secretin like	27	72	50
class C metabotropic glutamate pheromone	27	73	96
nuclear receptor	29	76	89
thyroid hormone-like	28	75	84
estrogen-like	28	75	93
glutamate cationic receptor	26	74	53
nicotinicoid	26	73	51
oxidoreductase	27	76	60
transferase	25	74	64
hydrolase	27	72	53
lyase	26	73	64
isomerase	25	71	68
ligase	26	73	75
mean	**27**	**73**	**69**
standard deviation	**1**	**2**	**15**
	(b)		
amine binding class A GPCR	9	79	57
peptide binding class A GPCR	10	80	49
prostanoid	9	82	37
nucleotide-like	10	81	50
class B secretin-like	9	81	33
class C metabotropic glutamate pheromone	9	81	87
nuclear receptor	9	81	90
thyroid hormone-like	10	81	82
estrogen-like	9	81	91
glutamate cationic receptor	9	82	58
nicotinocoid	9	81	47
oxidoreductase	8	83	48
transferase	9	82	57
hydrolase	10	83	39
lyase	9	82	67
isomerase	9	81	70
ligase	9	81	100
mean	**9**	**81**	**62**
standard deviation	**1**	**1**	**21**

bias has been removed, it could still differentiate between molecules in the context of a multivariate model.

Model Construction and Validation. For each property set, models were created systematically removing each ontology class from the data set. In each case the remaining WDI and ACD data sets were randomly split in two and combined, one-half becoming the training set and the other the test set. The model was constructed using the training set and then tested using both the "standard" randomly selected test set and the test set represented by the omitted ontology class. In this way the performance of the model in the context of standard model validation could be compared with that obtained using a more rigorous test set taken from the omitted drug class.

RESULTS

The results for the studies described above are shown in Table 2(a,b).

DISCUSSION

Models constructed using MDL substructure keys (Table 2a) show reasonable separation in druglikeness predictions for the two databases. For the random test sets 23% of the ACD (std. dev. 1%) are predicted to be druglike versus 73% for the WDI (std. dev. 2%). The ontology specific test sets show less stability, with an average prediction of 69% of compounds being druglike and a standard deviation of 15%. This higher standard deviation illustrates the difficulty in predicting druglikeness when the model is forced to jump between ontology classes. The variability in predictive ability is highlighted by the fact that certain drug classes show druglike prediction barely above random (hydrolase at 53% and nicotinicoid at 51%).

Classification through the application of all selected descriptors produces seemingly superior results, and high stability across the models was observed when measured against random test set data. Only 9% of the ACD is predicted druglike (std. dev. 1%) versus 81% for the WDI (std. dev. again 1%). The results of the ontology class tests sets suggest that the models are significantly overfitted, however, as the mean druglikeness prediction drops markedly to 62%, below that seen for the model using MDL keys alone. Standard deviations are also higher at 21%. Many of the ontology classes are now poorly predicted, including hydrolase (39%) and nicotinicoid (47%) once again.

While the predictions of the classifiers shown here are comparable to some of the models produced within the literature, they are by no means the best. The relative performance of the MDL keys only versus the all available descriptors model clearly highlights, however, that increased predictive performance in a random test set sense does not necessarily translate to true model utility. In Table 2a,b, the classic random test set behavior is very stable, irrespective of the drug class omitted (1−2% standard deviation in performance across all models). The much higher standard deviations (15−21%) of omitted class test set performance clearly highlights the overly optimistic view of performance presented by such random testing. These high standard deviations also illustrate the difficulty of predicting some drug ontologies using a single classification model derived from the remaining classes.

Lipinski et al.[7] found that only antibiotics, antifungals, vitamins, and cardiac glycosides fell outside their rule set (the reason for their omission being ascribed to transporter effects). The results from these studies suggest, however, that the discontinuities in druglike space are significantly greater at least in the context of classification. Lu et al. showed how the correlation of general properties with bioavailability is highly dependent on the therapeutic classes included in the training sets.[27] Vieth and Sutherland recently highlighted the limitations of the rule of 5 using an analysis of drugs broken down by target type.[28] In particular they suggest that molecular weight is more a function of target type than any intrinsic druglike nature. The review of Muegge[10] eloquently highlights that the approximate nature of druglikeness models often restricts their application to library design, where limited predictive capability may still provide value. Even here caution should be exercised, however, since libraries created for lead discovery should

by necessity be leadlike rather than druglike,[29] further restricting their utility.

The results presented here further highlight the issues of model applicability in the context of druglikeness. Constraining ACD data to the molecular weight distribution of our WDI compounds illustrates how consideration of the nature and choice of data set can profoundly effect compound selection. It is important to realize that the scope of such an analysis spreads well beyond these studies, however. Hann et al. illustrated the differences between leadlike and druglike compounds, highlighting a major issue inherent in compound selection for screening enrichment.[24,29] Nissink et al.[1] and Smith et al.[21] highlighted a number of problems intrinsic to many of the complexes and data sets extracted from the PDB[30] for scoring function design. The impact of data set selection on study conclusions can also clearly be seen in the context of scoring function design. Experiments reported by Bissantz et al. found that docking consensus scoring performances varied widely among targets.[31] In contrast Stahl and Rarey suggested that the combinations of FlexX and PLP scores worked best.[32] Clearly a careful analysis of the data being used with due consideration being given to data quality and history as well as volume has significant value.

Of potentially even greater importance is the issue of random test set selection. The results presented here highlight the dangers of such a procedure in the context of data containing large numbers of closely related analogues. As we have already pointed out, the issue of analogue bias runs through many of the data sets used in CAMD studies. In essence any study involving data sets culled from databases constructed using drug discovery project compounds are prone to this problem. Smith et al. highlighted how such bias is endemic to many data set selections extracted from the PDB for scoring function design.[21] Good et al. illustrated how analogue bias can have a major effect on virtual screening enrichment studies, both in the context of study design and conclusions.[6] Another example for the reader to consider is the blood brain barrier (BBB) permeability model of Narayanan and Gunturi.[33] The history of drugs with CNS activity is dominated by compounds that target biogenic amine binding GPCRs. As their name suggests, compounds that bind to these targets invariably contain an amine as part of the binding pharmacophore. One of the primary descriptors of the preferred BBB models in this study is the presence of a triply substituted nitrogen. It can be argued that this descriptor is more a reflection of the history of CNS drug targeting by the pharmaceutical industry, rather than any intrinsic preference of the BBB for basic moieties. Those wishing to test this possibility might consider designing a test set devoid of biogenic amine target compounds.

CONCLUSIONS

This study highlights how test set selection can profoundly alter the conclusion of a (Q)SAR analysis. The makeup of said selection is heavily dependent on the nature of the data being used and the desired end point of the study. As such, random test set selection is not sufficient to reflect said data and endpoints and can potentially afford misleading conclusions. Using constraints in data and test set construction and analysis tailored specifically to the goal of the model, we have illustrated these issues in the context of druglikeness classification. The systematic differences that exist between ACD and WDI databases have been both illustrated and mitigated for our studies. The overpredictions inherent in random test set selection have been highlighted through the application of ontology constrained test set selection. Further, these studies suggest discontinuities in druglike space beyond those defined by Lipinski and others into ontology classes that form mainstays of the WDI.

These issues are by no means confined to the druglikeness models discussed within this article. For virtually all (Q)-SAR analyses involving general model construction (e.g., BBB permeability, scoring functions) that utilize large data sets derived from multiple drug discovery projects, similar problems are likely to exist. Those wishing to construct such models would be well advised to undertake test set selections with this in mind.

REFERENCES AND NOTES

(1) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins* **2002**, *49*, 457−471.

(2) Hert, J.; Willett, P.; Wilton, D. J. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(3) Olah, M.; Mracec, M. L.; Ostopovici, R. R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: New York, 2004; pp 223−239.

(4) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins* **2005**, *60*, 333−340.

(5) Tai, K.; Murdock, S.; Wu, B.; Ng, M. H.; Johnston, S.; Fangohr, H.; Cox, S. J.; Jeffreys, P.; Essex, J. W.; Sansom, M. S. P. BioSimGrid: towards a worldwide repository for biomolecular simulations. *Org. Biomol. Chem.* **2004**, *2*, 3219−3221.

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

(7) 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.

(8) Clark, D. E. Prediction of intestinal absorption and blood-brain barrier penetration by computational methods. *Comb. Chem. High Throughput Screening* **2001**, *4*, 477−496.

(9) Egan, W. J.; Walters, W. P.; Murcko, M. A. Guiding molecules towards drug-likeness. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 540−549.

(10) Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **2003**, *23*, 302−321.

(11) Lipinski, C. A. Filtering in drug discovery. *Ann. Rep. Comp. Chem.* **2005**, *1*, 155−168.

(12) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharm. Toxicol.* **2001**, *44*, 235−249.

(13) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(14) Ajay, A.; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−24.

(15) Wagener, M.; Van Geerestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280−292.

(16) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882−1889.

(17) Mueller, K. R.; Raetsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying 'Drug-likeness' with Kernel-Based Learning Methods. *J. Chem. Inf. Model.* **2005**, *45*, 249−253.

(18) World Drug Index (WDI) index for marketed and development drugs, developed and distributed by Thompson Scientific. http://scientific.thomson.com/products/wdi/ (accessed Oct 2006).

(19) MDL Drug Data Report (MDDR) and the Available Chemical Database (ACD) are developed and distributed by MDL Information Systems. www.mdl.com (accessed Oct 2006).

(20) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.

(21) Smith, R.; Hubbard, R. E.; Gschwend, D. A.; Leach, A. R.; Good, A. C. Analysis and optimization of structure-based virtual screening protocols (3). New methods and old problems in scoring function design. *J. Mol. Graphics Modell.* **2003**, *22*, 41−53.

(22) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a Method for Evaluating Drug-Likeness and Ease of Synthesis Using a Data Set in Which Compounds Are Assigned Scores Based on Chemists' Intuition. *J. Chem. Inf. Comput. Sci.* **2003,** *43*, 1269−1275.

(23) Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J. J.; Lecchini, S.; Jacoby, E. An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 947−955.

(24) Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856−864.

(25) Varma-O'Brien, S.; Rogers, D. Bayesian modeling in Pipeline PilotTM: Application to structural analysis of CDK2 inhibitors. Abstracts of Papers, 232nd ACS National Meeting, San Francisco, CA, United States, Sept. 10−14, 2006 CINF-008.

(26) MDL Information Systems. www.mdl.com (accessed Oct 2006). *Molecular Database Administration Guide, Ver. 2.0.1*; MDL Information Systems, Inc.: 1996; pp 2−13, 8−57.

(27) Lu, J. J.; Crimin, K.; Goodwin, J. T.; Crivori, P.; Orrenius, C.; Xing, L; Tandler, P. J.; Vidmar, T. J.; Amore, B. M.; Wilson, A. G. E.; Stouten, P. F. W.; Burton, P. S. Influence of Molecular Flexibility and Polar Surface Area Metrics on Oral Bioavailability in the Rat. *J. Med. Chem.* **2004**, *47*, 6104 −6107.

(28) Vieth, M.; Sutherland, J. J. Dependence of molecular properties on proteomic family of marketed drugs. *J. Med. Chem.* **2006**, *49*, 3451−3453.

(29) Hann, M. M.; Oprea, T. I. Pursuing the lead-likeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255−263.

(30) 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. www.rcsb.org/pdb/ (accessed Oct 2006).

(31) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(32) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(33) Narayanan, R.; Gunturi, S. B. In Silico ADME Modelling; Prediction models for blood brain barrier permeation using a systematic variable selection method. *Bioorg. Med. Chem.* **2005**, *13*, 3017−3028.

CI6003493