Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data

Sebastian G. Rohrer and Knut Baumann*

Institute of Pharmaceutical Chemistry, Beethovenstrasse 55, Braunschweig University of Technology, 38106 Braunschweig, Germany

Received August 1, 2008

Refined nearest neighbor analysis was recently introduced for the analysis of virtual screening benchmark data sets. It constitutes a technique from the field of spatial statistics and provides a mathematical framework for the nonparametric analysis of mapped point patterns. Here, refined nearest neighbor analysis is used to design benchmark data sets for virtual screening based on PubChem bioactivity data. A workflow is devised that purges data sets of compounds active against pharmaceutically relevant targets from unselective hits. Topological optimization using experimental design strategies monitored by refined nearest neighbor analysis functions is applied to generate corresponding data sets of actives and decoys that are unbiased with regard to analogue bias and artificial enrichment. These data sets provide a tool for Maximum Unbiased Validation (MUV) of virtual screening methods. The data sets and a software package implementing the MUV design workflow are freely available at http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html.

INTRODUCTION

Today, virtual screening (*VS*) is routinely applied in many drug discovery campaigns. Starting with large virtual databases of available compounds, the main goal of VS is to generate small focused subsets with an enriched fraction of active compounds in order to speed up biological testing.[1] This study will focus on ligand based virtual screening (*LBVS*), which utilizes the knowledge about one or several active compounds. Virtual libraries of potential drugs are ranked by their similarity to the query substances represented by descriptors, i.e. numerical representations of the molecules.[1] Literally thousands of methods are available for the conduction of LBVS.[2] As a consequence, a common problem of most VS campaigns is to decide which method will be suited best for the task at hand. Retrospective validation against data sets of compounds with known activity has been established as the standard approach to this problem.[3−7] A typical validation procedure consists of three major steps: (1) A part of the data set set is chosen (often randomly) to act as query. (2) The rest is pooled with a large number of compounds presumed to be inactive (frequently called "*decoys*") to become the validation set. (3) Different VS methods are then assessed according to their ability to separate the actives from the decoys in the final ranking. Several metrics or figures of merit (*FoM*) for the assessment of the quality of the generated ranking exist. The most commonly used metrics are the recall or retrieval rate (*RTR*), i.e. the fraction of actives found in the first percent of the ranking, and the area under the receiver operating characteristic curve (*ROC*). In this study, both metrics will be used in a complementary manner.

A basic precondition for the conduction of VS validation experiments is the availability of benchmark data sets of compounds with experimentally determined activity. A number of such benchmark data sets have been published for both ligand based virtual screening and structure based virtual screening (SBVS).[3,4,6−10] They range in size from tens to several hundreds of active compounds. The sources of these data sets vary, including among others the medicinal chemistry literature, public databases like the PDB,[11] commercial databases like the MDL Drug Data Report (MDDR),[12] or proprietary data of pharmaceutical companies. As a consequence, the experimental conditions leading to the qualification of a compound as active usually vary within each data set. It is tedious, in the case of literature data sets or, impossible, in the case of proprietary data, to examine and compare these experimental conditions.

Recent research has shown that the composition of the decoy sets has an immense effect on the validation of SBVS methods. Verdonk and co-workers showed that if the decoy set differs significantly from the set of actives regarding "simple" properties like molecular weight or number of hydrogen bond donors/acceptors, it may lead to "artificial enrichment", i.e. the classification is actually caused by the differences in "simple" properties.[13] They conclude that focusing the library of decoys to the same range of "simple" properties as the actives is essential for the results of VS validation to be representative. Recently DUD ("Directory of Useful Decoys"), a collection of validation sets for SBVS seeking to fulfill these requirements, has become available for public use.[10]

In addition to the composition of the decoy set, the composition of the data set of actives itself has a major influence on VS validation results. Findings of Good et al. show that data sets constructed from the literature or collected from drug discovery projects are prone to over-representation of certain scaffolds or chemical entities. This so-called "analogue bias" may lead to overoptimistic estimates of VS performance.[14−16] Recently, Clark et al. proposed a weight-

* Corresponding author phone: +49-531-3912751; fax: +49-531-3912799; e-mail: k.baumann@tu-braunschweig.de.

10.1021/ci8002649 CCC: $40.75
Published on Web 01/22/2009

ing scheme based on the ROC metric to correct validation results for analogue bias, which requires the assignment of the actives to subclasses by clustering algorithms.[17]

Recent research from our laboratory showed that the validation of LBVS methods is affected by both artificial enrichment and analogue bias. Indeed, data sets usually exhibit a combination of both phenomena.[18] This combined bias will be referred to as "benchmark data set bias" throughout the text. The same study also provided a nonparametric methodology within the framework of spatial statistics[19,20] to quantify the effect of benchmark data set bias by an analysis of the data set's topology in chemical space. In this context, chemical space is defined as the set of coordinates that can be occupied by vectors belonging to chemical compounds in the vector space spanned by a structural descriptor. The methodology utilizes two cumulative distribution functions of distances, the "*nearest-neighbor function*" $G(t)$ and the "*empty space function*" $F(t)$, which reflect the distributions of active-active and decoy-active distances, respectively. Holliday et al. successfully employed empirically derived probability density functions resembling $G(t)$ based on pairwise distances in chemical descriptor space for the validation of bioisosteric similarity metrics.[21] Since $G(t)$ is based on active-active distances, it reflects the level of self-similarity within the data set of actives and can therefore be used to detect analogue bias in validations. In a corresponding fashion, the degree of separation between decoys and actives can be measured using $F(t)$, thereby facilitating the detection of artificial enrichment. (See Methods for more detail.) By numerical integration of these functions, global figures for data set self-similarity (ΣG) and the separation between actives and decoys (ΣF) can be computed. By subtracting $G(t)$ from $F(t)$, a third function $S(t)$ and its numerical integral ΣS can be calculated for each data set. $S(t)$ and ΣS provide an estimate of "*data set clumping*", which comprises a combination of both self-similarity of actives and separation from the decoys. ΣS has the additional value of incorporating this global estimate of a data set's topology in a single scalar. In particular, negative values of ΣS indicate data set clumping, positive values indicate dispersion, and values near zero indicate a spatially random distribution of actives and decoys. It was demonstrated that overoptimistic VS validation results and benchmark data set bias can be linked to data set clumping.[18]

For the design of unbiased benchmark data sets that are not affected by artificial enrichment and analogue bias, $G(t)$ and $F(t)$ can effectively be employed as objective functions, if a descriptor capturing the molecular properties associated with these phenomena is employed. For this purpose, "Simple" descriptors have proven to be very useful.[18] Basically, simple descriptors as defined here are a vectorized form of the respective counts of all atoms, heavy atoms, boron, bromine, carbon, chlorine, fluorine, iodine, nitrogen, oxygen, phosphorus, and sulfur atoms in each molecule as well as the number of H-bond acceptors, H-bond donors, the logP, the number of chiral centers, and the number of ring systems. As a first condition for unbiased validation experiments, data sets should always exhibit a *dispersed* topology in the chemical space spanned by these simple descriptors. More precisely, active-active distances should be larger as or equal to decoy-active distances in simple descriptor space, because otherwise similarity searching becomes trivial. Topologies of this type have been shown to prevent both analogue bias and artificial enrichment.[18]

A second condition for the design of unbiased validation data sets arises from the fact that the goal of VS benchmark experiments is to identify the method with the best VS performance across a range of data sets. Experiments of this kind are most often applied to show how well a novel VS method performs in relation with available ones[22−24] or by pharmaceutical companies in order to determine which software to acquire.[7,25] In order to maximize the information of such experiments regarding the abilities of different methods, it is desirable to exclude data set topology as a factor influencing the validation experiments. This can be achieved by setting the extent of data set clumping in all data sets used in the validation to a constant value. Hence, the *differences* in data set composition must be minimized, i.e. all data sets should be adjusted to a common level of dispersion in simple molecular property space. Any arbitrary level of dispersion would suffice for benchmarking experiments of this type, as long as it is common to all data sets. However, the state of *"spatial randomness"* ($\Sigma S \approx 0$) is especially advantageous, since it implies that the distribution of simple molecular properties between actives and decoys contains no information about the respective bioactivities. The benchmark data sets presented here were therefore designed to be as close to spatial randomness as possible given the topology of the original data sets. This was achieved using ΣG and ΣF as objective functions in optimization algorithms.

The data sets are based on bioactivity data available in PubChem.[26,27] PubChem is the central repository of small molecule data of the Molecular Libraries Initiative[28] of the NIH Roadmap for Medical Research[29,30] and is composed of three major databases. Two will be used here: *PCCompound* provides chemical structures of the compounds tested as part of the NIH Roadmap effort. *PCBioAssay* lists bioactivity data for presently (July 2008) more than 640,000 compounds, derived from readouts of more than 1100 bioassays. Each record in PubChem is assigned a unique ID (*UID*) by which it can be easily accessed and retrieved. For *PCCompound* and *PCBioAssay*, these IDs are termed compound ID (*CID*) and assay ID (*AID*), respectively. Compared to other databases of bioactivity data, PubChem features several major advantages with respect to the design of VS benchmark data sets: (1) All data in PubChem, including structures of compounds, bioassay conditions, and experimental readouts, are publicly accessible. (2) Due to the specifications of the NIH Roadmap initiative, the compound collections tested in each bioassay exhibit a remarkable level of diversity. (3) The vast majority of tested compounds are "druglike". (4) For each assay, compounds that were found to be *inactive* are listed in addition to those found to be *active*. These inactive compounds can be used as decoys in validation experiments, as done recently by Hsieh et al. for a single data set in the validation of a QSAR modeling approach.[31] This provides the unique opportunity to design decoy sets, for which the inactivity against the target is actually experimentally validated. (5) PubChem is fully integrated into the NCBI Entrez[32] database system. Using the Entrez Programming Utilities (E-Utilities)[27] and the PubChem Power User Gateway (*PUG*)[33] automated chemoge-

nomics analyses are feasible, linking compounds with their bioactivities and the protein or DNA information of their targets.[34]

On the downside however, most of the bioactivity data available from PubChem is based on High-Throughput Screening (*HTS*) experiments. HTS data are notoriously affected by experimental noise and artifacts.[35−38] Thus, for the design of benchmark data sets it is essential to scrutinize PubChem bioactivity data with extreme thoroughness, which was ensured by a cautious selection of the raw data sets from PubChem and further applying an assay artifacts filter that removed any compounds for which the specificity of the reported bioactivity might be subject to any doubts. (See Methods for details.) This included potential aggregators (Hill slope filter), frequent hitters (frequency of hits filter), and compounds interfering with optical detection methods (autofluorescence and luciferase inhibition filter).

Topological optimization in simple descriptor space using experimental design strategies monitored by spatial statistics functions was then used to generate a collection of data sets of actives and corresponding decoy data sets that are unbiased with regard to both analogue bias and artificial enrichment and thus allow Maximum Unbiased Validation (MUV) of virtual screening methods.

METHODS

MUV Benchmark Data Set Design Strategy. The design strategy for MUV data sets comprised three major steps (Figure 1): (1) A collection of bioassays was extracted from *PCBioAssay* that justifies high confidence in the respective bioactivities. The compounds found active and inactive, respectively, formed the basis for the subsequent design steps. The resulting data sets of compounds were termed "potential actives" (*PA*) and "potential decoys" (*PD*). Filters were applied to the *PA* data sets that further purged all compounds for which the specificity of the respective bioactivities might be subject to any doubts. (2) The chemical space around the *PA* compounds was examined statistically in order to determine if the *PA* compounds are well embedded in decoys, a precondition for validation set design. (3) Experimental design algorithms were applied to select subsets of $k=30$ actives and $d=15000$ decoys from the *PA/PD* data sets with a spatially random distribution of actives and decoys regarding simple molecular properties. With constant data set sizes of $k=30$ and $d=15000$, MUV data sets also minimize the variance of validation results as demonstrated by Truchon et al.[39] The numerical values of k and d were chosen arbitrarily based on the size of the available bioactivity data sets. In principle, k and d could also be set to other values.

Selection of Sets of Potential Actives and Associated Bioassays. All assays with a specified protein target were extracted from *PCBioAssay*. From these, pairs of primary and confirmatory bioassays against the same target were selected. In these pairs, the bioactivity against the target is first determined for a large set (>50000) of compounds in a primary HTS experiment. The hits from the primary screen are then subjected to a low-throughput confirmatory screen testing for dose−response relationships. To be selected for the MUV design process, the confirmatory screens were further required to contain associated EC_{50} values. The actives from the confirmatory screens, referred to as Potential Actives (*PA*) in the text, and the inactives from the primary screens, referred to as Potential Decoys (*PD*), formed the basis for the generation of MUV data sets as presented here. The resulting data sets are summarized in Table 1 (Results). SD-files of the data sets were downloaded from PubChem by a Perl script that utilizes the PubChem Power User Gateway.[33]

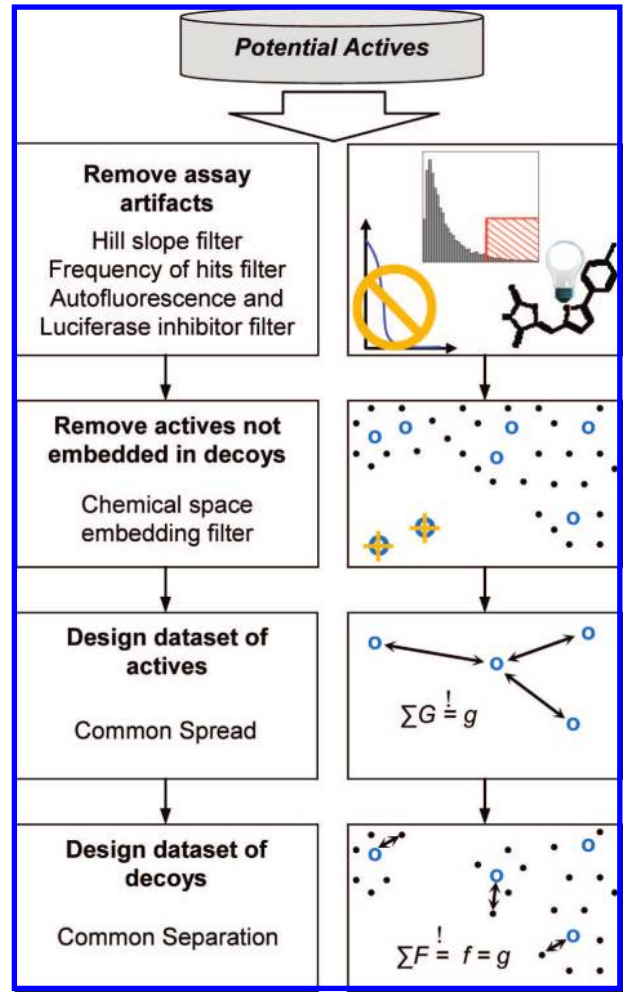

Figure 1. Synopsis of the MUV design workflow. Compounds with a potential for unspecific bioactivity are removed by the *assay artifacts filter*. Actives devoid of decoys are removed by the *chemical space embedding filter*. The spread of actives (ΣG) is adjusted to a common level g. Correspondingly, the separation between actives and decoys (ΣF) is adjusted to a common level f that is equal to g, thereby enforcing spatial randomness.

Assay Artifacts Filter. The requirement, that the bioactivity of all potential actives is determined by low-throughput dose−response experiments, justifies some confidence in the reliability of these assignments. Nevertheless, screening experiments are prone to artifacts caused by the tendency of some organic chemicals to form aggregates in aqueous buffers,[36] to exert off-target or cytotoxic effects or to interfere with the assay's optical detection method. In order to remove all compounds for which the specific mode of action could be subject to doubts from the *PA* data sets, a range of filters was applied, namely the "Hill slope filter", "frequency of hits filter", and the "autofluorescence and luciferase inhibition filter". Together these filters form the "assay artifacts filter". The particular filters were implemented as follows.

Hill Slope Filter. Aggregate formation is often associated with unusual Hill slope values[35,36,40,41] (also referred to as

Table 1. Bioactivity Data Sets from Pairs of Primary HTS and Confirmatory Dose-Response Experiments

target	mode of interaction	target class	prim. assay (AID)	confirm. assay (AID)	assay-type	Hill slope	PA	PD
S1P1 rec.	agonists	GPCR	449	466	reporter gene	PubChem	223[a]	55395
PKA	inhibitors	kinase	524	548	enzyme	PubChem	62	64814
SF1	inhibitors	nuclear receptor	525	600	reporter gene	PubChem	213	64550
Rho-Kinase2	inhibitors	kinase	604	644	enzyme	PubChem	67	59576
HIV RT-RNase	inhibitors	RNase	565	652	enzyme	PubChem	370	63969
Eph rec. A4	inhibitors	rec. tyr. kinase	689	689[b]	enzyme	PubChem	80	61480
SF1	agonists	nuclear receptor	522	692	reporter gene	PubChem	75	63683
HSP 90	inhibitors	chaperone	429	712	enzyme	calculated	91	63481
ER-a-coact. bind.	inhibitors	PPI[c]	629	713	enzyme	calculated	221	84656
ER-β-coact. bind.	inhibitors	PPI[c]	633	733	enzyme	calculated	194	84984
ER-a-coact. bind.	potentiators	PPI[c]	639	737	enzyme	not applicable[d]	64	84947
FAK	inhibitors	kinase	727	810	enzyme	PubChem	110	96070
Cathepsin G	inhibitors	protease	581	832	enzyme	PubChem	65	62007
FXIa	inhibitors	protease	798	846	enzyme	PubChem	70	218421
S1P2 rec.	inhibitors	GPCR	736	851	reporter gene	PubChem	54	96674
FXIIa	inhibitors	protease	800	852	enzyme	PubChem	99	216795
D1 rec.	allosteric modulators	GPCR	641	858	reporter gene	PubChem	226	54292
M1 rec.	allosteric inhibitors	GPCR	628	859	reporter gene	PubChem	231	61477

[a] A counterscreen was conducted using negative control cells not expressing the target (AID: 467). Actives as reported here are compounds active in assay 466 but not in assay 467. [b] Primary high-throughput assay and confirmatory dose response assay were carried out but reported as one record in *PCBioAssay*. [c] PPI: protein−protein interaction. [d] The mechanism of the PPI potentiation does not comply with Michaelis−Menten kinetics.

slope factors)[42] in dose−response curves. For competitive inhibition at a single-inhibitor binding site, the Hill slope h is expected to be approximately 1, based on Michaelis−Menten kinetics.[41] Although kinetic theory does not allow a ready application of this expectation to cell based assays or assays screening for allosteric modulation, very large Hill slopes nevertheless raise doubts about the specificity of the observed response. Generally, Hill slopes exceeding 2 are interpreted as harbingers of unspecific activity.[40,42]

Hill slopes for the dose response curves of all *PA* compounds were determined. If the Hill slopes were deposited in *PCBioAssay*, these values were used (Table 1). For all other *PA* compounds, Hill slopes were calculated directly from the PubChem dose−response data using GraphPad Prism 4.[43] For a *PA* compound to pass the Hill slope filter, its Hill slope h was required to be in the interval h=[0.5, 2]. The filter was supplemented by a list of experimentally verified aggregators recently deposited in *PCBioAssay* (AIDs: 584, 585).[40] Compounds identified as aggregators in this screen were removed from all *PA* data sets. The algorithm of the Hill slope filter is summarized in Figure 2.

Frequency of Hits (FoH) Filter. In addition to the special case of aggregate formation, bioassays are prone to a range of artifacts caused by unspecific activity of chemical compounds. For cell-based reporter gene assays, for instance, this is mainly caused by off-target or cytotoxic effects.[44] These artifacts have been associated with distinct molecular features[38] and cellular actions[44] of compounds showing unspecific activity, which have been termed "*frequent hitters*".[38] Several studies have tried to identify frequent hitters by first predicting their alleged mode of action as off-target promiscuous binders, aggregators, or cytotoxins. Frequent hitters are then flagged based on this prediction.[44−46] With a large resource of bioassay data such as PubChem, it is possible to flag a compound as an unspecific binder based on the ratio of the number of assays in which it occurs as a hit and the number of assays in which it was tested. This ratio is termed *frequency of hits (FoH)*. It is expected to be small for specific binders (tested in many assays but a hit in few assays) and large for unspecific binders (tested in many assays and a hit in most assays). The frequency distribution

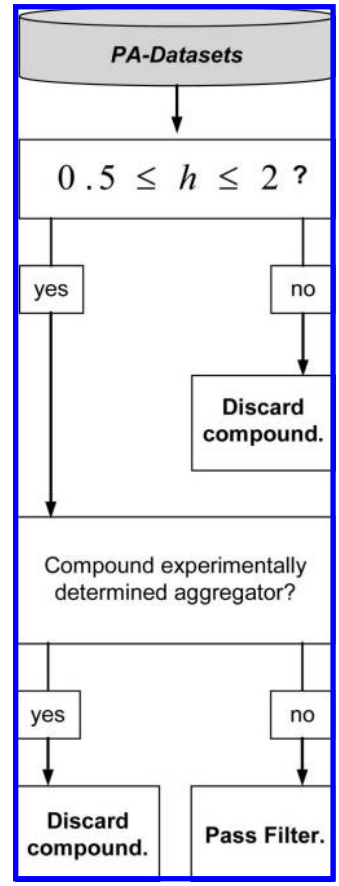

Figure 2. Flow-chart of the Hill slope filter. Compounds with undesirable Hill slopes h and compounds that have been experimentally shown to be aggregators are filtered from the *PA* data sets.

of *FoH* for a large set of compounds typically features two peaks: one at small values reflecting the population of specific binders and another one at large values of *FoH* indicating unspecific binders (Figure 3A). In order to distinguish the two populations, the first local minimum of the distribution can be utilized as a conservative, empirical cutoff. At this point, the descent of the *FoH* distribution after the first peak (specific binders) fades into the ascent to the second peak

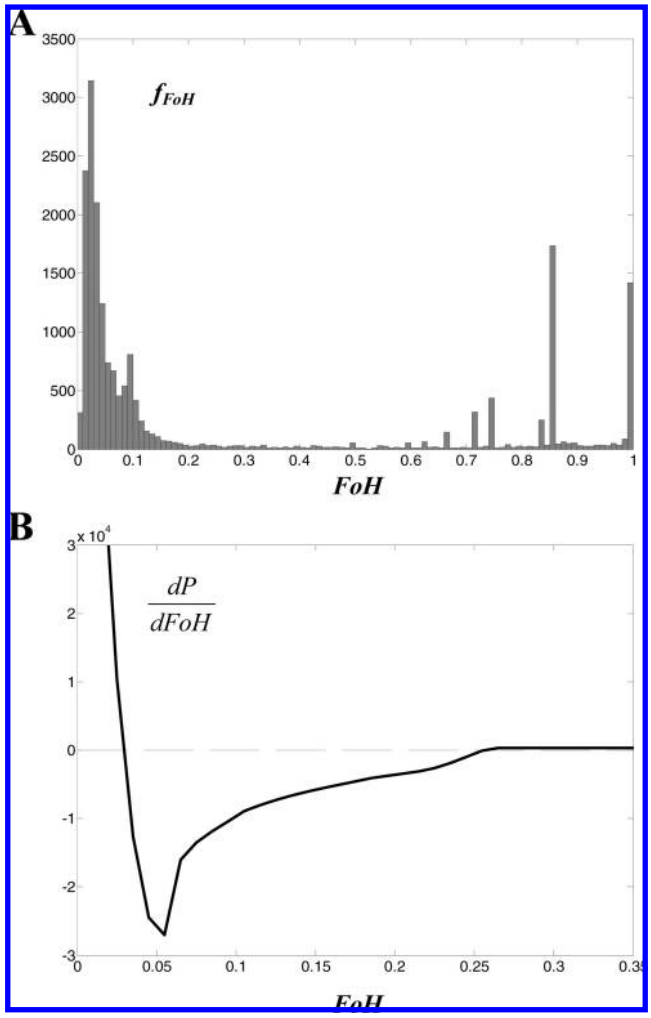

Figure 3. (**A**) Histogram of *FoH* for all compounds active in at least five *PCBioAssay* HTS screens. The histogram reflects two distinct populations of compounds: The left-hand part of the distribution (small *FoH*) is dominated by specific binders. The right-hand part (large *FoH*) reflects unspecific binders hitting in multiple assays. (**B**) The first derivative of a polynomial *P* fitted to the histogram has its 2nd zero crossing at *FoH*=0.26, corresponding to the first minimum of the *FoH* distribution. This was determined as the cutoff beyond which a compound was considered an unspecific binder.

(unspecific binders). Using piecewise bandwidth optimization,[47] a local polynomial *P* was fitted to the distribution of *FoH* determined for all PubChem compounds active in at least five bioassays (Figure 3B). The first minimum of *P* was determined as *FoH*=0.26. Consequently, compounds with a *FoH* larger than 0.26 were considered potentially unspecific binders and thus removed from the *PA* data sets.

In this analysis, it is important to take into account that some assays test for activity against very closely related targets. Whereas for example a compound that was found active in screens against four closely related cholinergic receptors might very well be a real binder to all of them, a compound found active against a kinase, a protease, and in two cytotoxicity screens is highly likely to be unspecific. Thus, it is desirable to correct the *FoH* for the presence of closely related assays in *PCBioAssay*. As a first measure, only large scale (>10000 compounds) HTS assays were considered in the *FoH* analysis, in order to exclude confirmation assays against identical targets from the statistic. The remaining assays were weighted according to the sequence

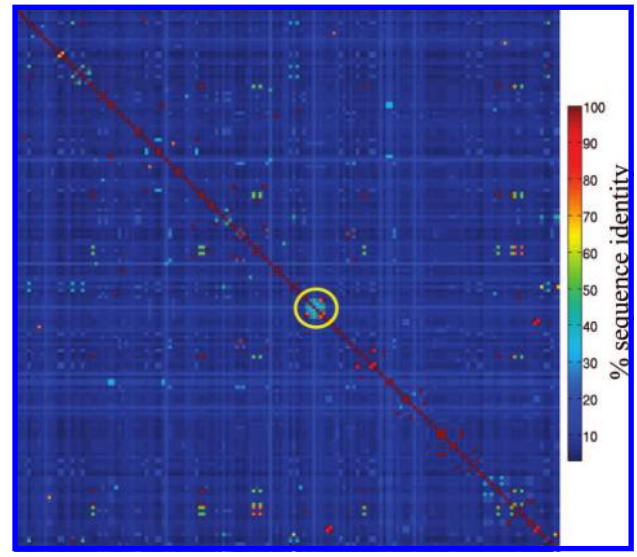

Figure 4. Heat-map visualization of the sequence similarity distance matrix of PubChem HTS assays with protein target annotation. Many groups of assays with closely related protein targets are perceptible, most notably a cluster of Ras-related GTPases (highlighted by the yellow circle).

identity of their respective protein targets. Of the 313 HTS assays in *PCBioAssay*, 163 were associated with protein target information. The respective protein sequences were downloaded from Entrez Protein[27] by a Perl script utilizing Entrez E-Utilities.[27] Using ClustalW[48,49] a multiple sequence alignment was constructed for all sequences. The resulting guide tree was converted into a distance matrix linking all assays by the pairwise percent sequence identity of their respective targets (Figure 4).

For each compound the set of assays, in which it was active, was determined. Weights were calculated for each assay as

$$w = 1 - (\%SI/100) \quad (1)$$

with %SI representing the percent sequence identity with the most closely related target associated to one of the assays in the set. All unrelated assays, including assays without protein target annotation, were weighted by 1. Using these weights, a weighted count of assays in which it was found active (*wAAC*) was calculated for each compound. The *FoH* of each compound was calculated as

$$FoH = wAAC/TAC \quad (2)$$

with *wAAC* representing the weighted number of assays in which the compound was found active, and *TAC* representing the number of assays in which it was tested. This weighted *FoH* score was actually used throughout the analysis of the PubChem data sets.

Autofluorescence and Luciferase Inhibition Filter. Assays based on optical detection are often affected by the chromo/fluorogenic properties of some compounds. Recently, a large scale fluorescence spectroscopic profiling of PubChem substances has been carried out and been reported in *PCBioAssay*.[50] Compounds that were found to exhibit undesirable properties in this profiling (AIDs: 587, 588, 590, 591, 592, 593, 594) were removed from the *PA* data sets. In a similar fashion, a recent screen tested a large set of PubChem substances for their potential to inhibit Luciferase.[51] Compounds found active in this screen (AID 411) were also removed from the PA data sets.

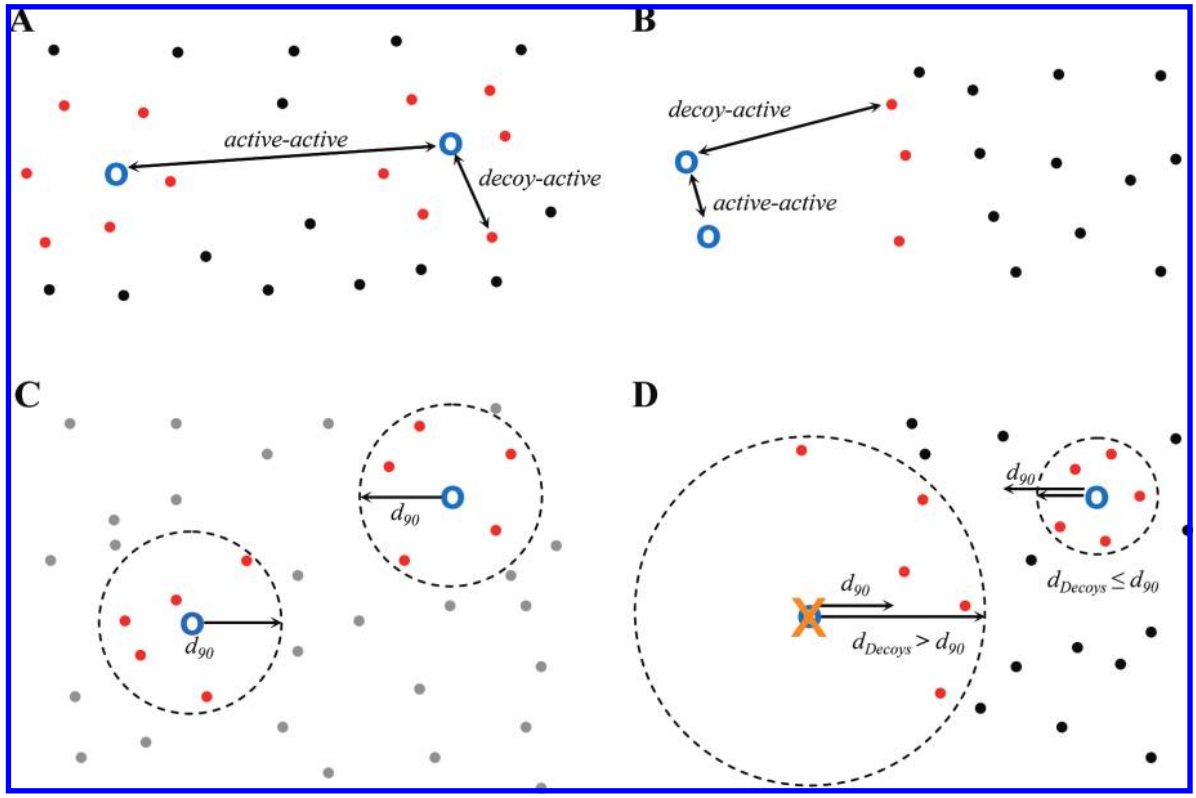

Figure 5. (A) A set of similar decoys (red) is selected for each active (blue circles). Since active-active distances are generally larger than decoy-active distances, artificial enrichment is prevented. (B) If actives are inadequately embedded in decoys, decoy-active distances result, that are larger than active-active distances even for the most similar decoys. Artificial enrichment is the consequence. (C) For each active the distance to the 500th nearest neighbor in a representative sample of compounds (gray dots) is determined. This corresponds to the radius of a hypersphere incorporating the 500 nearest neighbor compounds (red dots, five nearest neighbors are used here for the sake of clarity). d_{90} is determined as a 90% confidence boundary of this distance in 100 samples of compounds. (D) If an active is located in a region of chemical space that is significantly devoid of decoys (black dots), the hypersphere containing the 500th nearest decoys has a radius d_{decoys} larger than d_{90}. Such compounds are inadequately embedded in decoys and must be rejected.

Investigation of Potential False Negatives in the Data Sets of Decoys. In addition to false positives, which are addressed by the assay artifacts filter, HTS experiments are also affected by false negatives, i.e. active compounds are falsely designated as inactive.[52] This constitutes a potential source of error for the decoy sets presented in this study. In contrast to false positives, the data in *PCBioAssay* provide no means to detect potential false negatives. Furthermore, it is not possible to apply statistical methods such as bootstrapping in order to get an estimate of the error introduced by false negatives, since there is no way to reasonably estimate the rate of false negatives in PubChem single dose HTS assays. In the future, results of quantitative high throughput screening (*qHTS)* experiments may provide numerical data that might form the basis of such calculations.[53] At the moment however, these results are specific for the respective target, detection method, compound library, and tested compound concentration. Thus, they cannot be applied to the decoy sets utilized here (C. Austin, D. Auld, personal communication). In order to get at least some idea about the validity of the selected decoys, a similarity search employing simple descriptors and MAX-rule data fusion[3,4] was performed on each MUV decoy set, using the complete set of actives as query. For each MUV data set, the five decoys most similar to the actives were recorded. For each of the resulting 85 compounds, an extensive literature research was performed using SciFinder Scholar.[54] The results are summarized in Tables S1 and S2 (Supporting Information). In summary, no reports could be found in the literature that suggested a specific activity of any of the decoys against the target of its respective MUV data set. Although this evidence is at best anecdotal, together with the experimental result of inactivity in the HTS assay, it justifies some confidence about the inactivity of MUV decoys. Especially compared to traditional VS benchmark data sets, in which the inactivity of the decoys is merely assumed without any experimental evidence whatsoever,[3,4,7,10] this constitutes a considerable improvement.

Chemical Space Embedding Filter. In order to prevent artificial enrichment, decoys are selected similar to the set of actives regarding "simple" molecular properties. Usually, this is achieved by selecting a set of neighbors for each active **a** from a set of potential decoys (Figure 5A). However, if chemical space around **a** is devoid of decoys, no selection of decoys is possible that can prevent artificial enrichment (Figure 5B). Actives must be well embedded in decoys to allow unbiased decoy set design. Thus, actives inadequately embedded in decoys were removed from the *PA* data sets by a "*chemical space embedding filter*".

In order to quantitatively define "good embedding", a comprehensive sample of chemical space was compiled. Compounds were pooled from DrugBank,[55] Prous Drugs of the Future,[56] the Sigma-Aldrich chemistry catalogue,[57] and the MDDR.[12] This collection, which comprised 372021 unique compounds, will be referred to as the "*chemical space sample*" in the remainder of the text. It is safe to assume

that all actives are well embedded in this comprehensive collection of compounds. All data sets and the chemical space sample were encoded by "simple" descriptors. For each *PA* data set, a random subsample of the same size as the corresponding *PD* data set was drawn from the chemical space sample. For each active **a**, the distance to the 500th nearest neighbor in the random sample was determined. This was repeated 100 times, and the 90th percentile d_{90} was recorded as the 90% confidence boundary for a good embedding of **a** (Figure 5C). The distance of the active **a** to the 500th nearest neighbor in the decoy set, d_{Decoys}, was determined in an analogous fashion. Actives were considered inadequately embedded in decoys and thus removed from the *PA* data sets, if d_{Decoys} was larger than d_{90} (Figure 5D). The distance to the 500th nearest neighbor in the decoy set was chosen as a criterion for chemical space embedding, because 500 decoys were selected for each active in the process of MUV data set design (actives: k=30, decoys: d=15000).

Descriptors. SD-Files[58] of all *PA* and *PD* data sets were downloaded from PubChem as described above. Small fragments and counterions were removed and three-dimensional structures were generated using Molecular Networks CORINA.[59]

Simple descriptors are a vectorized form of the respective counts of all atoms, heavy atoms, boron, bromine, carbon, chlorine, fluorine, iodine, nitrogen, oxygen, phosphorus, and sulfur atoms in each molecule as well as the number of H-bond acceptors, H-bond donors, the logP, the number of chiral centers, and the number of ring systems. They were calculated using OpenEye BABEL3,[60] OpenEye FILTER,[61] and an in-house Perl script.

In order to validate how other descriptors perform on the MUV data sets, they were encoded by three additional classes of descriptors: SESP a class of versatile, alignment-independent 2D topological indices based on atom pairs,[62] MOE molecular properties descriptors,[63] and MACCS structural keys.[64] Bias introduced by the 3D conformation generator CORINA was excluded by using only the 2D class of MOE descriptors. Since the numerical values of properties in descriptors have significantly different ranges, all *PA/PD* descriptor matrices, except the ones encoded by MACCS keys, were autoscaled column-wise by subtraction of the mean and division by the standard deviation of the respective column in the chemical space sample. Columns with a standard deviation of 0 were removed from all descriptor matrices including MACCS keys.

After this pretreatment, the descriptor matrices of Simple, MACCS, MOE, and SESP had a dimensionality of 17, 154, 184, and 418, respectively. In order to reduce noise, principal components analysis (PCA)[65] was applied to the descriptor matrices of MOE and SESP, based on a singular value decomposition (SVD)[66] of the chemical space sample encoded by the respective descriptors. An analysis of the resulting eigenvalues showed that >90% of the total variance could be explained by the first 70 components for MOE and the first 94 components for SESP, respectively. Thus, the first 70 (MOE) and the first 94 (SESP) scores from the PCA were used as the final descriptors.

Spatial Statistics for Benchmark Data Set Design. Refined nearest neighbor analysis was recently introduced for the analysis of VS benchmark data set design.[18] It represents a mathematical framework for the analysis of mapped point patterns.[19,20] Here, it is applied to the vector representations of chemical compounds in descriptor space. In the context of virtual screening, refined nearest neighbor analysis is based on two functions from the positions of actives and decoys in descriptor space:

$G(t)$ is the proportion of actives for which the distance to the nearest neighbor active is less than t. $G(t)$ is called the "*nearest neighbor function*". Let n be the number of events, then $G(t)$ is given as

$$G(t) = \frac{\sum_i I_t(i,j)}{n};\ i = 1,...,n \tag{3}$$

with $I_t(i,j)$=1 if the distance of active i to its nearest neighbor j is smaller than t. $G(t)$ can be directly calculated from a descriptor representation of a data set of actives according to eq 3. It measures the level of self-similarity in the data set.

$F(t)$ is the proportion of decoys for which the distance to the nearest active is less than t. $F(t)$ is called the "*empty space function*", because it is able to detect gaps, the presence of multiple clusters, and their spacing. It is usually calculated based on points generated by bootstrapping from a large background database of compounds.[18] In this study, the compounds of the decoy sets were used as points in the calculation of $F(t)$. Thereby, $F(t)$ directly measures the degree of separation between decoys and actives. Let m be the number of points (decoys), and then $F(t)$ is given as

$$F(t) = \frac{\sum_j I_t(j,i)}{m};\ j = 1,...,m \tag{4}$$

with $I_t(j,i)$=1 if the distance of point j to the nearest event i is smaller than t.

By using $G(t)$ and $F(t)$ in a complementary fashion, the topology of a data set can be characterized comprehensively. Summing up the values of both functions over a range of distances, t_i, yields two scalars

$$\sum G = \sum_i G(t_i) \tag{5}$$

$$\sum F = \sum_i F(t_i) \tag{6}$$

that are robust estimates for the self-similarity (ΣG) and the separation from the decoys (ΣF). Large values of ΣG indicate a high level of self-similarity among the actives, whereas small values of ΣF indicate a high degree of separation from the decoys. Both figures can be combined to provide a single scalar ΣS that characterizes the degree of *data set clumping*:

$$\sum S = \sum_i (F(t_i) - G(t_i)) \tag{7}$$

Here, negative values indicate clumping in the data set, values of ΣS near zero indicate random-like distribution of actives and decoys, and positive values indicate that active-active distances are actually larger than decoy-active distances.

All topological figures discussed so far (ΣG, ΣF, ΣS) are based on cumulative distribution functions of distances in descriptor spaces with possibly different dimensionality. High-dimensional vector spaces are subject to the "*curse of dimensionality*", i.e. in higher dimensions distances are

inherently larger.[67] For ΣS to provide comparable results for different descriptors, it is therefore necessary to scale all figures characterizing data set topology ($G(t)$, $F(t)$) by a descriptor space specific factor. Preliminary experiments provided the median nearest neighbor distance d_{mnn} in a representative sample of chemical space as a good empirical estimate for this factor. The chemical space sample (see above) was encoded by all descriptors used in this study, and d_{mnn} was determined for each descriptor space. The range of values for t depends on the particular application of refined nearest neighbor analysis and must be determined empirically. Preliminary experiments showed that a maximum value of $t_{max}=3*d_{mnn}$ and a resolution of $\Delta t=t_{max}/500$ was sufficient to represent all topologies encountered throughout this study. Thus, $G(t)$, $F(t)$, and $S(t)$ were calculated according to eqs 3 and 4 with $t=[0,0.006,0.012,...,3]*d_{mnn}$.

Design of MUV Data Sets. The goal of MUV design is to generate sets with a spatially random distribution of actives and decoys in simple descriptor space. This was accomplished by a two step procedure. First the data sets of actives were adjusted to a common level of spread. Subsequently, decoy sets were selected with a common level of separation from the actives.

Subsets of $k=30$ actives with the maximum spread possible in each *PA* data set were generated using the well established Kennard-Stone[68] algorithm. Since this algorithm generates the maximum spread for each individual data set, the maximum *common* level of spread to which all data sets can be adjusted is the lowest level of spread observed among all Kennard-Stone subsets. This corresponds to the maximum value of ΣG among the subsets of $g=312$. A row-exchange algorithm[69] was applied to reduce the spread of all data sets with $g<312$ to a level of $g\approx312$. With the respective Kennard-Stone subsets of actives as a starting design, compounds were exchanged with the remaining *PA* compounds until the objective function

$$D = g - \sum G_{dataset} \tag{8}$$

reached values smaller than $D=2$, with $\sum G_{dataset}$ representing ΣG of the data set at the respective iteration. The search was also terminated, if the set remained constant for more than 10 iterations. As an additional constraint, only compounds were allowed to be selected that had a nearest neighbor distance larger than $r=0.8$ to the compounds already in the selection. The dissimilarity constraint r is used in analogy to the well-established OptiSim algorithm[70] for diverse subset selection. The application of this dissimilarity constraint is necessary in order to prevent a selection of actives that would cause an early but flat ascent in $G(t)$. Data sets with such curves of $G(t)$ might contain clusters of self-similar actives and still exhibit values of ΣS close to 0, if the respective curve for $F(t)$ featured a complementary late, but steep ascent. Thereby the self-similarity present in the data set would be masked by the cancelation effects associated with the numerical integration used in the calculation of ΣS. Both cutoff values of $D=2$ and $r=0.8$ were determined empirically by preliminary experiments.

For actives and decoys to exhibit a spatially random distribution, ΣF must be equal to ΣG. Therefore, a starting design of decoys was generated by selecting the 500 most similar decoys for each active, resulting in a set of $d=15000$ decoys for each data set. ΣF was adjusted to $f=g=312$ by a genetic algorithm with

$$D = f - \sum F_{dataset} \tag{9}$$

as the fitness function, with $\sum F_{dataset}$ representing ΣF of the data set at the respective iteration. Convergence was reached if $D<2$ or D remained constant for more than 10 iterations. Again, $D=2$ was determined as a reasonable fitness cutoff by preliminary experiments.

Retrospective Virtual Screening Simulations. In order to demonstrate the application of the MUV data sets in VS validation experiments, retrospective VS simulations were carried out. The data sets of actives and decoys were encoded by all descriptors (Simple, MOE, SESP, MACCS). For each run, a query of one or ten compounds, respectively, was chosen randomly from the data sets of actives. The remaining actives were pooled with the respective decoys to form the validation set. For each MUV data set of actives, 100 such random splits were generated, in order to obtain a mean value of VS performance that is not affected by the random choice of the query molecules. Similarity was measured by Euclidean distance, and the validation sets were ranked accordingly. For queries consisting of ten compounds, MAX-rule data fusion[3,4] was applied to the ranking.

Figures of Merit (FoM) for Virtual Screening Performance. VS performance was measured by the area under the receiver operating characteristic curve (*ROC*).[71,72] Additionally, the ability for early recognition of active substances was quantified by the fraction of retrieved actives (Retrieval Rate, *RTR*) in the first percent of the ranked validation set. The mean areas under the receiver operating characteristic curves and mean Retrieval Rates obtained from the 100 random query/validation set splits generated for each data set will be denoted *mean*(*ROC*) and *mean*(*RTR*) throughout the text.

RESULTS AND DISCUSSION

Bioactivity Data Sets from PubChem. From the bioactivity data available in *PCBioAssay*, 18 pairs of assays against pharmaceutically relevant targets were selected. Each of these pairs included a primary high-throughput screen and a low-throughput confirmation assay against the same target. *Active* compounds from the confirmation assays were used as the data set of potential actives (*PA*) for the subsequent design steps. In a complementary fashion, *inactive* compounds from the corresponding primary screen were used as the data set of potential decoys (*PD*). An overview of the respective pairs of bioassays is provided by Table 1. High standards regarding the specificity of the bioactivities in the data sets of actives were enforced by selecting only low-throughput confirmatory assays with associated dose−response information and EC_{50} values as *PA* data sets. An assay artifacts filter further removed compounds with a potential for unspecific activity, including aggregators, promiscuous binders, and compounds interfering with optical detection methods (Table 2). In addition to the data sets of actives, the specificity of the data sets of decoys was enforced by utilizing only compounds for potential decoy (*PD*) data sets, whose inactivity against the respective target was experimentally determined by HTS. Although decoys that were false negatives in the HTS cannot be detected by these procedures, the level of confidence in

Table 2. Effect of MUV Design Strategy on *PA* Data Sets

data set		compounds in data set after application of filter/design				distinct molecular scaffolds[a]	
target	AID	PA	assay artifacts filter	ch. space emb. filter	MUV design	PA	MUV
S1P1 rec.	466	223	185 (−38)	180 (−5)	30	127	28
PKA	548	62	51 (−11)	50 (−1)	30	37	27
SF1	600	213	71 (−142)	70 (−1)	30	47	24
Rho-Kinase2	644	67	58 (−9)	57 (−1)	30	39	27
HIV RT-RNase	652	370	178 (−192)	169 (−9)	30	121	27
Eph rec. A4	689	80	58 (−22)	56 (−2)	30	48	29
SF1	692	75	37 (−38)	37 (−0)	30	36	30
HSP 90	712	90	46 (−44)	44 (−2)	30	32	27
ER-a-coact. bind. inh.	713	221	98 (−123)	92 (−6)	30	81	26
ER-β-coact. bind. inh.	733	194	104 (−90)	101 (−3)	30	78	28
ER-a-coact. bind. pot.	737	64	48 (−16)	42 (−6)	30	39	28
FAK	810	110	71 (−39)	62 (−9)	30	51	28
Cathepsin G	832	65	51 (−14)	51 (−0)	30	31	24
FXIa	846	70	61 (−9)	60 (−1)	30	31	21
S1P2 rec.	851	54	28 (−36)	23 (−5)	23	16	16
FXIIa	852	99	81 (−18)	80 (−1)	30	39	24
D1 Rec.	858	226	140 (−86)	138 (−2)	30	106	24
M1 Rec.	859	234	149 (−85)	133 (−16)	30	103	29

[a] Molecular scaffolds were determined using MOE[63] following the approach by Bemis and Murcko.[73]

the inactivity of the MUV decoys is still higher than for benchmark data sets that merely assume compounds without any annotated activity to be inactive.

In order to efficiently prevent artificial enrichment by decoy set design, it is necessary that enough decoys are available in the chemical space around each active, i.e. each active must be adequately *embedded* in decoys (Figure 5). Active compounds that were not adequately embedded in decoys were removed from the *PA* data sets by a *chemical space embedding filter*. In some cases, the amount of inadequately embedded actives exceeded 10% of the size of the respective unfiltered data sets (Table 2). This highlights the importance of the removal of these compounds, which might otherwise cause substantial levels of artificial enrichment.

MUV Benchmark Data Sets: General Properties. Subsets of k=30 actives and d=15000 decoys were selected from each *PA/PD* pair of data sets that were optimal regarding the criterion of spatial randomness (see below). The resulting data sets contain a remarkably high number of distinct molecular scaffolds (Table 2). On average, MUV data sets contain only 1.16 compounds per scaffold class, a ratio that effectively eliminates analogue bias. Here, scaffolds were defined as reduced molecular graphs as proposed by Bemis and Murcko.[73] Further, MUV data sets provide a good representation of druglike chemical space. As shown by Table S3 (Supporting Information), violations of Lipinski's Rule of 5 occur with very low frequency, both in the data sets of actives and decoys.

Only 23 compounds in the *PA* data set of S1P2 receptor inhibitors (AID 851) passed all filters, a number that is insufficient for MUV design of the data set of actives. The data set is kept here for illustrative purposes but will not be part of the final MUV collection, which consequently consists of 17 benchmark data sets.

Spatial Statistics Analysis of MUV Data Sets. Figure 6 gives an overview of data set clumping in simple descriptor space before and after MUV data set design. Before MUV design, all *PA/PD* data sets show considerable clumping indicated by negative values of ΣS. Overoptimistic validation results would be the consequence, if these data sets were used for VS validation without further design. MUV data sets, on the other hand, exhibit positive ΣS values close to 0, indicating mildly dispersed distributions of actives and decoys close to spatial randomness. With only small variations in topology between the MUV data sets, differences in VS performance between the data sets are largely independent of simple molecular properties. Table S4 (Supporting Information) provides ΣS, ΣG, and ΣF for all MUV data sets.

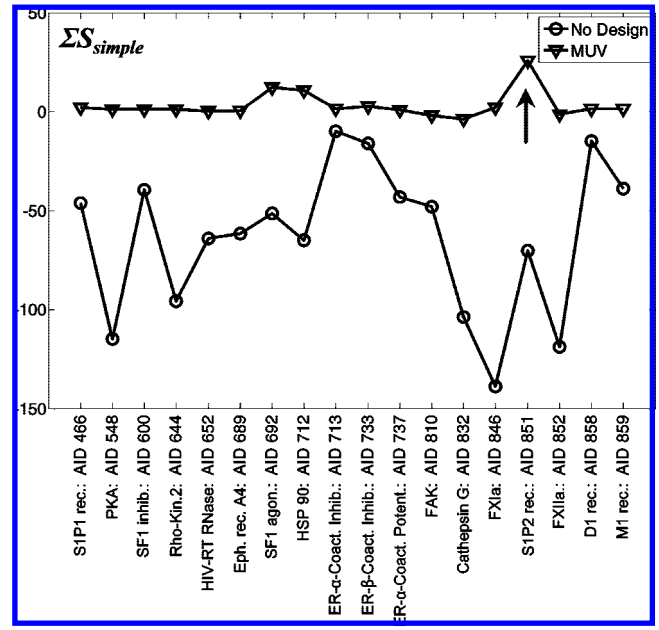

Figure 6. Effect of MUV design on data set topologies in simple descriptor space. Without MUV design all data sets show clumping (negative values of ΣS_{simple}), indicating the danger of benchmark data set bias. MUV data sets exhibit ΣS_{simple} values close to 0, indicating spatial randomness of actives and decoys. Data set AID 851 (indicated by the arrow) does not fulfill the criteria. No design is feasible on the respective data set of actives, since it contains less than 30 compounds after application of the assay artifacts filter.

Figure 7 illustrates the effect of MUV design on the topology of data sets visualized in more detail for two example data sets (AID 548, AID 858). Graphs of the

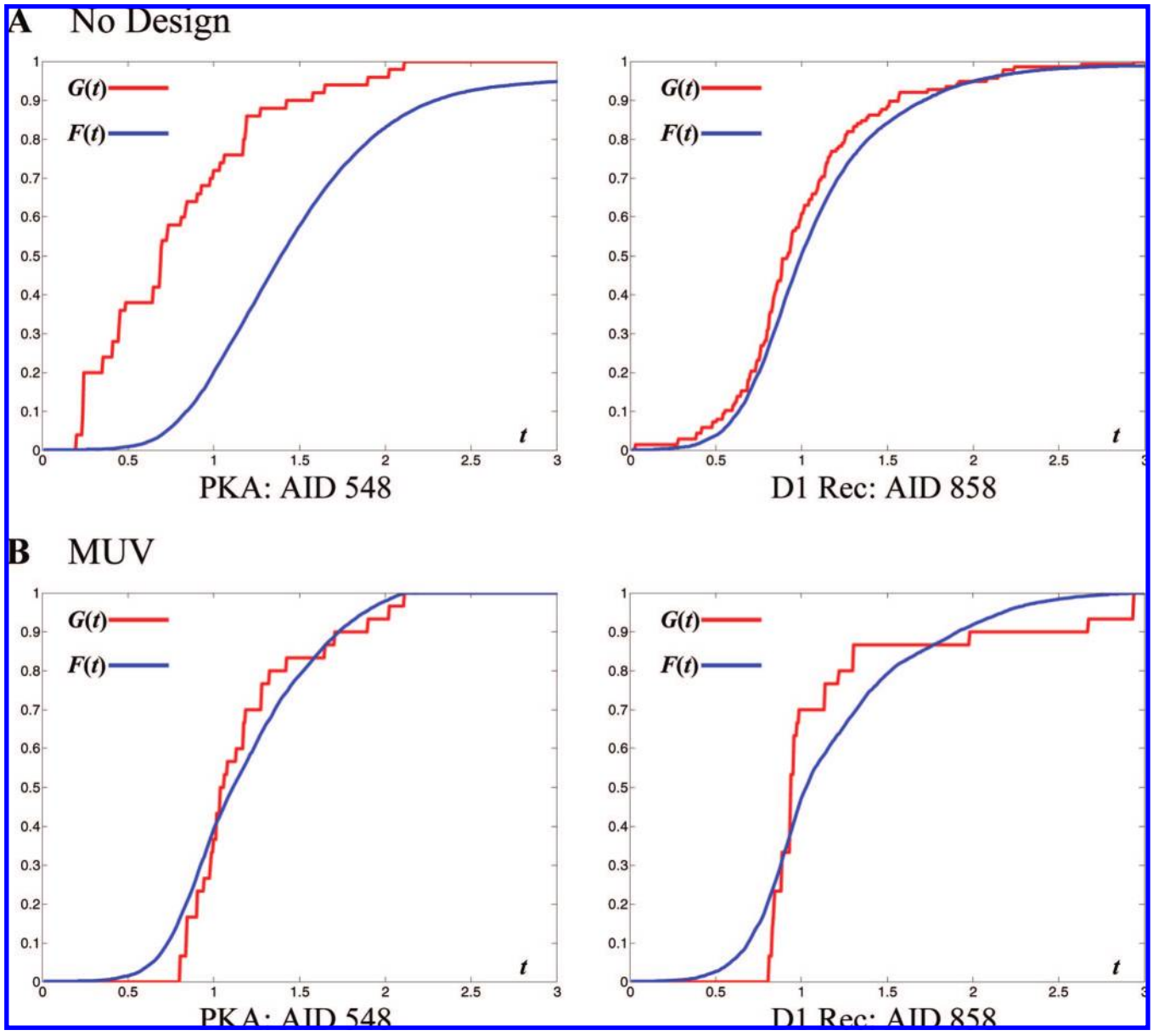

Figure 7. Effect of MUV design on the topology of benchmark data sets AID 548 and AID 858 in simple descriptor space. (A) Different degrees of data set clumping are observable in the raw data sets. The large shift of $F(t)$ relative to $G(t)$ in AID 548 indicates extensive clumping. (B) MUV design smoothes the differences in topology between the data sets. Both, $G(t)$ and $F(t)$, show similar curves for both data sets.

nearest-neighbor function $G(t)$ (red) and the empty space function $F(t)$ (blue) provide information about the self-similarity in the set of actives and the separation between actives and decoys before (Figure 7A) and after (Figure 7B) MUV design. Here, an early ascent in the cumulative distribution of nearest neighbor distances $G(t)$, i.e. the presence of many actives with a very small distance t to their nearest neighbor, indicates high self-similarity among actives. On the other hand, a late ascent in $F(t)$ implies the presence of a high proportion of decoys with a large distance t to the nearest active, i.e. a high level of separation. Any rightward shift of $F(t)$ relative to $G(t)$ results in negative values of ΣS and thereby indicates data set clumping. The objective of MUV design is 2-fold: (i) minimize the differences in data set clumping between the data sets, i.e. generate similar curves of $G(t)$ and $F(t)$ for all data sets. (ii) Generate spatially random topologies ($\Sigma S \approx 0$), i.e. minimize the shift between $G(t)$ and $F(t)$. Figure 7 shows that both goals of MUV design are achieved for the two data sets examined. Before MUV design (Figure 7A) both data sets exhibit the rightward shift in $F(t)$ relative to $G(t)$ that indicate data set clumping. Moreover, the topologies of both data sets are clearly different, with a large extent of clumping in data set AID 548 and only mild clumping in data set AID 858. After MUV design (Figure 7B) both data sets show spatial randomness, i.e. no rightward shift between $F(t)$ and $G(t)$. Furthermore, the curves of both functions are similar for both data sets, i.e. differences in topology are minimized.

Application of MUV Data Sets for LBVS Benchmarking. In order to demonstrate the utility of the MUV approach, the PubChem bioactivity data sets before and after MUV design were encoded by different descriptors: the aforementioned simple descriptors and three additional classes of descriptors: SESP,[62] which is based on 2D atom pairs, MOE molecular properties descriptors[63] and MACCS structural keys,[64] both of which have found extensive use in the

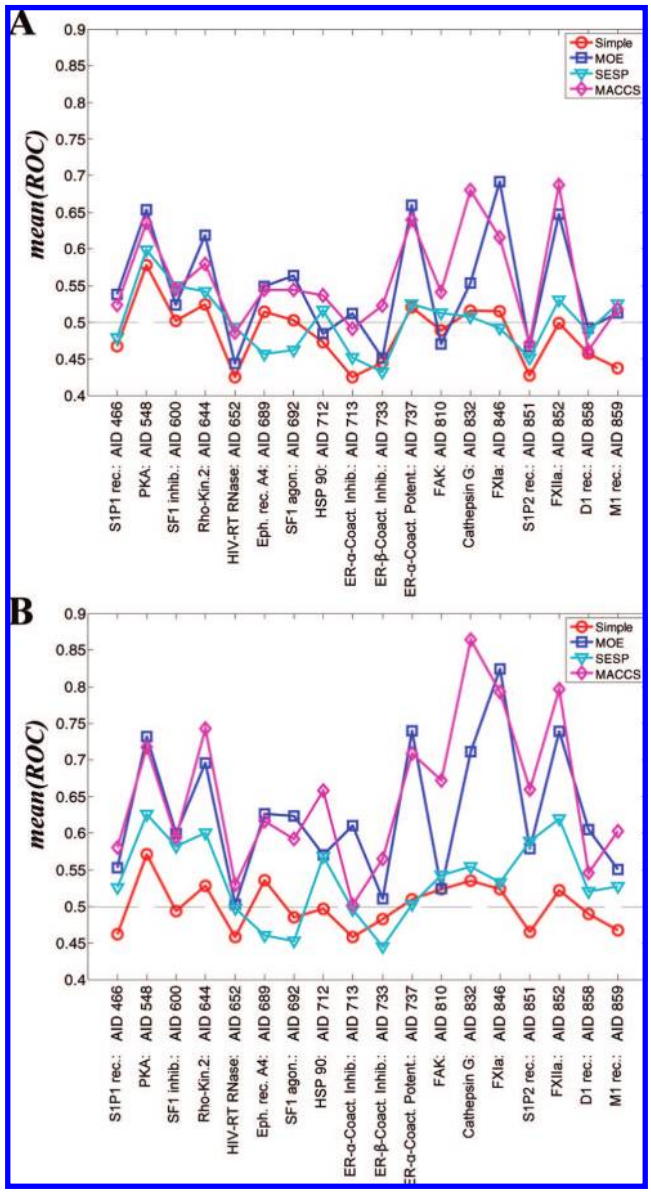

Figure 8. Performance of different VS methods in retrospective VS simulations on MUV data sets: (A) 1 query compound and (B) 10 query compounds. Generally, MACCS and MOE perform best. The expectation of *mean*(*ROC*)=0.5 for random rankings is indicated by the dashed line.

literature for VS validation tasks.[18,74,75] Retrospective LBVS simulations were carried out on the original and on the MUV data sets. From each data set one or ten actives, respectively, were chosen randomly as query molecules. The rest of the actives was pooled with the corresponding decoys, and similarity searching was carried out. For searches with 10 query molecules, MAX-rule data fusion[3,4] was applied to the ranking. This was repeated 100 times. Average VS performance for the 100 query/validation set splits was measured by the mean area under the receiver operating curve (mean(*ROC*)) and the mean retrieval rate (mean(*RTR*), Table S5, Supporting Information) to minimize variability caused by the random selection of the query set. The *mean*(*ROC*) values obtained on the MUV data sets are summarized by Figure 8.

As a first, more general observation, Figure 8 shows that when encoded by simple descriptors, none of the data sets exhibited a *mean*(*ROC*) significantly exceeding the random ranking expectation of 0.5. This indicates that the spatial randomness of MUV data sets in simple descriptor space is well reflected in the respective VS ranking. MOE descriptors and MACCS keys generally performed best. Similarity searching with 10 query compounds and data fusion worked better than single query searches. Moreover, the performance of SESP was found to be superior to simple descriptors only in very few cases. This is expected, since SESP is based on count statistics of atom pairs, which are highly correlated with the atom counts employed by simple descriptors.

However, the goal of this paper is not to determine the most superior descriptor or searching method. The objective of MUV data set design is to prevent bias introduced by benchmark data set composition from distorting validation experiments. Such benchmark data set bias occurs whenever the results of VS validation experiments largely depend on the simple property composition of the employed benchmark data set, rather than the actual performance of the tested methods. Thus, the validation of a method is affected by benchmark data set bias, if its performance is highly correlated with the topology of the data sets in simple descriptor space. Consequently, benchmark data set bias can be detected numerically using the correlation coefficient $\rho(\Sigma S_{simple}, mean(ROC))$ between the performance of a given VS method ($mean(ROC)$) and data set clumping in simple descriptor space ΣS_{simple}. Since negative values of ΣS_{simple} indicate a higher degree of clumping, VS performance and ΣS_{simple} are negatively correlated. Therefore, values of $\rho(\Sigma S_{simple}, mean(ROC))$ close to -1 indicate that the validation results are biased by the simple molecular properties of the benchmark data sets. Values near 0, on the other hand, indicate that the validation is not influenced by simple molecular properties.

Figure 9 provides graphs visualizing the correlation between the $mean(ROC)$ of the tested descriptors and ΣS_{simple} for similarity searches using 10 query compounds. A rank transformation was applied to the data, since $mean(ROC)$ and ΣS_{simple} span considerably different numerical regions and because there is no evidence for a linear relation between them. Before MUV design, a tight correlation between ΣS_{simple} and the VS performances of all tested descriptors is easily observable, both in the original data domain and after the rank transformation. Thus, before MUV design, data set composition regarding simple molecular properties dominates the validation results of all descriptors, i.e. the original data sets are affected by benchmark data set bias. After MUV design, this correlation no longer exists, and benchmark data set bias is effectively prevented. These observations are supported by Table 3, which provides the respective Spearman rank correlation coefficients $\rho(\Sigma S_{simple}, mean(ROC))$ for single and multiple query searches. On the original data sets all tested descriptors exhibit large negative correlation coefficients that indicate considerable benchmark data set bias. MUV design reduces this correlation below the level of statistical significance and thereby prevents benchmark data set bias. Here, it might be noteworthy that in all tested cases MOE descriptors exhibited the smallest extent of correlation with ΣS_{simple}. These results suggest that of the descriptors tested here, MOE descriptors are least susceptible to benchmark data set bias.

Comparison of MUV with DUD. The directory of useful decoys DUD is a collection of VS benchmark data sets

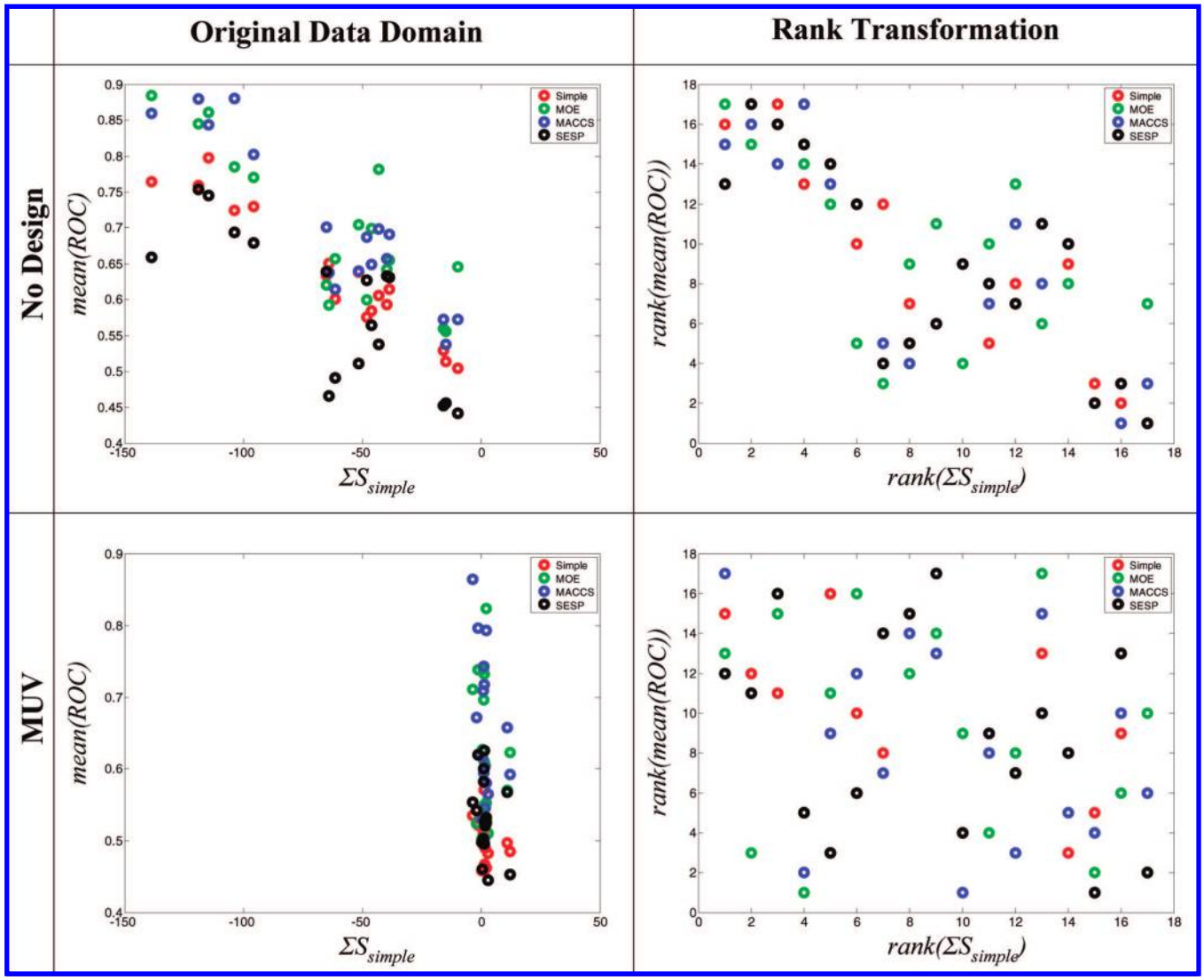

Figure 9. Plots of the VS performance (*mean*(*ROC*)) obtained with four different descriptors (Simple, MOE, MACCS, SESP) on the 17 benchmark data sets vs data set clumping in simple descriptor space (ΣS_{simple}) before and after MUV design (similarity searches, 10 query compounds). Before MUV design (top) a clear correlation between *mean*(*ROC*) and ΣS_{simple} is observable, both in the original data domain (left) and after a rank transformation (right). MUV design (bottom) causes a decorrelation of *mean*(*ROC*)) and ΣS_{simple}. Benchmark data set bias is prevented.

Table 3. Correlation Coefficients between Data Set Clumping in Simple Descriptor Space (ΣS_{simple}) and VS Performance (*mean*(*ROC*)) of Tested Descriptors Before and After MUV Design

	no design				MUV			
$\rho(\Sigma S_{simple}, mean(ROC))$[a]	Simple	MOE	SESP	MACCS	Simple	MOE	SESP	MACCS
1 query cmpd	−0.84	−0.43	−0.72	−0.80	−0.22	−0.08	−0.37	−0.41
10 query cmpds	−0.91	−0.69	−0.78	−0.79	−0.41	−0.17	−0.32	−0.41
conf. itv. boundary[b]				−0.41				

[a] ρ: Spearman rank correlation coefficients. The observed effect is even more pronounced when using the commonly used Pearson correlation coefficient. However, there is no hard evidence for a direct linear relation between ΣS_{simple} and *mean*(*ROC*). Therefore Spearman rank correlation coefficients were considered more appropriate. [b] The boundary of the one-sided 95% confidence interval of ρ for rejecting the null hypothesis of no correlation was calculated as the 5th percentile (left tail) of a distribution of correlation coefficients generated by 100000-fold random permutation of the respective ΣS_{simple} and *mean*(*ROC*) values.[76]

designed to prevent artificial enrichment in the validation of structure based virtual screening methods.[10] Since its publication, DUD has become the *de facto* standard for the validation of docking methods. Briefly, DUD comprises a collection of 40 data sets of actives with differing sizes compiled from various sources. All compounds in the ZINC database[77] not described to be active against one of the 40 targets were used as potential decoys. Potential decoys were required to have a Tanimoto similarity of less than 0.9 to any of the actives based on CACTVS type 2 substructure keys[78] in order to prevent the selection of false negatives as decoys. Using Schrödinger QikProp,[79] a vector of physical properties quite similar to the simple descriptor used here was calculated for actives and potential decoys. Based on

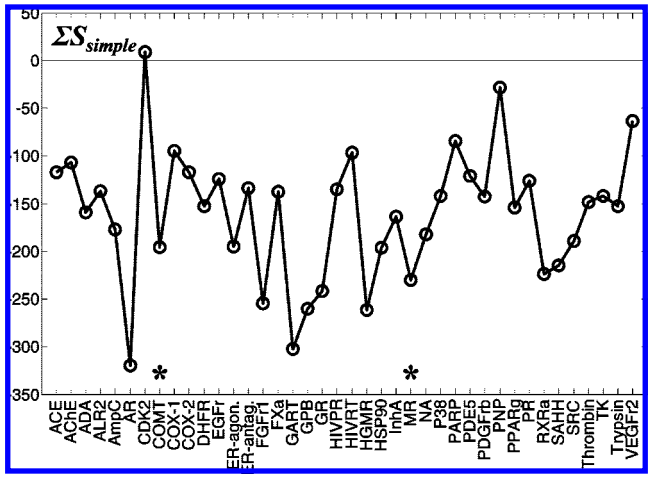

Figure 10. Data set clumping of DUD benchmark data sets measured by ΣS in simple descriptor space. Large negative values indicate considerable self-similarity and separation from the decoys regarding simple molecular properties. Asterisks indicate data sets with less than 20 compounds, for which results of VS runs using 10 query compounds are not representative.

the QikProp descriptor, the 36 most similar decoys were selected for each active from the set of potential decoys in order to generate decoy sets with minimum separation.

In order to compare the DUD collection of benchmark data sets with MUV, SD-files[58] of the DUD data set (Release 2) were downloaded from the DUD Web site[80] and encoded by simple and MOE descriptors. MOE descriptors were chosen, because they were found to be the least susceptible to benchmark data set bias. (See above.) Spatial statistics analysis and retrospective LBVS simulations were conducted in an analogous fashion as for the MUV data sets. The results are summarized in Figures 10 and 11 and Table 4. Most DUD data sets exhibit high levels of clumping in simple descriptor space, indicated by large negative values of ΣS_{simple} (Figure 10). Considerable differences in data set topology exist between the data sets. This corresponds with significant correlation between the VS performance of both simple and MOE descriptors with ΣS_{simple} (Figure 11, Table 4). The large number of datapoints with $mean(ROC) \geq 0.8$ for simple descriptors (Figure 11A) is particularly striking. It indicates that retrieval of actives from the DUD data sets is not very challenging, even for simple descriptors that do not encode any type of higher level molecular information like pharmacophore features or substructure elements. Apparently, DUD is subject to considerable benchmark data set bias. Apart from errors in the chemical structures of several DUD ligands[81] and the presence of duplicates in the DUD data sets,[23] two main factors can be identified. (i) As shown recently by Good et al., DUD is seriously affected by analogue bias.[16] Applying the same algorithm by Bemis and Murcko as in the analysis of MUV data sets (see above), the average number of compounds per scaffold class in DUD was determined to be 4.56 (MUV: 1.16). This is in accordance with high values of ΣG observed for all DUD data sets of actives (Table S6, Supporting Information), indicating high levels of self-similarity within the data sets. Here, it is important to point out that DUD was designed for the validation of docking programs, and its application to LBVS is strongly discouraged by the DUD authors.[10,82] Consequently, no diversity selection was applied to the data

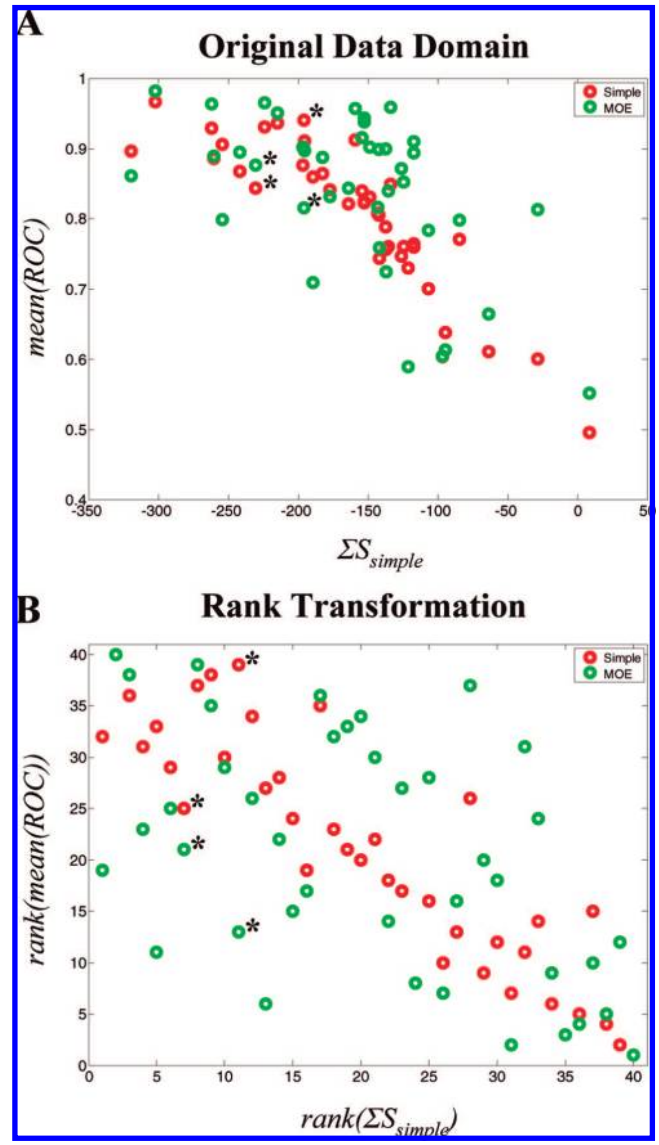

Figure 11. ($mean(ROC)$) plotted against ΣS_{simple} for MOE and simple descriptors (10 query cmpds) on DUD benchmark data sets. (A) Original data domain. A tight correlation is observable for both simple and MOE descriptors in the original data domain. (B) The correlation persists also after rank transformation, although less obvious for MOE descriptors. Asterisks indicate data sets with less than 20 compounds.

Table 4. Correlation Coefficients of DUD Data Set Clumping in Simple Descriptor Space (ΣS_{simple}) and VS Performance of Tested Descriptors

$\rho(\Sigma S_{simple}, mean(ROC))$[a]	Simple	MOE
10 query cmpds	−0.92	−0.54
conf. itv.boundary[b]	−0.27	

[a] ρ: Spearman rank correlation coefficient. Data sets with less than 20 compounds were excluded from the calculation. [b] Boundary of the one-sided 95% confidence interval of ρ for rejecting the null hypothesis of no correlation. Also see Table 3.

sets of actives in the generation of DUD, which explains the high number of compounds per scaffold class. Since docking tools are less sensitive to self-similarity in the data set of actives, the influence of such analogue bias is most likely limited. However, for the validation of methods that incorporate any form of ligand information, such as recently by Reid et al.,[23] the analogue bias in DUD is critical. (ii)

DUD decoy sets also exhibit considerable levels of separation from the actives in simple descriptor space, indicated by small values in ΣF (Table S6, Supporting Information). This is surprising, because the explicit principle of DUD design is the selection of decoys that are minimally separated from the actives. Since we do not have access to the original DUD potential decoy set, it is difficult to precisely specify causes for this problem. One possible reason might be that, according to our analysis, on average 13% of the actives in DUD are inadequately embedded in decoys, with peak values of 60% for the data set of PDGFrb inhibitors and 51% for InhA inhibitors. It is quite probable that this causes a certain degree of separation. Another possible reason is that the Tanimoto dissimilarity criterion applied to potential decoys in the design of DUD is too harsh. This might create "bubbles" devoid of decoys in chemical space around actives.

MUV data sets utilize decoys, for which inactivity against the respective targets is experimentally determined, which renders a minimum dissimilarity between actives and decoys obsolete. The associated problems in the design of decoy data sets are thus circumvented.

CONCLUSION

A collection of benchmark data sets for ligand based virtual screening methods was generated from PubChem bioactivity data. These data sets minimize the influence of benchmark data set bias on validation results and therefore provide a tool for the Maximum Unbiased Validation (MUV) of virtual screening methods, which is publicly available and employs decoys whose inactivity is supported by experimental data.

The MUV approach specifically addresses the validation of ligand based virtual screening approaches. Therefore both components of benchmark data set bias, i.e. analogue bias and artificial enrichment, are minimized in the MUV data sets. With these properties, however, MUV data sets also fulfill the criteria postulated by Verdonk et al.[13] for the unbiased benchmarking of molecular docking methods. Three dimensional structures are available from the PDB for seven of the MUV targets (PKA, SF1, HIV RT-RNase, HSP90, FAK, Cathepsin G, FXIa). The respective data sets are readily applicable to the validation of SBVS methods. Furthermore, it can be safely expected that the number of MUV data sets with associated 3-D protein target structures will rise, given the rapid growth of both PubChem and the PDB. Hence, MUV is a collection of benchmark data sets that is equally unbiased for SBVS and LBVS methods. This might constitute an important progress toward comparing the performance of docking programs with ligand based VS techniques in an unbiased manner.

A workflow is presented that allows the generation of spatially optimized benchmark data sets from raw bioactivity data. As a special benefit, this workflow provides a data centered approach to detect HTS assay artifacts. The workflow is readily applicable to custom data sets of prospective users. Thereby users can generate MUV data sets customized to their specific virtual screening problems from their own in-house bioactivity data. Moreover, the filters implemented in the workflow can also be used to purge data sets for applications other than VS validation from potential unspecific binders.

The workflow is modular and easily extendable regarding two important aspects: (i) New bioactivity data sets deposited in PubChem can readily be fed into the workflow, filtered for assay artifacts and inadequately embedded actives, optimized spatially, and thus be integrated into the collection of MUV benchmark data sets. (ii) Because of the data centered nature of the assay artifacts filter, new experimental screening and profiling data for assay specificity can quickly be integrated. Thus, driven by the fast growth of screening data in PubChem, the MUV data set collection will continuously be extended by new targets and data sets. Much the same way, the efficiency of the assay artifacts filter will be considerably augmented by the rapidly accumulating knowledge available in the PubChem database.

The MUV collection of VS benchmark data sets and a MATLAB[83] toolbox for the spatial statistics analysis of chemical data sets are available for download from our Web site: http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html.

Abbreviations: %SI, percent sequence identity; AID, assay ID; CID, compound ID; D1 rec., dopamine D1 receptor; DUD, directory of useful decoys; Eph rec. A4, Eph receptor A4; ER-a-coact. bind., estrogen receptor alpha coactivator binding; ER-a-coact. bind., estrogen receptor beta coactivator binding; FAK, focal adhesion kinase; FoH, frequency of hits; FoM, figure of merit; FXIa, coagulation factor Xia; FXIIa, coagulation factor XIIa; HIV RT-RNase, human immunodeficieny virus reverse transcriptase RNase; HSP90, heat shock protein 90; HTS, high throughput screening; LBVS, ligand based virtual screening; M1 rec., M1 muscarinic receptor; MDDR, MDL Drug Data Report; MUV, maximum unbiased validation; PA, potential actives; PD, potential decoys; PDB, Protein Data Bank; PKA, protein kinase A; PPI, protein−protein interaction; PUG, PubChem Power User Gateway; ROC, receiver operating characteristic; RTR, retrieval rate; S1P1, sphingosine-1-phosphate receptor 1; S1P2, sphingosine-1-phosphate receptor 2; SBVS, structure based virtual screening; SF1, nuclear receptor steroidogenic factor 1; TAC, tested assay count; UID, unique ID; VS, virtual screening; wAAC, weighted active assay count; ρ, Spearman rank correlation coefficient.

ACKNOWLEDGMENT

We thank the anonymous referees whose fruitful comments helped to improve an earlier draft of this paper. Furthermore, we thank Dr. Christopher Austin and Dr. Douglas Auld, NIH Chemical Genomics Center, Rockville, MD for helpful comments regarding potential false negatives in PubChem HTS data. We also thank Almuth Kaune for invaluable help with the collection of bioactivity and literature data.

Supporting Information Available: Investigation of potentially false negative decoys in PubChem primary HTS assays and summary of the literature research (Table S1), investigation of potentially false negative decoys in PubChem primary HTS assays and found references (Table S2), occurrence of Lipinski's Rule of Five violations in the MUV data sets (Table S3), ΣS, ΣG, and ΣF for all MUV data sets measured in simple descriptor space (Table S4), VS performance (*mean*(*ROC*), *mean*(*RTR*)) and the respective standard

deviations for all retrospective VS simulations carried out (Table S5), and ΣS, ΣG, ΣF for all DUD data sets measured in simple descriptor space (Table S6). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Böhm, H. J.; Schneider, G. *Virtual Screening for Bioactive Molecules*; Wiley-VCH: Weinheim, 2000.
(2) Stahura, F. L.; Bajorath, J. New Methodologies for Ligand-Based Virtual Screening. *Curr. Pharm. Des.* **2005**, *11*, 1189–1202.
(3) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
(4) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.
(5) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (Molprint 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
(6) 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.
(7) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of Automated Docking Programs as Virtual Screening Tools. *J. Med. Chem.* **2005**, *48*, 962–976.
(8) Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2006**, *49*, 5856–5868.
(9) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.
(10) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(11) 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.
(12) MDL Drug Data Report (MDDR); Symyx Technologies, Inc.: Santa Clara, CA, 2005.
(13) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–806.
(14) 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.
(15) Good, A. C.; Hermsmeier, M. A. Measuring CAMD Technique Performance. 2. How "Druglike" Are Drugs? Implications of Random Test Set Selection Exemplified Using Druglikeness Classification Models. *J. Chem. Inf. Model.* **2007**, *47*, 110–114.
(16) Good, A. C.; Oprea, T. I. Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: A Help or Hindrance in Tool Selection. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.
(17) Clark, R.; Webster-Clark, D. Managing Bias in ROC Curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141–146.
(18) Rohrer, S. G.; Baumann, K. Impact of Benchmark Data Set Topology on the Validation of Virtual Screening Methods: Exploration and Quantification by Spatial Statistics. *J. Chem. Inf. Model.* **2008**, *48*, 704–718.
(19) Upton, G. J. G.; Fingleton, B. *Spatial Data Analysis by Example*; Wiley & Sons Ltd.: New York, NY, 1985.
(20) Fortin, M.-J.; Dale, M. R. T. *Spatial Analysis: A Guide for Ecologists*; Cambridge University Press: Cambridge, U.K., 2005.
(21) Holliday, J. D.; Jelfs, S. P.; Willett, P.; Gedeck, P. Calculation of Intersubstituent Similarity Using R-Group Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 406–411.
(22) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74–82.
(23) Reid, D.; Sadjad, B.; Zsoldos, Z.; Simon, A. LASSO-Ligand Activity by Surface Similarity Order: A New Tool for Ligand Based Virtual Screening. *J. Comput.-Aided Mol. Des.* **2008**, *6−7*, 479–487.
(24) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D Pharmacophore Descriptions for Scaffold Hopping. *J. Chem. Inf. Model.* **2006**, *46*, 208–220.
(25) Schulz-Gasch, T.; Stahl, M. Scoring Functions for Protein-Ligand Interactions: A Critical Perspective. *Drug Discovery Today: Technol.* **2004**, *1*, 231–239.
(26) National Center for Biotechnology Information (NCBI). Pubchem. http://pubchem.ncbi.nlm.nih.gov (accessed Feb 14, 2008).
(27) Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Edgar, R.; Federhen, S.; Feolo, M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Miller, V.; Ostell, J.; Pruitt, K. D.; Schuler, G. D.; Shumway, M.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R. L.; Tatusova, T. A.; Wagner, L.; Yaschenko, E. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2008**, *36*, D13–D21.
(28) National Institutes of Health (NIH). Molecular Libraries Initiative. http://mli.nih.gov/mli/ (accessed Feb 14, 2008).
(29) National Institutes of Health (NIH). NIH Roadmap for Medical Research. http://nihroadmap.nih.gov/molecularlibraries/ (accessed Feb 14, 2008).
(30) Zerhouni, E. Medicine. The NIH Roadmap. *Science* **2003**, *302*, 63–72.
(31) Hsieh, J.-H.; Wang, X.; Teotico, D.; Golbraikh, A.; Tropsha, A. Differentiation of AmpC Beta-Lactamase Binders vs. Decoys Using Classification kNN QSAR Modeling and Application of the QSAR Classifier to Virtual Screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 593–609.
(32) Schuler, G. D.; Epstein, J. A.; Ohkawa, H.; Kans, J. A. Entrez: Molecular Biology Database and Retrieval System. *Methods Enzymol.* **1996**, *266*, 141–162.
(33) Pubchem Power User Gateway (PUG). ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_pug.pdf (accessed Feb 25, 2008).
(34) Zhou, Y.; Zhou, B.; Chen, K.; Yan, S. F.; King, F. J.; Jiang, S.; Winzeler, E. A. Large-Scale Annotation of Small-Molecule Libraries Using Public Databases. *J. Chem. Inf. Model.* **2007**, *47*, 1386–1394.
(35) Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-Throughput Assays for Promiscuous Inhibitors. *Nat. Chem. Biol.* **2005**, *1*, 146–148.
(36) Shoichet, B. K. Screening in a Spirit Haunted World. *Drug Discovery Today* **2006**, *11*, 607–615.
(37) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
(38) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.-M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjögren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G. Development of a Virtual Screening Method for Identification Of "Frequent Hitters" In Compound Libraries. *J. Med. Chem.* **2002**, *45*, 137–42.
(39) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for The "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
(40) Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A High-Throughput Screen for Aggregation-Based Inhibition in a Large Compound Library. *J. Med. Chem.* **2007**, *50*, 2385–2390.
(41) Motulsky, H. *Analyzing Data with Graphpad Prism*; Graphpad Software Inc.: San Diego, CA, 1999.
(42) Walters, W. P.; Namchuk, M. Designing Screens: How to Make Your Hits a Hit. *Nat. Rev. Drug Discovery* **2003**, *2*, 259–266.
(43) *Graphpad Prism, 4*; GraphPad Software, Inc.: San Diego, CA, 2003.
(44) Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Kang, Z. B.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M. Understanding False Positives in Reporter Gene Assays: In Silico Chemogenomics Approaches to Prioritize Cell-Based HTS Data. *J. Chem. Inf. Model.* **2007**, *47*, 1319–1327.
(45) Pearce, B. C.; Sofia, M. J.; Good, A. C.; Drexler, D. M.; Stock, D. A. An Empirical Process for the Design of High-Throughput Screening Deck Filters. *J. Chem. Inf. Model.* **2006**, *46*, 1060–1068.
(46) Merkwirth, C.; Mauser, H.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble Methods for Classification in Cheminformatics. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1971–1978.
(47) Fan, J.; Gijbels, I. Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation. *J. Roy. Stat. Soc.* **1995**, *57*, 371–394.
(48) Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. Clustal W and Clustal X Version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948.
(49) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. Clustal W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Res.* **1994**, *22*, 4673–4680.

(50) Simeonov, A.; Jadhav, A.; Thomas, C. J.; Wang, Y.; Huang, R.; Southall, N. T.; Shinn, P.; Smith, J.; Austin, C. P.; Auld, D. S.; Inglese, J. Fluorescence Spectroscopic Profiling of Compound Libraries. *J. Med. Chem.* **2008**, *51*, 2363–2371.
(51) Auld, D. S.; Southall, N. T.; Jadhav, A.; Johnson, R. L.; Diller, D. J.; Simeonov, A.; Austin, C. P.; Inglese, J. Characterization of Chemical Libraries for Luciferase Inhibitory Activity. *J. Med. Chem.* **2008**, *51*, 2363–2371.
(52) Malo, N.; Hanley, J. A.; Cerquozzi, S.; Pelletier, J.; Nadon, R. Statistical Practice in High-Throughput Screening Data Analysis. *Nat. Biotechnol.* **2006**, *24*, 167–175.
(53) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative High-Throughput Screening: A Titration-Based Approach That Efficiently Identifies Biological Activities in Large Chemical Libraries. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11473–11478.
(54) *SciFinder Scholar*, *2007*; Chemical Abstracts Service: Columbus, OH, 2007.
(55) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. Drugbank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906.
(56) *Prous Drugs of the Future*; Prous Science: Philadelphia, PA, 2008.
(57) Sigma-Aldrich. Chemistry Product Catalog. http://www.sigmaaldrich-.com/homepage/Site_level_pages/CatalogHome/Chemistry_Catalog.html (accessed Mar 7, 2008).
(58) *CTFile Formats*; Symyx Technologies, Inc.: Santa Clara, CA, 2005.
(59) *3D Structure Generator CORINA: Generation of High-Quality Three-Dimensional Molecular Models*; Molecular Networks GmbH Computerchemie: Erlangen, Germany, 2006.
(60) *BABEL3, 2.2*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2006.
(61) *FILTER, 2.2.1*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2007.
(62) Baumann, K. An Alignment-Independent Versatile Structure Descriptor for QSAR and QSPR Based on the Distribution of Molecular Features. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 26–35.
(63) *Molecular Operating Environment (MOE), 2007.09*; Chemical Computing Group: Montreal, Canada, 2007.
(64) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using MDL "Keys" As Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.
(65) Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, NY, 2002.
(66) Mandel, J. Use of the Singular Value Decomposition in Regression Analysis. *Am. Statist.* **1982**, *36*, 15–24.
(67) Bellman, R. E. *Adaptive Control Processes: A Guided Tour*; Princeton University Press: Princeton, NJ, 1961.
(68) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137–148.
(69) Atkinson, A. C.; Donev, A. N. *Optimum Experimental Designs*; Oxford University Press: Oxford, U.K., 1992.
(70) Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181–1188.
(71) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J. Chem. Inf. Model.* **2001**, *41*, 1395–1406.
(72) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.
(73) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
(74) Vogt, M.; Bajorath, J. Introduction of an Information-Theoretic Method to Predict Recovery Rates of Active Compounds for Bayesian in Silico Screening: Theory and Screening Trials. *J. Chem. Inf. Model.* **2007**, *47*, 337–341.
(75) Schmuker, M.; Schneider, G. Processing and Classification of Chemical Data Inspired by Insect Olfaction. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 20285–20289.
(76) Edgington, E. S. *Randomization Tests*; Marcel Dekker, Inc.: New York, NY, 1980.
(77) Irwin, J. J.; Shoichet, B. K. ZINC-a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(78) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109–116.
(79) *QikProp 3.0*; Schrödinger, LLC: New York, NY, 2007.
(80) Shoichet Laboratory, Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). DUD - a Directory of Useful Decoys. http://dud.docking.org/ (accessed Oct 1, 2007).
(81) Liebeschuetz, J. Evaluating Docking Programs: Keeping the Playing Field Level. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 229–238.
(82) Irwin, J. Community Benchmarks for Virtual Screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193–199.
(83) *Matlab 7*; The Mathworks: Natick, MA, 2006.

CI8002649