Impact of Benchmark Data Set Topology on the Validation of Virtual Screening Methods: Exploration and Quantification by Spatial Statistics

Sebastian G. Rohrer and Knut Baumann*

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

Received March 20, 2007

A common finding of many reports evaluating ligand-based virtual screening methods is that validation results vary considerably with changing benchmark data sets. It is widely assumed that these data set specific effects are caused by the redundancy, self-similarity, and cluster structure inherent to those data sets. These phenomena manifest themselves in the data sets' representation in descriptor space, which is termed the data set *topology*. A methodology for the characterization of data set topology based on spatial statistics is introduced. The method is nonparametric and can deal with arbitrary distributions of descriptor values. With this methodology it is possible to associate differences in virtual screening performance on different data sets with differences in data set topology. Moreover, the better virtual screening performance of certain descriptors can be explained by their ability of representing the benchmark data sets by a more favorable topology. Finally it is shown, that the composition of some benchmark data sets causes topologies that lead to overoptimistic validation results even in very "simple" descriptor spaces. Spatial statistics analysis as proposed here facilitates the detection of such biased data sets and may provide a tool for the future design of unbiased benchmark data sets.

INTRODUCTION

Today, virtual screening (VS) is a standard technique used in almost every drug discovery campaign, both in industrial and academic environments. Its main target is to narrow down the huge chemical space to a level, where experimental scientists or automated systems can cope with testing the substances for biological activity.[1] Virtual screening methods can be roughly divided into structure-based and ligand-based approaches.[2] In structure-based VS, potential drug molecules are first docked into the three-dimensional structure of the given target and then ranked according to the estimated binding affinity of the predicted complex.[3] Ligand-based VS utilizes the knowledge about one or several substances active against the target. Libraries of potential drugs are ranked by their similarity to the query substances following one of medicinal chemistry's most fundamental paradigms: chemically similar substances are likely to share similar biological activity.[4–6] This study will focus on ligand-based virtual screening.

A prerequisite for conducting ligand-based VS is a set of molecules, that exhibit the desired activity, encoded by a numerical structure descriptor.[2] A large variety of such descriptors with different levels of sophistication and complexity has been developed.[2] Subsequent to the computation of descriptors, chemical similarity is quantified by applying one of several similarity or distance measures to the descriptor vectors of two molecules. The most widely used similarity measures are the "Tanimoto coefficient" for binary vectors and the Euclidean distance for continuous representations of molecules, respectively.[6] A consequence of this quantification of similarity in descriptor space is the need for an estimate of the numerical cutoff value of similarity that is sufficient to ensure similar biological activity. Martin et al. have shown in their much-cited paper that it is not possible to define an explicit numerical similarity cutoff in descriptor space that ensures common biological activity across target classes.[5]

In two earlier papers, Brown and Martin assessed the ability of various descriptors to discriminate actives from inactives[7] and to predict physicochemical properties relevant to receptor binding[8] using data sets of molecules for which these parameters were experimentally determined. This validation of VS methods and descriptors on benchmark data sets has become a standard methodology, because it facilitates two tasks deemed critical for prospective virtual screening campaigns: (i) from the plethora of available methods, to determine the one most apt to the particular VS problem and (ii) to estimate the number of expected hits.

A validation procedure normally starts with the selection of a set of molecules with known activity against the target under scrutiny. Part of this set is chosen (often randomly) to act as query. The rest is pooled with a usually large number of molecules presumed to be inactive (frequently called "background") to become the validation set. Both sets, the query and the validation set, are then encoded by the descriptor to be validated and the validation set is ranked according to its similarity with the query. A common way of visualizing the retrieval of known active substances in a VS ranking is to plot the fraction of found actives against the fraction of the ranking containing it. These plots are frequently called "enrichment plots". The recall or retrieval rate (RTR) at one percent of the ranked validation set has been established as a widely used figure of merit for the

* Phone: +49-531-3912751. Fax: +49-531-3912799. E-mail: k.baumann@tu-braunschweig.de.

10.1021/ci700099u CCC: $40.75
Published on Web 04/02/2008

performance of a given VS method.[9,10] It is evident, that the RTR fulfills both goals of VS validation as stated above: comparability of methods and estimation of the expected hit number. Often so-called "enrichment factors" (EF)[11] are calculated from the RTR, that are meant to normalize to the null hypothesis of uniformly distributed active molecules in the final ranking list. Bender and Glen have shown that enrichment factors tend to overestimate performance, since the assumption of a random ranking of the validation set as a baseline is not realistic.[12] Moreover, several authors have criticized the RTR and EF metrics for the fact of being susceptible to changes in the ratio of the sizes of benchmark data set vs background and the inability to reflect the position of the found actives before the threshold.[13,14] In order to avoid these issues, the area under the receiver operating characteristic curve (ROC) was used in a number of studies for the analysis of VS validation rankings.[15–18] However, the ROC metric has one important shortcoming: it is unable to address the so-called "early recognition" problem.[13,14] Since usually only a small fraction of a VS ranking can be tested experimentally, a good metric for VS should reflect the enrichment of actives at the beginning of the ranking. Recent efforts have sought to develop VS performance metrics that combine the statistic stability of ROC with the early recognition properties of RTR or EF.[13,14] However, due to their novelty, these metrics have not yet found extensive use in VS validation studies and it is therefore difficult to compare the results obtained with these metrics to those of other works. Thus, in this study, RTR and ROC were used in a complementary manner for the analysis of VS rankings.

A basic condition for the comparability of results of VS validation is the availability of commonly employed benchmark data sets. In two seminal papers Hert and Willet provided 11 data sets extracted from the MDL Drug Data Report (MDDR),[19] each consisting of several hundreds of substances, with known activities against a range of therapeutically relevant targets.[9,10] These data sets have become widely accepted for the validation of virtual screening methods[12,20–22] and will also be used in this study.

Although the validation and calibration of VS methods on real data sets is now an established methodology,[23–26] a number of problems arise from its empirical nature.

In a study on the evaluation of the docking program GOLD,[27,28] Verdonk et al. have shown that results are highly influenced by the composition of the data set of inactive molecules.[29] They proved that if the background significantly differs from the set of actives regarding "low dimensional" 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 lower dimensions. They conclude that focusing the library of inactives to the same range of low dimensional 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 molecular docking fulfilling these requirements, has become available for public use.[30]

As mentioned above, Bender and Glen showed that the random ranking hypothesis underlying the calculation of enrichment factors often leads to overoptimistic estimations of a method's performance. They suggest normalizing the RTR of any VS method by the RTR of a so-called "dumb" descriptor like molecular weight or atom counts, basically a vectorized form of the chemical sum formula.[12] Since the sum formula is highly correlated with molecular weight and hydrogen bond donor/acceptor counts, this approach implicitly covers many of the phenomena discussed by Verdonk et al. and can effectively be used as a negative control when validating VS methods utilizing similarity searching.

While the works by Verdonk et al. and Bender et al. precisely highlight the effect the composition of the background data set has on the outcome of VS validation, two papers by Good and co-workers[31,32] concluded that validation sets extracted from databases constructed from drug discovery projects, such as the MDDR,[19] are prone to overrepresentation of certain scaffolds or chemical entities. It was shown, that unless the benchmarking data sets and the background are chosen with care, the figures of merit for ligand based virtual screening may be overoptimistic due to the so-called "analogue bias".[31,32] Recently, Vogt and Bajorath[33,34] proposed a measure of divergence between the descriptor distributions of the benchmark data set and the background that relates to VS performance and can thus be used to estimate performance rates in descriptor-based virtual screening. Their approach is based on a number of assumptions about the descriptor distribution. As opposed to this, the method presented here is nonparametric, i.e. it can deal with arbitrary distributions of benchmark data set and background. The target of this study is to complement the findings of Verdonk et al.,[29] Bender et al.,[12] and Good et al.[31,32] on the validation of virtual screening methods by quantitative data.

When comparing the benchmark data sets proposed by Hert et al.[9,10] with data sets of actives that form the input of real-life virtual screening campaigns, the most striking feature is their size. Whereas the benchmark data sets consist of several hundreds to more than thousand substances, usually only a small number (~10–20) of active substances are available at the beginning of a VS campaign, rendering the benchmark data sets redundant with respect to real-life VS. Furthermore, the potential presence of large scaffold families in the data sets introduces analogue bias, as stated above. These phenomena and their variations across data sets manifest themselves as spread, self-similarity, patchiness and clustering of the data sets' representation in descriptor space. We will refer to this mapping of data set composition in descriptor space as the data set *topology*. Both, the calculation of molecular similarity and the subsequent ranking of the screened database, which are the most crucial steps in the conduction of virtual screening, are based on the respective molecules' representation by descriptors. It is therefore reasonable to quantify analogue bias, redundancy and other phenomena influencing VS performance by their representation in descriptor space, i.e. data set topology. The field of spatial statistics offers a broad array of methods for the analysis of mapped point patterns, which are well established in sciences as diverse as astronomy, geography and biology and have been extensively reviewed.[35] For our purpose, refined nearest neighbor analysis[36] proved especially useful, because a distribution statistic of intraset nearest neighbor distances is augmented by information about the empty space intervening clusters of actives. With this methodology, it is possible to summarize complex properties of a data set, e.g.

Table 1. Benchmark Data Sets of Substances with Activity against the Specified Targets Extracted from the MDDR

activity	data set size
angiotensin converting enzyme (ACE) inhibitors	355
acetylcholine esterase (AChE) inhibitors	701
angiotensin II type 1 receptor blockers	943
cyclooxygenase (COX) inhibitors	636
D2 antagonists	395
HIV protease inhibitors	750
5HT1A agonists	827
5HT3 antagonists	752
5HT reuptake inhibitors	359
protein kinase C (PKC) inhibitors	452
renin inhibitors	1130
substance P inhibitors	1246
thrombin inhibitors	803

the number of clusters, their respective density and the distances between clusters, collectively by a single scalar.

The basic idea of this paper is to sample subsets with different topologies from available benchmark data sets. By comparing the results of retrospective VS simulations on these subsamples, the influence of data set topology on the validation results can be observed.

Various sampling strategies were employed to generate archetypal subsamples from the literature benchmark data sets: (1) maximum diversity subsets, (2) space filling samples, and (3) subsets with minimum intraset diversity. The analysis of the varying VS performance on these prototype data sets allowed us to assess if and to what extent data set topology affects the validation of virtual screening. Here, it is essential to ensure that no other factors influence the outcome of the VS runs, which was achieved by a combination of careful experimental setup and bootstrapping methods. Spatial statistics techniques are introduced here to explore and quantify the effect of benchmark data set topology in more detail. In order to get an idea how encoding by different descriptors influences data set topology in descriptor space, two kinds of descriptors were used for the experiments: molecular operating environment (MOE)[37] molecular descriptors and a "simple" descriptor acting as a negative control. After using artificial subsamples to show that data set topology has an effect on the outcome of VS validation and that it can be quantified by spatial statistics methods, the methodology is applied to the complete benchmark data sets.

METHODS

Compilation and Preparation of Benchmark Data Sets. The data sets as described by Hert et al.[9,10] and two additional data sets containing inhibitors of angiotensin converting enzyme (ACE) and acetylcholineesterase (AChE) were extracted from the MDDR (MDL Drug Data Report)[19] by their activity indices. An overview of the data sets is provided by Table 1. The remaining 93925 molecules in the MDDR that did not belong to at least one of the activity classes, were assumed to be inactive and were used as the screening background as described by Hert et al.[9,10] This excludes potential overlap between the sets of actives as a factor distorting validation results. SD-Files[38] of all data sets and the background were cleaned from small fragments like counterions and solvents using MOE (molecular operating environment)[37] and converted to three-dimensional structures with CORINA.[39] For all data sets and the background, MOE molecular properties descriptors[37] were computed. Properties

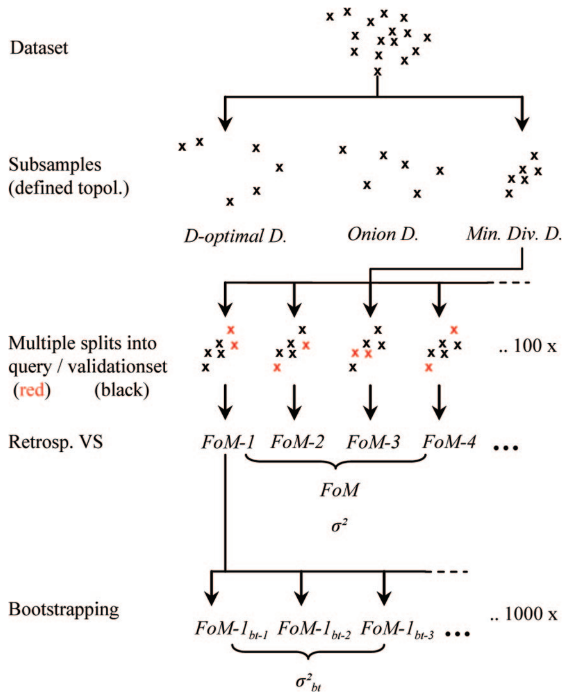
Dataset
Subsamples (defined topol.)
D-optimal D.
Onion D.
Min. Div. D.
Multiple splits into query / validationset (red) (black)
.. 100 x
Retrosp. VS
FoM-1 FoM-2 FoM-3 FoM-4 ...
FoM
σ^2
Bootstrapping
.. 1000 x
$FoM\text{-}1_{bt\text{-}1}$ $FoM\text{-}1_{bt\text{-}2}$ $FoM\text{-}1_{bt\text{-}3}$ ...
σ^2_{bt}

Figure 1. Schematic outline of the experimental setup. Subsamples with controlled topology are extracted from literature benchmark data sets. One hundred random splits of the subsample into query and validation set are generated for retrospective VS. The arithmetic mean and the variance of the figure of merit (FoM) under scrutiny over all splits estimate the VS performance on the subsample and its statistical error, respectively. Bootstrapping is used to determine to what extent this statistical error is caused by factors other than topology.

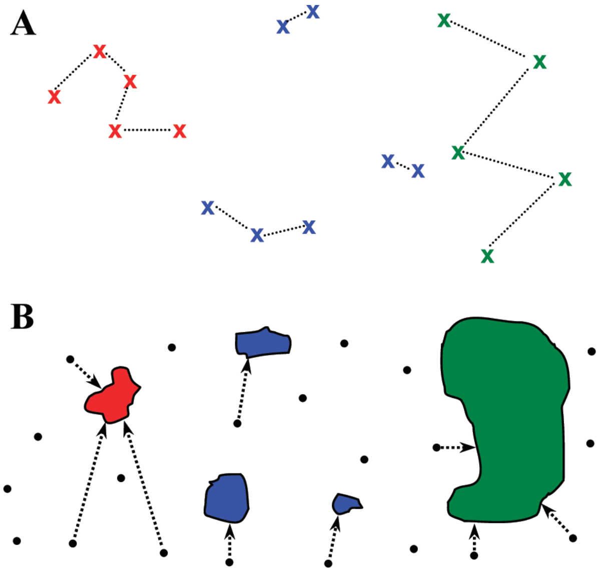

Figure 2. Representative data sets and their characterization by refined nearest neighbor analysis. (A) Nearest neighbor distances in concentrated (red) and patchy (blue) sets are smaller than those in dispersed sets (green). However, patchy and concentrated sets can not be distinguished by nearest neighbor distances, since these are not affected by the presence of multiple clusters. (B) By "flooding" the analyzed space with random points and measuring their distances to the nearest event, gaps in the data can be detected and the overall spread of the sets can be quantified, thereby differentiating patchy from concentrated sets.

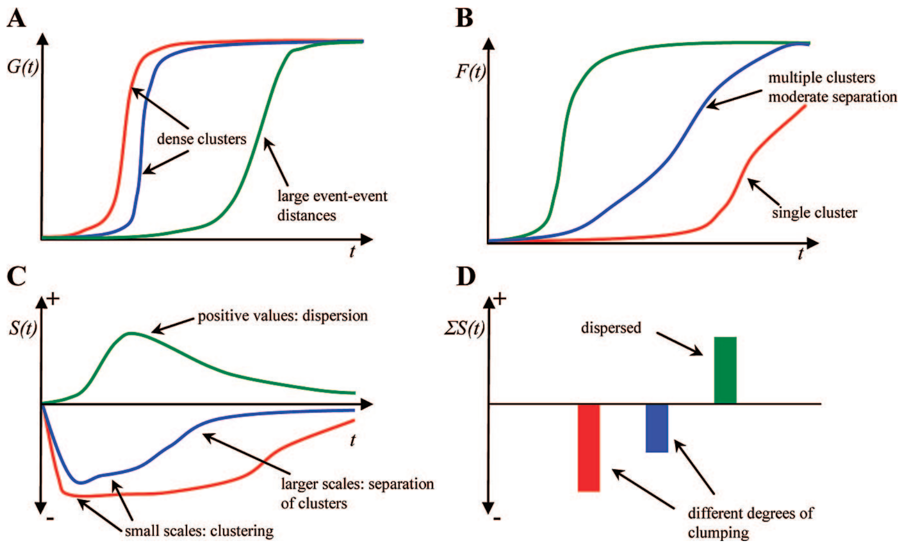

Figure 3. Exemplary graphs of refined nearest neighbor analysis functions. (A) Graphs of $G(t)$ for both concentrated (red) and patchy (blue) sets feature a steep ascent at small values of t, caused by clusters of high density. Dispersed sets (green) can easily be recognized. (B) Patchy sets (blue) can be distinguished from concentrated ones (red) by the earlier rise in $F(t)$. For large separation between the clusters, the graph for patchy sets would converge to the graph for dispersed ones (green). (C) Using $S(t)$ the topology of sets can be characterized unambiguously. Whereas dispersed sets (green) are marked by positive values for $S(t)$, different types of clustering exhibit characteristic curves in the negative region. (D) ΣS provides an easily interpretable estimate of the degree of clumping or dispersion of a set.

are grouped into three classes: 2D (computed from the two-dimensional topology of a molecule), i3D (properties that depend on internal three-dimensional coordinates), and x3D (calculated from a grid surrounding the three-dimensional structure of the molecule). Since the x3D class depends on a common spatial frame of reference for all molecules, which is not feasible for VS applications, only the 2D and i3D classes were used for experiments. The numerical values of properties in MOE descriptors have significantly different ranges. Molecular weight for instance typically varies between 0 and ~1000, whereas logP often has a value roughly between −1 and 5 for drug-like molecules. Therefore the data matrix consisting of MOE descriptor vectors for the complete MDDR was autoscaled columnwise by subtracting the mean and dividing by the standard deviation of each column. Columns whose properties had constant values for the complete MDDR were removed from the matrix. After this pretreatment, the matrix had a dimensionality of 180. In order to reduce noise in the descriptor matrix, principal components analysis (PCA)[40] was applied. An analysis of the resulting eigenvalues showed that >99% of the total variance could be explained by the first 54 components. Thus the first 54 scores from the PCA were used as the final descriptors for the VS simulations.

Following the approach suggested by Bender and Glen,[12] the database was also encoded by a negative control simple descriptor, which consisted 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 log P, the number of chiral centers and the number of ring systems. These properties were generated from the SD-Files[38] of the data sets using OpenEye BABEL3[41] (atom counts) and OpenEye FILTER[42] (acceptors, donors, logP, chiral centers, ring systems). The respective output files were parsed with an in-house PERL script to provide the final descriptors. For the sake of brevity, this representation of molecules will be referred to as simple descriptors from now on in the text.

Generation of Subsets with Defined Topologies. Various methods exist to generate subsets of substances based on a descriptor representation of the original data set. Based on the MOE-PCA representation of the data, from each of the benchmark data sets a subset of k = [50, 100, 150, 200, 250, 300] substances was generated for each of the three following sampling strategies. *D-optimal design* (DOD)[43,44] was used to provide subsets with the maximum intraset diversity for the respective number of substances. For the generation of subsets sampling the respective data set in a space filling manner, *D-optimal onion design* (OD)[45] was applied. Around the center of mass of each data set 5 shells were defined, of which each contained 20% of the data. In contrary to shells of equal distance, this approach ensures the presence of an adequate number of datapoints in each shell in spaces of high dimensionality.[46,47] In order to reflect the data sets' density distribution in the subsamples, an equal number of $k/5$ substances were chosen from each shell by the D-optimality criterion. Both, D-optimal design and D-optimal onion design were implemented using the statistics toolbox of Mathworks Matlab 7.[48] Subsets with a minimum sum of intraset all against all distances, i.e. by a *minimum diversity design* (MDD), were generated using an in-house row exchange algorithm[49] also implemented in Matlab 7. All three sampling strategies are deterministic, i.e. for a given k one optimum subsample from the original data set is selected. By this procedure three prototypes of subsamples were created for every k: (i) A "worst-case scenario" for which VS using MOE-PCA descriptors would be very difficult was generated by the maximum diversity criterion using D-optimal design. (ii) An intermediate case with active compounds equally spread across MOE-PCA descriptor space with varying density depending on k was generated by the onion design approach. (iii) Finally, a "best-case

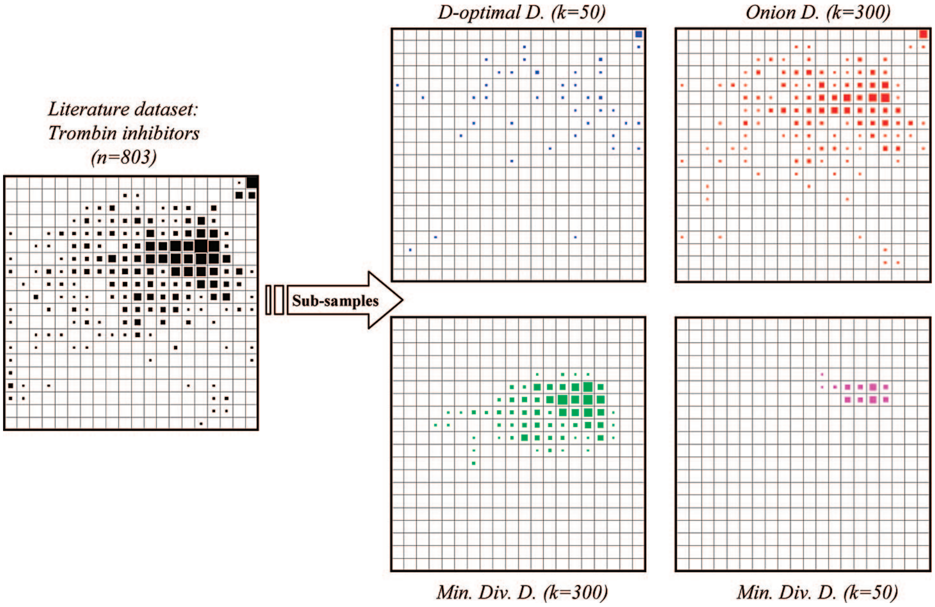

Figure 4. Visualization of the topology of subsamples from the data set of thrombin inhibitors using self-organizing maps (SOMs). Subsamples are generated from the original data set (black). The topology of the subsamples varies according to sampling strategy and sample size (k). D-optimal design with small k (blue) generates a set with the maximum degree of dispersion. Onion design with large k (red) results in moderately patchy data sets with comprehensive coverage of the original data set. Minimum diversity design with large (green) or small (magenta) k generates strongly patchy or concentrated data sets, respectively.

scenario" with multiple active substances concentrated in a tight cluster resulted from the minimum diversity design.

In order to observe how data set topology and VS performance are affected by different descriptor representations, the identical subsamples, i.e. the same substances chosen by the different design strategies for MOE-PCA descriptors, were compiled for the simple-descriptor representation. This excluded subsample composition as a factor of variance when comparing the performance of different descriptor representations.

Retrospective Virtual Screening Simulations. A variety of methods is available for the conduction of ligand based virtual screening. It was not the goal of this study to figure out the descriptor, similarity measure, or searching method that presumably generates best results for virtual screening, but to determine the influence of data set topology on method performance. Therefore we settled for one type of searching procedure and one similarity measure, which we kept constant for all experiments.

For similarity searching with multiple query molecules, data fusion according to the "MAX"-rule has proven to be very powerful, both in our own experience with the conduction of virtual screenings and in comparative studies in the literature.[9,10] The similarity of each molecule in the screened database (here the validation set) with each molecule in the query is calculated, and the maximum of these values of similarity—i.e. the nearest neighbor similarity—is used to rank-order the molecules in the search output.

All simulations of virtual screening described in this paper were carried out using ten active substances chosen randomly from the subsamples described above as query and pooling the remaining actives with the background. Similarity was measured by Euclidean distance and the respective validation sets were ranked accordingly. This was repeated 100 times to assess the variability of the results and to obtain a mean value that is not affected by the random choice of the query molecules. To obtain an estimate of the statistical error of the respective rankings, 1000 bootstrapping runs were carried out on the ranking resulting from each of these query/validation set combinations, randomly leaving out 20% of both the actives and the background substances in each run, respectively.[13]

Figures of Merit (FoM) for Virtual Screening Performance. The ability for early recognition of active substances was measured by the mean fraction of retrieved actives (retrieval rate, RTR) in the first percent of the ranked validation set. Additionally, the area under the receiver operating characteristic curve (ROC) was determined for all rankings in order to rule out any bias introduced by the hard 1% cutoff of the RTR. The mean retrieval rates and mean areas under the receiver operating characteristic curves obtained from the 100 random query/validation set splits, which were generated for each data set subsample, will be denoted mean(RTR) and mean(ROC) throughout the text.

Variance Decomposition for Figure of Merit Statistical Errors. The standard deviation of both RTR and ROC is often used to estimate the statistical error of VS validation results.[9,10,13,20,25] In the context of this study, it is desirable to isolate the component of error that is caused by the topology of the data set used in the respective validation experiment.

In their recent paper, Truchon and Bayly[13] not only introduced a new metric for the validation of VS methods, but also provided a remarkably concise and comprehensive analysis of the factors influencing the variance of VS rankings and the resulting figures of merit. According to them, the variance is mainly influenced by the following parameters: N, the absolute number of background molecules, and R_a, the fraction of actives in the ranking (i.e., benchmark data set size). Furthermore, the "goodness" of a ranking itself (denoted λ by Truchon and Baily) has considerable impact on the variance of the figures of merit. This is caused by the fact, that for a very "successful" screening run the actives are far less spread over the ranking, since they are concentrated at its beginning. Furthermore, they discuss a "saturation effect" that affects metrics measuring early recognition if the number of active substances is higher than the number of ranks in the part of the ranking that is considered "early".

In our setting, the number of background molecules was kept constant ($N = 93925$) for all experiments and can therefore be ruled out as a factor affecting the results of the VS simulations. Since the size of the biggest subsets of actives used here is 300, which is clearly smaller than the number of ranks in the first percent of the rankings (~940), the RTRs reported here are not subject to the saturation effect. The ROC metric has no early recognition features and consequently is not prone to the respective saturation effects.

There are, however, two parameters that can not be kept constant in this experimental setting: λ and R_a. The parameter λ is used by Truchon and Bayly to denote rankings with varying VS performance. However, the rankings discussed in their paper are derived from sampling a model probability density function rather than real VS runs. In our setting, the goodness λ of the ranking can not be determined before the experiment. For R_a, the sampling strategy for the design of subsets with different topology requires subsets of actives with differing sizes.

The magnitude of the variance component introduced by changing R_a and λ can be estimated by the bootstrapping procedure described above. As indicated in Figure 1, 1000 bootstrap samples were drawn from each ranking obtained from a particular query/validation set split. The variance ${\sigma_{\mathrm{bt},i}}^2$, with $i = 1 \ldots 100$, is an estimate of the statistical error caused by the particular combination of λ and R_a for each of the 100 splits. According to the law of total variance for experiments with a nested design,[50] the overall variance σ^2 can be decomposed to yield a variance component ${\sigma_{\mathrm{top}}}^2$ associated with data set topology:

$$\sigma^2 = {\sigma_{\mathrm{top}}}^2 + \mathrm{mean}({\sigma_{\mathrm{bt},i}}^2) \tag{1}$$

which can be written as

$${\sigma_{\mathrm{top}}}^2 = \sigma^2 - \mathrm{mean}({\sigma_{\mathrm{bt},i}}^2) \tag{2}$$

With all other factors constant, the only factor affecting ${\sigma_{\mathrm{top}}}^2$ is the data set topology. A corrected standard deviation providing an estimate of the statistical error of the validation results associated with data set topology can then be calculated as follows:

$$\mathrm{std}_{\mathrm{top}}(\mathrm{FoM}) = \sqrt{{\sigma_{\mathrm{top}}}^2} \tag{3}$$

Characterization of Data Set Topology by Spatial Statistics. The concept of chemical space is well established in chemoinformatics research.[51–53] In this study, chemical space is defined as the part of the descriptor space occupied by the representation of the complete MDDR, i.e. all sets of actives and the background. Accordingly, the positions of the compounds in the benchmark data sets mapped to the respective descriptor space form the basis of the analysis. In the terminology of spatial statistics, the position of a substance belonging to the sample under examination is an "event". On the other hand, "points" denote arbitrary coordinates in chemical space. Regarding virtual screening, there are three basic categories of topology for a set of events (Figure 2A). (i) "Concentrated" sets consist of a single dense cluster, well separated from the rest of chemical space. (ii) "Patchy" sets are composed of multiple dense but separated clusters. (iii) "Dispersed" sets are regularly distributed in chemical space, with comparatively large event–event distances. It should, however, be kept in mind that this rather coarse categorization of data set topologies is mainly used here for the illustration of the basic properties of the spatial statistics functions presented below. Real chemical data sets will usually incorporate all kinds of nuances and combinations of the three basic categories of data set topology.

Refined nearest neighbor analysis is a mathematical framework for the analysis of mapped point patterns. It is based on the calculation of two functions from the position of points and events (Figure 2):[36]

$G(t)$ is the proportion of events for which the distance to the nearest neighbor is less than t. $G(t)$ is called the "nearest neighbor function" and is a cumulative probability distribution of the distance of any event to its nearest neighbor event. Using $G(t)$, it is possible to distinguish dispersed from concentrated and patchy sets (green vs blue and red in Figure 2A). In some cases, it is however not possible to differentiate between patchy and concentrated sets (blue vs red in Figure 2A). Since only the nearest neighbor of each event is considered, the spacing between several dense clusters has no effect on $G(t)$, because the nearest neighbor of any event will always be located in the same cluster. As a consequence, $G(t)$ is neither sensitive to the presence nor to the spacing of multiple clusters. This distinguishes the nearest neighbor function from approaches based on average intraset distances[9,10,54] or methods based on more than one neighbor.[55,56]

Let n be the number of events, then $G(t)$ is given as follows:

$$G(t) = \frac{\sum_i I_t(i, j)}{n}; \quad i = 1 \ldots n \tag{4}$$

with $I_t(i, j) = 1$ if the distance of event i to its nearest neighbor j is smaller than t. Representative graphs of $G(t)$ for concentrated, patchy and dispersed sets are shown in Figure 3A.

In order to distinguish between concentrated and patchy sets, a large number of points are sampled randomly from chemical space. $F(t)$ is the proportion of these points for which the distance to the nearest event is less than t (Figures 2B and 3B). $F(t)$ is a cumulative probability distribution of the distance from a randomly chosen point to the nearest event and is often called the "empty space function", because it is sensitive to gaps in the data.[36] For a patchy set, the average distance from a random point to the nearest event will be smaller than for a concentrated set. Depending on the number and the degree of separation between the clusters less chemical space is unoccupied for patchy sets. On the other hand, because $F(t)$ does not take into account event–event distances, it can not differentiate between dispersed and patchy sets, if the clusters in the latter are far apart. Figure 3B provides representative graphs of $F(t)$ for concentrated, patchy and dispersed sets.

Let m be the number of random points, then $F(t)$ is given as follows:

$$F(t) = \frac{\sum_j I_t(j, i)}{m}; \quad j = 1 \ldots m \tag{5}$$

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

Incorporating the information of both functions into one equation, it is possible to differentiate and quantify all types of set topology:

$$S(t) = F(t) - G(t) \tag{6}$$

Figure 3C shows typical graphs of $S(t)$ for the major topology types. By summing up the values of $S(t)$ over a range of distances t, the clustering behavior of the set can be estimated by a single scalar:

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

where t_i represents a series of distances.[57] The scalar value of ΣS provides a quick and easily interpretable estimate of a set's clumping or dispersion, with values of $\Sigma S < 0$ indicating clumping and values of $\Sigma S > 0$ indicating dispersion. Most real-life chemical sets can not be strictly assigned to one of the basic categories of topology, but differ in the number of clusters, their respective density and scaling. This is perfectly reflected by the scalar ΣS, which provides a quantitative measure for the degree of clumping in a sample.

Implementing $G(t)$ for this study was straightforward. For any given subsample, the distance of each event (substance) to its nearest neighbor event was determined and $G(t)$ was calculated according to eq 4 for $t = [0.01, 0.02, \ldots 12]$. The range of values for t depends on the particular application of refined nearest neighbor analysis and must be determined empirically. Preliminary experiments identified $t = [0.01, 0.02, \ldots 12]$ as the best choice for the analyses presented here.

The implementation of $F(t)$ however faces the problem, that the descriptor spaces used in virtual screening are of rather high dimensionality, in our case 54 for MOE-PCA and 17 for the simple descriptors. High dimensional spaces are subject to the "empty space phenomenon", i.e. they are inherently sparsely populated.[47] As a consequence, if $F(t)$ was calculated based on random points uniformly distributed in such a high dimensional space, its value would be dominated by the empty space inherent to the dimensionality, not by the gaps in the data. Thus any sampling of points must ensure that these points lie in the portion of chemical space actually populated by substances but nevertheless provide representative coverage of chemical space (here defined by the MDDR database). This was achieved using three approaches for the generation of sets of random points. (i) Bootstrapping from the complete MDDR: 10 000 substances were randomly chosen from the union of the background and all benchmark data sets. (ii) Boostrapping from the background: 10 000 substances were chosen by random from the set of inactives. (iii) Convex pseudodata: Following the approach described by Breiman et al.,[58] 10 000 pseudodata points were generated from 20 000 substances chosen randomly from the MDDR. Briefly, from two datapoints x_1, x_2 a new pseudodatapoint x_3 is generated by selecting a random number v from the interval [0,1]. Then x_3 is given by the linear combination:

$$x_3 = vx_1 + (1 - v)x_2 \tag{8}$$

Thereby, artificial pseudodatapoints are created that occupy the same region of chemical space as the original population of datapoints.

For each subsample under scrutiny, this was repeated 20 times and $F(t)$ was calculated using eq 5 with $t = [0.01, 0.02, \ldots 12]$ for all samples of random points. The final value of $F(t)$ was determined as the arithmetic mean for all 20 sets of random points for each sampling method. Results for $F(t)$ were equal within the margin of statistical error for all methods of random point sampling. Therefore, all results will be reported for $F(t)$ with bootstrapping from the complete MDDR. An additional advantage of this procedure is that there is no need for terms of edge correction in $F(t)$ and $G(t)$, which are necessary in traditional spatial statistics applications on confined, two-dimensional maps. From $G(t)$ and $F(t)$, $S(t)$, and ΣS resulted as given by eqs 6 and 7.

Self-Organizing Maps (SOMs) as a Visual Support for Refined Nearest Neighbor Analysis. Self-organizing maps (SOMs)[59] are a special type of artificial neural networks, that project high-dimensional data onto a two-dimensional lattice while preserving the topology of the input data space.[60] It is obvious, that this feature makes SOMs the ideal tool for the visual perception of data set topology. Comprehensive reviews of the SOM-algorithm and its variations exist in the literature.[59–62]

All SOMs used here were generated and analyzed using the SOM-Toolbox 2.0[63] in Mathworks Matlab 7.[48] Map topology was chosen to be nontoroidal with a rectangular lattice. Since SOMs were only used for visualization supporting a more detailed analysis of data set topology by refined nearest neighbor analysis, a map grid of 20 × 20 units was considered sufficient. For training, the batch algorithm as implemented in the SOM-Toolbox was used.

RESULTS AND DISCUSSION

Characterization of Benchmark Data Set Subsamples by Refined Nearest-Neighbor Analysis. Using the sampling procedure described above, six subsamples of different size were generated for every sampling strategy from any of the benchmark data sets. Doing so, a total of 234 (6 sample sizes × 3 sampling strategies × 13 data sets) subsamples were generated featuring all types and nuances of topology, ranging from dispersed (D-optimal design) and patchy subsamples (onion design) to concentrated ones (minimum diversity design). For all samples, topology was characterized using refined nearest-neighbor analysis. The differences in subsample topology were well reflected by $G(t)$, $F(t)$, $S(t)$, and ΣS. The degree of clumping for each subsample as quantified by ΣS is shown in Table 2 (in the Supporting Information). An example for the visualization of subsample topologies and the results of refined nearest neighbor analysis for four subsamples from the data set of Thrombin inhibitors is provided by Figures 4 and 5, respectively.

Highly Correlated VS Performance and Data Set Clumping. As described above, retrospective VS validation experiments were carried out for 100 query/validation set splits of each subsample using MOE-PCA and simple descriptors. The resulting figures of merit as well as their

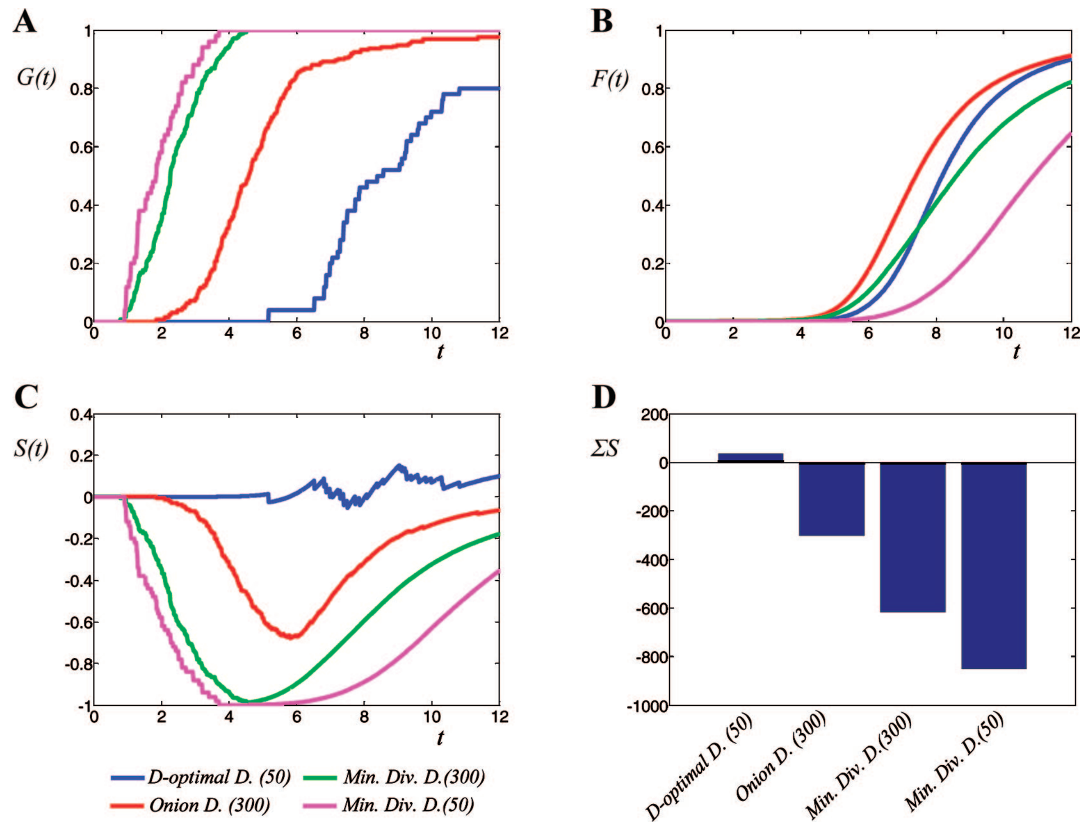

Figure 5. Topology analysis of four subsamples from the data set of thrombin inhibitors. (A) Data sets from the minimum diversity design (green, magenta) show the early and steep ascent in $G(t)$ characteristic for data sets with dense clumping. (B) The larger portion of chemical space occupied by the onion design subsample (red) is indicated by the earlier and steep ascent in $F(t)$ as opposed to the concentrated subsample (magenta) from the minimum distance design. It is difficult to differentiate patchy (green) from dispersed (blue) data sets by $F(t)$ alone. (C) $S(t)$ facilitates an unambiguous characterization of subsample topology. The subsamples are identified as dispersed (blue), moderately patchy (red), patchy with small separation (green), and concentrated (magenta). (D) ΣS reflects the dispersion of the D-optimal sample and the varying degree of clumping in the other samples.

variances and standard deviations (mean(RTR), mean(ROC), σ^2, σ_{top}^2, $\sigma_{\mathrm{bt},i}^2$, std$_{\mathrm{top}}$(FoM)) are given in Table 3 (in the Supporting Information). Although mean(RTR) and mean(ROC) are numerically different representations of VS performance, they agreed well in the relative rating of VS rankings for our experiments. In all the VS runs conducted here, there was no case where performance was rated high by mean(RTR) and intermediate or low by mean(ROC) and *vice versa*. The correlation coefficient of both figures of merit was ρ(mean(RTR), mean(ROC)) = 0.92 for MOE-PCA and ρ(mean(RTR), mean(ROC)) = 0.85 for simple descriptors. All figures of statistical error were generally higher for RTR than for ROC. On the other hand, the discriminatory power of the RTR was found to be higher, particularly for the upper and lower ends of the VS performance spectrum. This kind of behavior is known and expected for these figures of merit and highlights again the advantages of their complementary use.

Analyzing the degree of clumping ΣS for each subsample as shown in Table 2 and the figures of merit for VS performance obtained on the respective subsample (Table 3), a strong correlation was detected. In general, a higher degree of clumping (indicated by large negative values of ΣS) accounts for better VS performance as measured by mean(RTR) and mean(ROC), for both MOE-PCA and simple descriptors. The overall correlation coefficients were found to be $\rho(\Sigma S$, mean(RTR)) = −0.93 and $\rho(\Sigma S$, mean(ROC)) = −0.91 for MOE-PCA and $\rho(\Sigma S$, mean(RTR)) = −0.89 and $\rho(\Sigma S$, mean(ROC)) = −0.96 for simple descriptors, respectively. All correlation coefficients reported here and throughout the text were calculated as Spearman rank correlation coefficients (also see Table 4). A summary of the observed correlation coefficients is given by Table 4, and an example for the relation of VS performance and subsample clumping is shown in Figure 6.

Table 4. Correlation Coefficients of VS Performance Figures of Merit with Data Set Clumping (ΣS)

	$\rho(\Sigma S$, mean(RTR))[a]	$\rho(\Sigma S$, mean(ROC))
MOE-PCA	−0.93	−0.91
simple descriptors	−0.89	−0.96

[a] ρ: Spearman rank correlation coefficients. It should be noted that the numerical differences to the commonly used Pearson correlation coefficient are marginal. However, since the relationship between ΣS and VS performance after the rank transformation is approximately linear, which is not the case in the original data domain, Spearman rank correlation coefficients are preferred.

The component of variance introduced by topology σ_{top}^2 was found to be about 15 times larger (average over all subsamples) in magnitude than the component associated with the experimental setup mean($\sigma_{\mathrm{bt},i}^2$) for both, RTR and ROC. Thus, subsample topology has a considerable impact on the statistical error of VS validations. However, it was not possible to deduce any regularity or correlation with the measures for data set clumping introduced here.

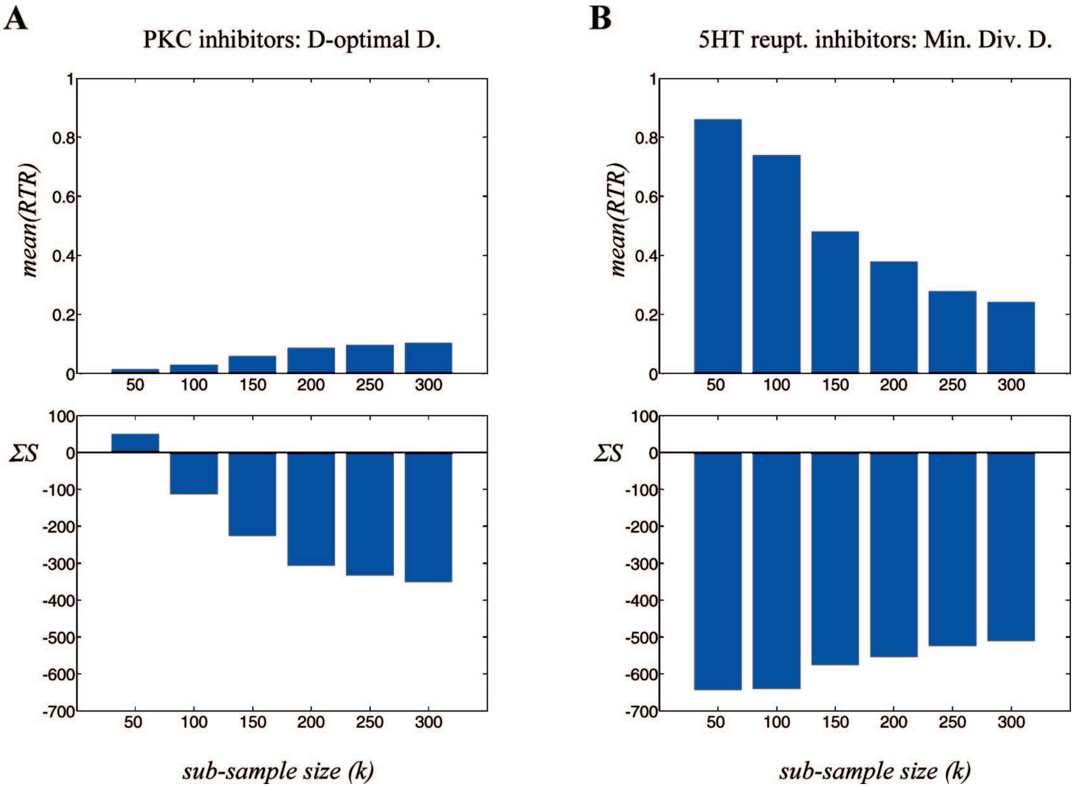

Figure 6. Strong correlation between mean(RTR) and ΣS, shown here for MOE-PCA descriptors. (A) Subsamples generated by D-optimal design from the data set of PKC inhibitors show a small degree of clumping indicated by positive or small negative values of ΣS and low mean(RTR). (B) The high degree of clumping observed in minimum distance design subsamples of 5HT reuptake inhibitors is associated with high mean(RTR).

ΣS only coarsely characterizes the degree of clumping in a data set of substances. The complex inter-relation of features like the number of clusters, their respective size, density, and separation can not be reflected in detail by the simple scalar ΣS. In some pathological cases, the specific topology of two subsamples can lead to very similar values of ΣS but to different outcomes of VS validation. This is illustrated by the extreme example of two subsamples extracted from the data set of Renin inhibitors (OD, $k = 50$) and 5HT reuptake inhibitors (OD, $k = 300$). (Figure 7) Both samples have a similar ΣS of -454.4 and -453.8, but the mean(RTR) is 0.58 on the Renin inhibitors subsample and 0.18 on the subsample of 5HT reuptake inhibitors. When analyzing the graphs of $G(t)$ (Figure 7A), it is obvious that the average distances between actives in the sample of 5HT reuptake inhibitors are smaller, i.e. the density in the sample is higher. However, the respective graph for $F(t)$ (Figure 7B, red) with its prevalence of lower point-event distances shows quite clearly that this is mainly caused by small, local clusters evenly spread across chemical space. On the other hand, the subsample of Renin inhibitors is well separated from the rest of chemical space, a fact that dominates the validation result. (Figure 7B, blue) This is also well reflected in the graphs of $S(t)$, in which the separation of the Renin inhibitors from the background is indicated by a rightward shift (Figure 7C). A SOM representation of both subsamples (Figure 7D) provides an intuitive visualization of the respective conditions in the subsamples.

ΣS usually provides a robust and easily interpretable measure for data set clumping and its effect on VS performance. However, to assess the topology in more detail, the graphs of $G(t)$, $F(t)$, and $S(t)$ have to be inspected. Here, visualization of data set topology by self-organizing maps can be of great benefit as it facilitates the intuitive perception of the information provided by the spatial statistics functions.

Relation of ΣS to Other Approaches for the Estimation of Data Set Self-Similarity in VS Validation. A commonly used measure to quantify data set self-similarity in descriptor spaces is the average of pairwise distances (denoted as avD). In order to investigate if the characterization of data set topology by ΣS really provides additional information over avD, the latter was computed for all subsamples in MOE-PCA descriptor space. Since $S(t)$ and thus also ΣS combine a statistic of the distances between nearest neighbors in a data set ($G(t)$) and a statistic of distances to random points in chemical space ($F(t)$), the question arises whether any one of them can explain the differences in VS performance alone. In order to express the information of $G(t)$ and $F(t)$ by a scalar, the median nearest neighbor distance g was obtained as the value for t, where $G(t) = 0.5$. In an analogous fashion, the median distance f of a point to the nearest event was determined as the value of t where $F(t) = 0.5$, for each of the 234 subsamples encoded by MOE-PCA descriptors.

The values obtained for avD, g, and f were correlated with VS performance on all subsamples in the same way as ΣS. Additionally, the correlation coefficients were also determined separately for the different subsample design strategies (DOD, OD, MDD), in order to determine if any of the methods can capture data set topology of a certain type (dispersed, patchy or concentrated) especially well.

Besides the already noted high correlation of mean(RTR) and ΣS, the results shown in Table 5 indicate moderate levels of correlation of avD and g with mean(RTR) over all

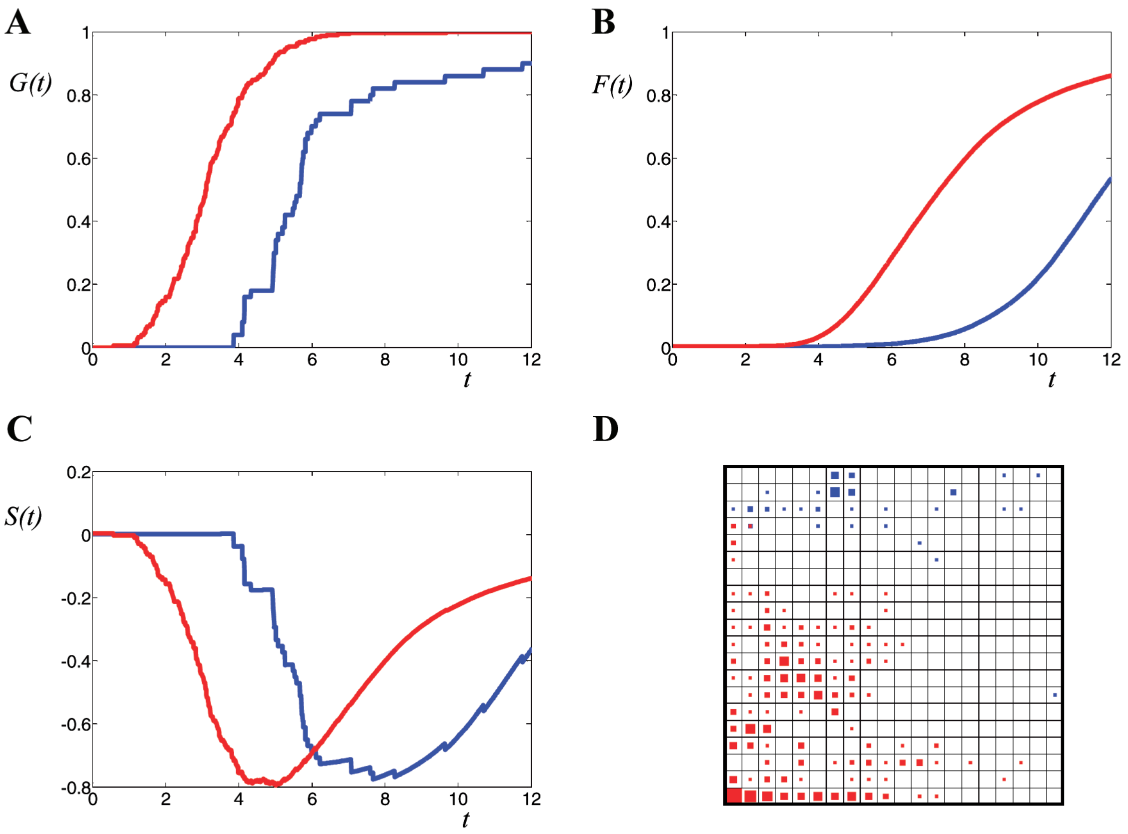

Figure 7. Estimation of clumping by Σ*S* leading to ambiguous results in some pathological cases. (A) The graph of *G*(*t*) for the subsample of 5HT reuptake inhibitors (OD, $k = 300$; red) exhibits regions of higher density than the sample of Renin inhibitors (min dist d, $k = 50$, blue). (B) The smaller amount of empty space for the 5HT reuptake inhibitors subsample is evident from its graph for *F*(*t*) (red). The larger amount of empty space present for the sample of Renin inhibitors (blue) indicates better separation of the sample from the background, causing better VS performance. (C) In this particular case, the graphs for *S*(*t*) of both subsamples are similar in shape and shifted on the *x*-axis. Summing over *S*(*t*) does not capture the rightward shift of the blue curve (Renin inhibitors) that indicates separation from the background. (D) The phenomena discussed in (A−C) are easily visualized on a SOM projection of the subsamples. (Renin inhibitors: blue; 5HT reuptake inhibitors: red).

Table 5. Correlation Coefficients of Several Measures of Data Set Topology and mean(RTR)[a]

ρ(mean(RTR), . . .	Σ*S*	avD	*g*	*f*
all subsamples	−0.93	−0.77	−0.53	0.44
DOD	−0.92	−0.19	−0.07	0.31
OD	−0.74	−0.04	−0.21	0.39
MDD	−0.95	−0.79	−0.82	0.27

[a] Spearman rank correlation coefficients. avD: average pairwise intraset distance. *g*: median intraset nearest neighbor distance. *f*: median point-event distance.

subsamples. The respective numbers for the different design strategies show however, that these levels of correlation are mostly caused by subsamples with concentrated topologies (MDD). This result can be explained by the fact that for concentrated subsamples, empty space will always be quite large with only marginal variations. Consequently, the main factor for their discrimination is the distribution of event-event distances, which is well captured by statistics like avD and *g*. However, this also shows, that avD and *g* can explain differences in VS performance only for concentrated data sets. On the other hand, *f* holds considerable information for dispersed (D-optimal) and patchy (OD.) subsamples. The large event−event distances in these subsamples render VS a difficult task. However, if a subsample is located in a region of chemical space, which is sparsely populated by inactives (i.e., more empty space, large *f*) VS performance will increase. This is a general feature that explains the moderate, but fairly constant (over all sampling strategies) level of correlation of *f* with VS performance. Whenever a set of actives is well separated from the rest of chemical space, VS performance increases. As discussed above, this is exactly the effect that causes better VS performance on the Renin inhibitors subsample vs the subsample of 5HT reuptake inhibitors.

Summarizing, all of these measures are able to reflect the impact of data set topology on VS performance under certain conditions, but only Σ*S* provides a comprehensive coverage across all types of data set topology, as indicated by the consistently higher correlation coefficients for Σ*S* in Table 5. Put another way, neither statistics about intraset distances, nor information about empty space can explain differences in VS performance alone. The augmentation of the nearest-neighbor function *G*(*t*) by the empty space function *F*(*t*) provides a substantial amount of additional information.

Better Descriptors Generate More Clumping. It was shown above, that the success rate of virtual screening is higher on data sets that have a clumpy topology in descriptor space. In a way, this is not surprising, since this is exactly what chemical descriptors were invented for. Good descriptors map substances with similar bioactivities to similar points in descriptor space, thereby introducing clumping. However, some descriptors are better at this task and some are worse. Put another way: If the same data set (or in our

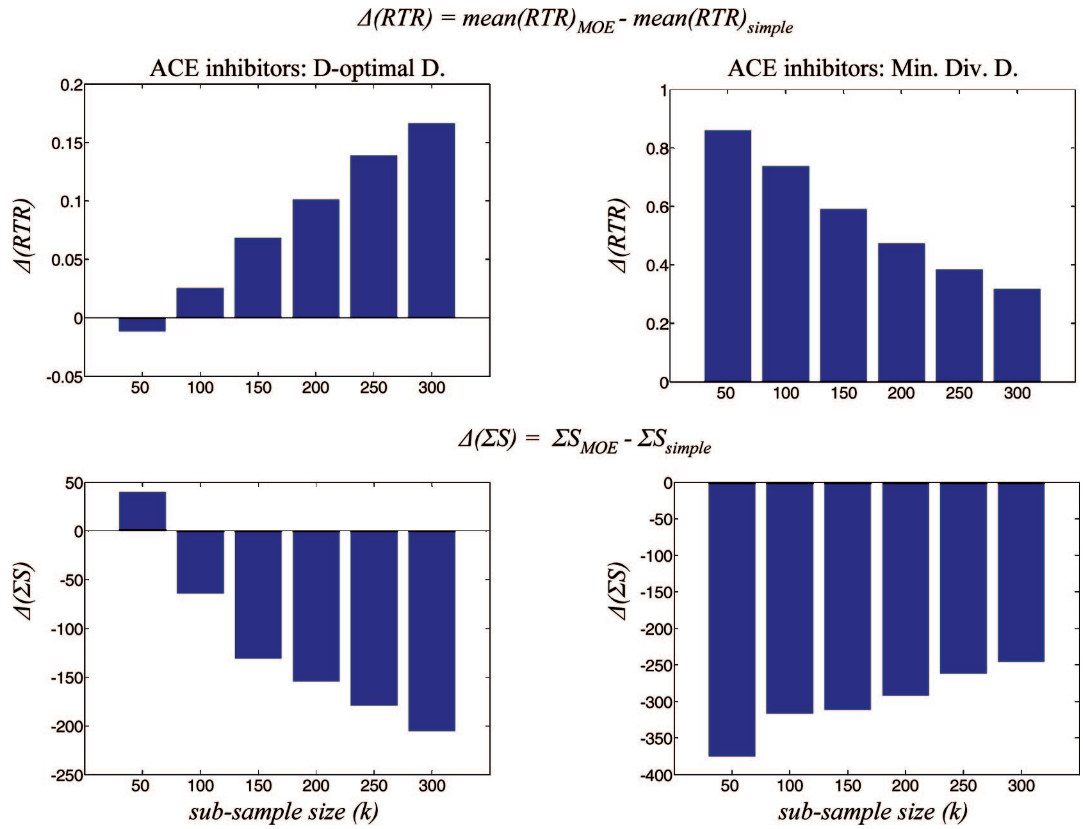

Figure 8. Gain in clumping of MOE-PCA descriptors vs simple descriptors for subsamples of the ACE inhibitors data set generated by D-optimal design (A) and minimum diversity design (B), respectively. (A) Δ(Σ*S*) is positive for $k = 50$, meaning that the representation of the subsample by simple descriptors features a higher degree of clumping. Accordingly, Δ(RTR) is negative for that subsample, indicating the better performance of simple descriptors. Δ(Σ*S*) shows negative values for the rest of the subsamples and is associated with positive values for Δ(RTR), which indicates that MOE-PCA outperforms simple with respect to mean(RTR) and clumping. (B) The minimum diversity design constitutes the best-case scenario for MOE-PCA descriptors. Accordingly both, Δ(RTR) and Δ(Σ*S*) indicate higher performance of MOE-PCA descriptors in these subsamples.

Table 6. VS Performance and Degree of Data Set Clumping for Complete Benchmark Data Sets

	MOE-PCA			simple		
	mean(RTR)	mean(ROC)	Σ*S*	mean(RTR)	mean(ROC)	Σ*S*
ACE inhibitors	0.34	0.79	−541.5	0.10	0.65	−316.7
AChE inhibitors	0.17	0.81	−455.4	0.09	0.77	−386.4
angio. r. blockers	0.23	0.84	−474.8	0.15	0.84	−445.8
COX inhibitors	0.11	0.78	−387.4	0.08	0.82	−387.5
D2 antagonists	0.19	0.87	−457.3	0.11	0.83	−411.6
HIV P. inhibitors	0.16	0.78	−454.7	0.09	0.73	−352.5
5HT1A agonists	0.16	0.87	−484.1	0.11	0.82	−421.1
5HT3 antagonists	0.29	0.92	−507.5	0.26	0.92	−495.5
5HT reup. inhibitors	0.20	0.86	−488.6	0.15	0.83	−433.4
PKC inhibitors	0.19	0.69	−409.6	0.07	0.57	−281.3
Renin inhibitors	0.45	0.93	−602.8	0.50	0.96	−692.2
subst. P inhibitors	0.16	0.79	−438.2	0.07	0.73	−367.3
thrombin inhibitors	0.23	0.82	−443.6	0.13	0.76	−342.8

case a subsample of a data set) is encoded by two different descriptors, VS performance should be better using the descriptor representation that introduces the more favorable topology, i.e. the higher degree of clumping. For the two descriptors (MOE-PCA and simple) used here, we calculated the difference in VS performance for each sample as

$$\Delta(\mathrm{FoM}) = \mathrm{mean}(\mathrm{FoM}_{\mathrm{MOE}}) - \mathrm{mean}(\mathrm{FoM}_{\mathrm{simple}}); \quad \mathrm{FoM: RTR\ or\ ROC} \quad (9)$$

for both RTR and ROC, respectively. Accordingly, Δ(FoM) is positive, whenever MOE-PCA performs better and negative if the simple descriptors generate superior figures of merit. The difference in subsample clumping was calculated in an analogous manner as follows:

$$\Delta(\Sigma S) = \Sigma S_{\mathrm{MOE}} - \Sigma S_{\mathrm{simple}} \quad (10)$$

Since a higher degree of clumping is associated with negative values of Σ*S*, positive values of Δ(Σ*S*) indicate more clumping for the representation by simple descriptors and vice versa. For all subsamples examined in this study, a strong correlation was found between Δ(Σ*S*) and Δ(FoM). The correlation coefficients were $\rho(\Delta(\Sigma S), \Delta(\mathrm{RTR})) = -0.93$ and $\rho(\Delta(\Sigma S), \Delta(\mathrm{ROC})) = -0.88$ for RTR and ROC, respectively. An example for subsamples taken from the ACE

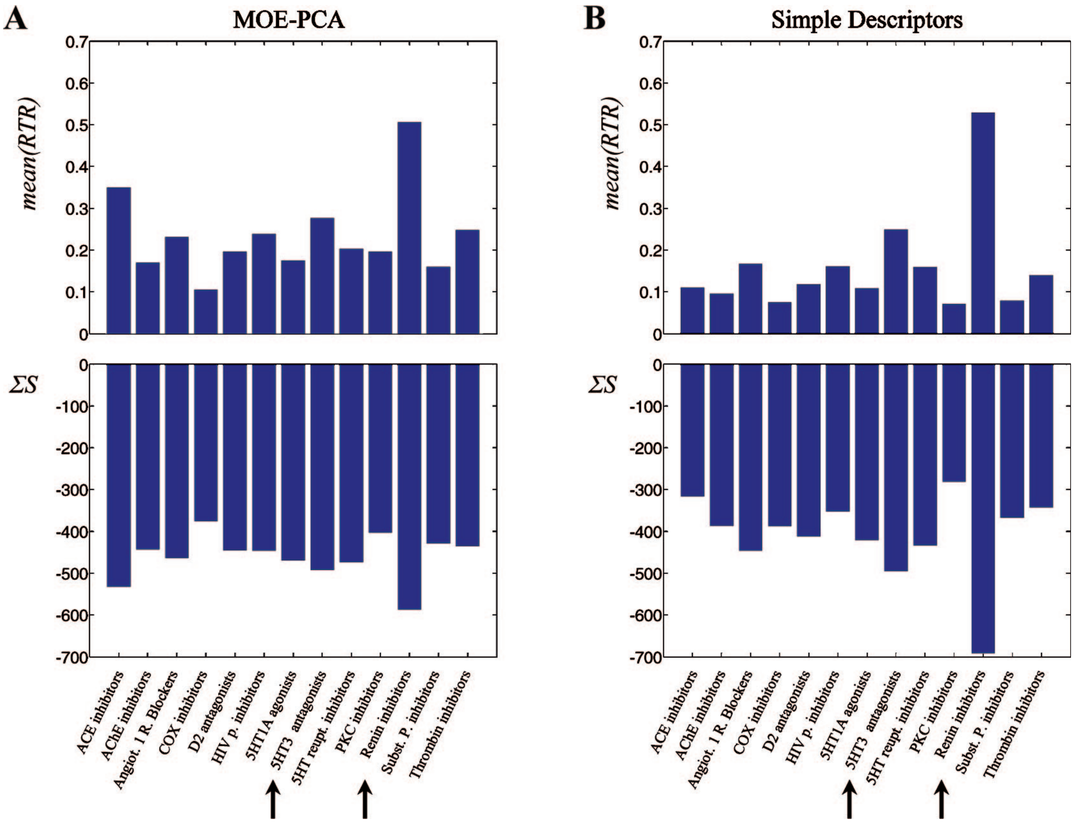

Figure 9. VS performance and data set clumping as observed on the complete benchmark data sets. As shown for the subsamples of controlled topology, a strong correlation exists between data set clumping and VS performance. Arrows indicate data sets discussed in more detail in the text.

Table 7. Correlation Coefficients of VS Figures of Merit with Data Set Clumping (ΣS) for the Complete Benchmark Data Sets

	$\rho(\Sigma S$, mean(RTR))	$\rho(\Sigma S$, mean(ROC))
MOE-PCA	−0.72	−0.77
simple descriptors	−0.77	−0.97

inhibitors data set is given in Figure 8. Better VS performance for a certain type of descriptors is closely associated by a gain in clumping of the respective subsample. It should be recalled, that the DOD subsamples with $k = 50$ constitute an artificial absolute worst-case scenario for performing VS with MOE-PCA descriptors. Thus, since it cannot get any worse, it is highly likely that the representation of the same substances in simple descriptor space is more clumpy. Therefore, the better performance of simple descriptors on these subsamples is mainly an effect of the sampling strategy.

Application of the Method to Whole Data Sets. Using subsamples generated by different design strategies, it was possible to observe the impact of data set topology on the results of VS validation. Furthermore, it was shown, that this impact can be quantified using spatial statistics methods. However, when performing a real-life validation of a ligand-based VS technique, one would not be interested in the VS performance on artificial subsamples, but on the complete benchmark data sets.

Table 6 and Figure 9 summarize the results of retrospective VS simulations following the procedure stated above (100 query/validation set splits, 10 query substances, MAX-rule data fusion) and refined nearest neighbor analysis on the complete benchmark data sets. Again, a strong correlation between data set clumping and VS performance as measured by ΣS, mean(RTR), and mean(ROC), respectively, can be observed (Table 7). The somewhat smaller values for the correlation coefficients are mainly due to the much smaller number of samples (13 complete data sets vs 234 subsamples) available for their calculation, so that deviations from the generally observed relationship have a higher impact. Once more, it must be stated that ΣS is a robust but rough estimate of data set clumping, which can not explain all effects of data set topology on VS performance perfectly. Discrepancies from the overall correlation of ΣS and VS performance such as those observed in Figure 9, can usually be explained by a more detailed inspection of $G(t)$, $F(t)$, and $S(t)$.

Also, the correlation of a gain in clumping and superior VS performance persists for the complete data sets. (Figure 10). Here, the respective correlation coefficients are $\rho(\Delta(\Sigma S)$, $\Delta(\text{RTR})) = -0.80$ and $\rho(\Delta(\Sigma S), \Delta(\text{ROC})) = -0.94$.

Apart from minor discrepancies, $\Delta(\Sigma S)$ is able to explain most of the differences in VS performance. Large gains in clumping are always associated with much better performance of the respective descriptor and small differences in topology coincide with small or no changes in VS performance. For answering the question if a particular descriptor really improves VS performance in a real-life VS campaign, these more general trends are of higher importance.

Benchmark Data Set Bias. From the chart of ΣS for the benchmark data sets encoded by the simple descriptors (Figure 9), the high degree of clumping observed for the data sets of Renin inhibitors and 5HT3 antagonists is striking. The values of ΣS for these data sets show, that they can easily be distinguished from the background even by these absolutely simple descriptors that do not encode molecular

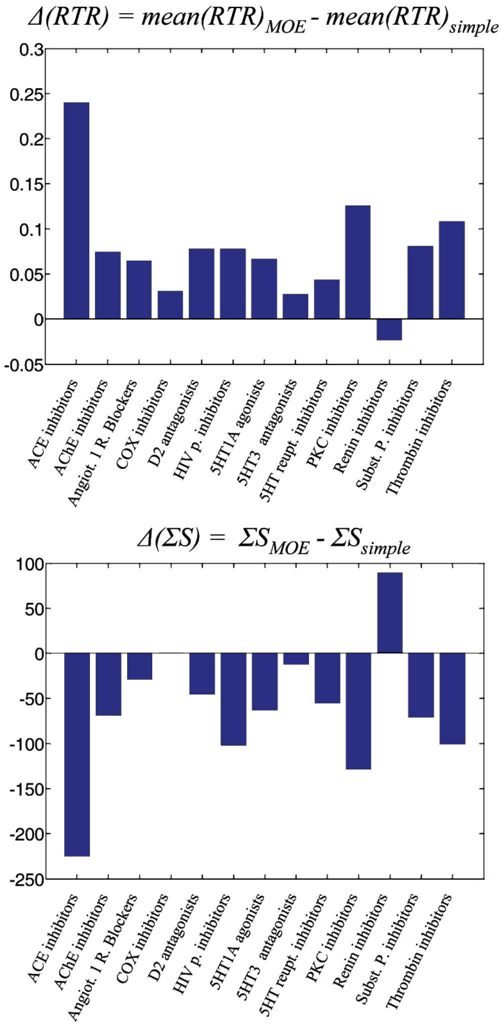

Figure 10. Differences in VS performance and gain in clumping of MOE-PCA vs simple descriptors on the complete benchmark data sets. A strong overall correlation of $\Delta(\Sigma S)$ and differences in VS performance can be observed.

connectivity. Accordingly, it should not be difficult to achieve good results when performing VS validation on these data sets. So, these data sets introduce a bias toward good validation results, which is also reflected by the fact that the highly complex MOE-PCA descriptors can not generate any significant gain in clumping over simple descriptors on these data sets. Actually both, clumping and VS performance are lower for the Renin inhibitors data set with MOE-PCA than with simple descriptors. Thus, the mean(RTR) values of ~0.3 and ~0.5 achieved with MOE-PCA on these data sets, which would normally be considered quite acceptable, are of no real value for the evaluation of the VS capabilities of MOE-PCA descriptors. This is exactly the effect that has been termed artificial enrichment by Verdonk et al.[29] for validations of molecular docking programs. For docking algorithms, this effect is caused by the fact that actives and inactives are too dissimilar regarding simple molecular properties, i.e. they are separated in simple descriptor space. Our results show, that for ligand-based virtual screening artificial enrichment does not only depend on the separation between actives and inactives (i.e. $F(t)$) but also on the distances between actives themselves (i.e. $G(t)$), a fact that should be taken into account in the design of validation experiments comparing ligand based virtual screening methods (Figure 11). As a first consequence, data sets that show a high degree of clumping in a simple reference descriptor space should be avoided in VS validation experiments. If, for some reason they cannot be evaded, the value of ΣS determined in a simple reference descriptor space can effectively be used to estimate an expectation for VS performance. Moreover, using the information provided by $F(t)$ and $G(t)$, it is possible to determine if the clumping in simple descriptor space is mainly caused by a high degree of separation of the data set from the background, e.g. the data set of Renin inhibitors (Figure 11). In this case, the problem can be solved simply by choosing a more appropriate background database.

CONCLUSION

In the course of this study we were able to show that the topology of benchmark data sets in descriptor space has a considerable impact on the results of VS validation. Our results point out that in contrast to molecular docking, both, the mutual distances of the active substances and their separation from the inactives in descriptor space are of

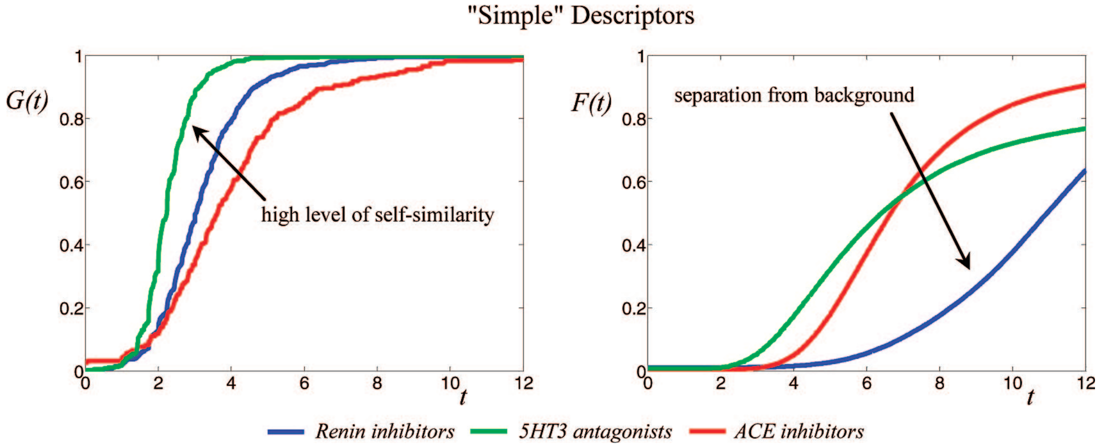

Figure 11. Benchmark data set bias is caused by both, high levels of intraset self-similarity in the data sets of actives and separation of actives from the background. Renin inhibitors (blue) and 5HT3 antagonists (green) both show benchmark data set bias caused by a high degree of clumping in simple descriptor space. A comparison of the respective graphs for $G(t)$ and $F(t)$ shows that for the 5HT3 antagonists this is mainly caused by small event−event distances. On the Renin data set, it is mainly due to a high degree of separation from the rest of chemical space. In comparison, the data set of ACE inhibitors (red), which is not subject to benchmark data set bias, exhibits neither self-similarity nor separation from the background.

importance for the validation of ligand-based virtual screening techniques. Both factors can effectively be described by the spatial statistics methodology introduced here. Based on this framework, data set topology can be examined on all levels of detail, ranging from a quick estimate of clumping to an in-depth analysis of clustering behavior. An indicator for data set clumping, ΣS, is introduced that can effectively be utilized to estimate an expectation of VS performance of a given benchmark data set. Furthermore, the methodology proposed in this paper can provide insights about the reasons for differences in the VS performance of different descriptors.

The methodology for the characterization of data set topology presented here does not imply any prior assumptions or preconditions about the composition of the data sets. On the contrary, our methodology is actually suited to provide exactly this information, i.e. if the data set is composed of a single or multiple clusters and if these are close or separated in chemical space. An obvious field of potential future use for this piece of information would be the rational design of validation experiments comparing different algorithms for similarity searching. On a patchy or dispersed data set, for instance, any algorithm based on multiple query molecules should be superior to an algorithm with only one query as an input. This advantage would be annihilated on a concentrated data set. In this context, an unbiased selection of benchmark data sets could be greatly facilitated by the topology analysis proposed in this paper.

Our methodology uses two basic functions for the elucidation of benchmark data set topology: The nearest-neighbor function $G(t)$ reflects the distribution of intraset "active-to-active" distances, whereas the empty-space function $F(t)$ represents the distribution of "active-to-background" distances. This opens up another field of potential future use for our methodology: By maximizing intraset nearest-neighbor distances and minimizing the separation from the background in a simple reference descriptor space, benchmark data sets for ligand-based virtual screening can be designed, that prevent artificial enrichment as postulated by Verdonk et al.[29] Current work is underway in our laboratory to provide a collection of such data sets.

ACKNOWLEDGMENT

We thank the anonymous referees whose fruitful comments helped to improve an earlier draft of this paper. Furthermore, we thank Dr. Nikolaus Stiefl, Novartis, Basel, Switzerland, for helpful discussions. Financial support of the German Research Foundation (DFG) is acknowledged (SFB-630-TPC5).

Supporting Information Available: Table 2 containg ΣS for all subsamples in MOE-PCA and simple descriptor space, respectively and Table 3 showing mean(RTR), mean(ROC), and the respective σ^2, $\sigma_{\mathrm{top}}{}^2$, $\sigma_{\mathrm{bt},i}{}^2$, std$_{\mathrm{top}}$ for MOE-PCA descriptors and mean(RTR) and mean(ROC) for simple descriptors for all subsamples. 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, Germany, 2000.

(2) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189–1202.

(3) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug. Discov.* **2004**, *3*, 935–949.

(4) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.

(5) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity. *J. Med. Chem.* **2002**, *45*, 4350–4358.

(6) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(7) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.

(8) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.

(9) 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.

(10) 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.

(11) Pearlman, D. A.; Charifson, P. S. Improved Scoring of Ligand-Protein Interactions Using OWFEG Free Energy Grids. *J. Med. Chem.* **2001**, *44*, 502–511.

(12) Bender, A.; Glen, R. C. A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369–1375.

(13) 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.

(14) 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.

(15) Cleves, A. E.; Jain, A. N. Robust Ligand-Based Modeling of the Biological Targets of Known Drugs. *J. Med. Chem.* **2006**, *49*, 2921–2938.

(16) Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2006**, *49*, 5856–5868.

(17) Klon, A. E.; Glick, M.; Davies, J. W. Application of Machine Learning To Improve the Results of High-Throughput Docking Against the HIV-1 Protease. *J. Chem. Inf. Model.* **2004**, *44*, 2216–2224.

(18) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.

(19) *MDL Drug Data Report (MDDR)*; Elsevier MDL: San Ramon, CA, 2005.

(20) Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Training similarity measures for specific activities: application to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46*, 577–586.

(21) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold hopping using clique detection applied to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503–511.

(22) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D Pharmacophore Descriptors for Scaffold Hopping. *J. Chem. Inf. Model.* **2006**, *46*, 208–220.

(23) 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.

(24) Yang, J.-M.; Chen, Y.-F.; Shen, T.-W.; Kristal, B.; Hsu, D. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134–1146.

(25) 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.

(26) Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S. The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2145–2156.

(27) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43–53.

(28) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(29) 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.

(30) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(31) 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.

(32) 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.

(33) 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.

(34) Vogt, M.; Bajorath, J. Introduction of a Generally Applicable Method to Estimate Retrieval of Active Molecules for Similarity Searching using Fingerprints. *ChemMedChem* **2007**, *2*, 1311–1320.

(35) Fortin, M.-J.; Dale, M. R. T. *Spatial analysis: a guide for ecologists*; Cambridge University Press: Cambridge, UK, 2005.

(36) Diggle, P. J. Statistical methods for spatial point patterns in ecology. In *Spatial and temporal analysis in ecology*; Cormack, R. M., Ord, J. K., Eds.; International Cooperative Publishing House: Fairland, MD, 1979; pp 95−150.

(37) *MOE Molecular Operating Environment*, 2002.03; Chemical Computing Group, Inc.: Montreal, Canada, 2002.

(38) *CTFile Formats*; Elsevier MDL: San Ramon, CA, 2005.

(39) *3D Structure Generator CORINA: Generation of High-Quality Three-Dimensional Molecular Models*; Molecular Networks GmbH Computerchemie: Erlangen, Germany, 2006.

(40) Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer-Verlag New York, Inc.: New York, 2002.

(41) *BABEL3*, 2.2; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2006.

(42) *FILTER*, 2.0.1; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2007.

(43) De Aguiar, P. F.; Bourguignon, B.; Khots, M. S.; Massart, D. L.; Phan-Than-Luu, R. D-Optimal designs. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 199–210.

(44) Johnson, M. E.; Nachtsheim, C. J. Some guidelines for constructing exact D-optimal designs on convex design spaces. *Technometrics* **1983**, *25*, 271–277.

(45) Olsson, I.-M.; Gottfries, J.; Wold, S. D-optimal onion designs in statistical molecular design. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 3746.

(46) Olsson, I. M.; Gottfries, J.; Wold, S. Controlling coverage of D-optimal onion designs and selections. *J. Chemom.* **2004**, *18*, 548–557.

(47) Scott, D. W.; Thompson, J. R. Probability density estimation in higher dimensions, In *Interface: Computer Science and Statistics*; Proceedings of the Fifteenth Symposium, Houston, TX, 1983; Gentle, J., Ed.; North-Holland: Amsterdam, The Netherlands, 1983; pp 173–179.

(48) *Matlab 7*; The Mathworks: Natick, MA, 2006.

(49) Atkinson, A. C.; Donev, A. N. *Optimum Experimental Designs*; Oxford University Press: Oxford, UK, 1992.

(50) Box, G. E. P.; Hunter, J. S.; Hunter, W. G. *Statistics for Experimenters: Design, Discovery and Innovation*, 2nd ed.; Wiley & Sons: Hoboken, NJ, 2005.

(51) Godden, J. W.; Bajorath, J. A distance function for retrieval of active molecules from complex chemical space representations. *J. Chem. Inf. Model.* **2006**, *46*, 1094–1097.

(52) Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model.* **2006**, *46*, 717–727.

(53) Klabunde, T.; Hessler, G. Drug design strategies for targeting G-protein-coupled receptors. *ChemBioChem.* **2002**, *3*, 928–944.

(54) Ripley, B. D. 2nd-Order analysis of stationary point processes. *J. Appl. Probab.* **1976**, *13*, 255–266.

(55) Guha, R.; Dutta, D.; Jurs, P. C.; Chen, T. R-NN curves: an intuitive approach to outlier detection using a distance based method. *J. Chem. Inf. Model.* **2006**, *46*, 1713–22.

(56) Guha, R.; Dutta, D.; Wild, D. J.; Chen, T. Counting clusters using R-NN curves. *J. Chem. Inf. Model.* **2007**, *47*, 1308–18.

(57) Upton, G. J. G.; Fingleton, B. *Spatial Data Analysis by Example*; Wiley & Sons Ltd: New York, 1985.

(58) Breimann, L. *Using convex pseudo-data to improve prediction accuracy*; Statistics Department, University of California Berkeley: Berkeley, CA, 1998.

(59) Kohonen, T. *Self-organizing maps*; Springer-Verlag New York, Inc.: Secaucus, NJ, 1997.

(60) Vesanto, J. *Data Mining Techniques Based on the Self-Organizing Map*; Helsinki University of Technology: Helsinki, Finland, 1997.

(61) Vesanto, J. SOM-Based Data Visualization Methods. *Intell. Data Anal.* **1999**, *2*, 111–126.

(62) Vesanto, J.; Alhoniemi, E. Clustering of the Self-Organizing Map. *IEEE Trans. Neural. Netw.* **2000**, *11*, 586–600.

(63) Alhoniemi, E.; Himberg, J.; Parhankangas, J.; Vesanto, J. *SOM Toolbox*, 2.0; SOM Toolbox Team, Laboratory of Computer and Information Science: Helsinki, Finland, 2005.

CI700099U