Relationships between Molecular Complexity, Biological Activity, and Structural Diversity

Ansgar Schuffenhauer,* Nathan Brown, Paul Selzer, Peter Ertl, and Edgar Jacoby

Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland

Received August 31, 2005

Following the theoretical model by Hann et al. moderately complex structures are preferable lead compounds since they lead to specific binding events involving the complete ligand molecule. To make this concept usable in practice for library design, we studied several complexity measures on the biological activity of ligand molecules. We applied the historical IC_{50}/EC_{50} summary data of 160 assays run at Novartis covering a diverse range of targets, among them kinases, proteases, GPCRs, and protein−protein interactions, and compared this to the background of "inactive" compounds which have been screened for 2 years but have never shown any activity in any primary screen. As complexity measures we used the number of structural features present in various molecular fingerprints and descriptors. We found generally that with increasing activity of the ligands, their average complexity also increased, and we could therefore establish a minimum number of structural features in each descriptor needed for biological activity. Especially well suited in this context were the Similog keys and circular substructure fingerprints. These are those descriptors, which also perform especially well in the identification of bioactive compounds by similarity search, suggesting that structural features encoded in these descriptors have a high relevance for bioactivity. Since the number of features correlates with the number of atoms present in the molecule, also the number of atoms serves as a reasonable complexity measure and larger molecules have, in general, higher activities. Due to the relationship between feature counts and densities on one hand and biological activity on the other, the size bias present in almost all similarity coefficients becomes especially important. Diversity selections using these coefficients can influence the overall complexity of the resulting set of molecules, which has an impact on the biological activity that they exhibit. Using sphere-exclusion based diversity selection methods, such as OptiSim together with the Tanimoto dissimilarity, the average feature count distribution of the resulting selections is shifted toward lower complexity than that of the original set, particularly when applying tight diversity constraints. This size bias reduces the fraction of molecules in the subsets having the complexity required for a high, submicromolar activity. None of the diversity selection methods studied, namely OptiSim, divisive *K*-means clustering, and self-organizing maps, yielded subsets covering the activity space of the IC_{50} summary data set better than subsets selected randomly.

INTRODUCTION

The discovery of lead structures is an important and still challenging task in the overall drug discovery process. It is possible to discover lead structures by virtual screening techniques based on known ligands (similarity[1−3] or pharmacophore searching[4]) or on target protein structures (docking[5]) or de novo design.[6,7] However, there are still many situations where neither any ligand nor the target structure is known. Also, there is always an interest in discovering new ligand chemotypes to identify new modes of action or to establish new intellectual property. Therefore, most lead-finding projects in the pharmaceutical industry use High Throughput Screening (HTS) methodologies. HTS aims to examine the chemistry space experimentally in a thorough manner for the desired activity. The size of the druglike chemistry space, which is estimated to contain 10^{60} structures,[8,9] prevents an exhaustive screening since the typical size of a screening collection that a pharmaceutical company can handle is currently in the range of 10^6 compounds.[10] In addition, the costs of protein production needed for screening can limit this number still further for targets where protein expression is difficult. In the HTS-based lead discovery practice, one is confronted by two challenges: how best to select from the commercially offered screening compounds those that should be added to the corporate screening collection; and, if a complete HTS is not possible, how best to select the subset out of the corporate screening collection, which should be screened in a lower-throughput assay. In the absence of additional information that can be used for virtual screening, a frequently chosen approach is to select a subset of maximum diversity based on chemical structure descriptors.[11−13] Furthermore, recently the aspect of molecular complexity has been discussed as an important parameter in the design of screening collections.[14−18] Given the fact that most molecular similarity measures on which diversity selections are based are influenced by the number of bits set in the fingerprints being compared[19,20] and that the number of bits-set is a measure for the complexity of a molecule, one may expect that diversity selection procedures can bias the complexity of the subsets generated with them. If there is a relationship between complexity and biological activity,

* Corresponding author phone: +41 61 32 45385; e-mail: ansgar.schuffenhauer@novartis.com.

10.1021/ci0503558 CCC: $33.50
Published on Web 12/14/2005

Table 1: Subset Size and Dissimilarity Criteria for the Subsets Selected with OptiSim/UNITY

maximum similarity	number of compounds	percentage of database (%)
1.0	191828	100.0
0.95	170655	89.0
0.88	123815	64.5
0.80	77909	40.6
0.73	48411	25.2
0.65	26511	13.8
0.58	14404	7.5
0.50	6498	3.4

as we have reported previously,[18] one can then expect a subsequent influence on the chance of discovering biological activity in these subsets. In this manuscript we will compare the bias on several molecular complexity measures resulting from diversity selections obtained by several procedures; we will then analyze the relation between molecular complexity and the distribution of biological activity in the diversity selection subsets.

EFFECT OF DIVERSITY SELECTION ON MOLECULAR COMPLEXITY

This study was conducted on a random selection from commercial compound vendor catalogues without any prefiltering steps except the elimination of such structures, for which the generation of SMILES (Simplified Molecular Input Line Entry Specification) failed or which contained multiple fragments. Replicate structures were also removed. This set comprises 191 828 unique structures. See the Supporting Information for a list of vendor collections included.

Different diversity selection or clustering procedures were used.

1. Diversity selection with the OptiSim algorithm[21] as implemented in the dbdiverse program from Tripos.[22] This program is based on Tanimoto similarity[20] comparisons of the UNITY fingerprints by Tripos. dbdiverse was executed in the exhaustive mode with increasingly tight constraints on the maximal diversity two molecules were allowed to have (Table 1). This method is herein referred to as OptiSim/UNITY.

2. The clustering component in PipelinePilot[23] uses an algorithm similar to Tripos' OptiSim and is used to select a maximally diverse subset of a given size. Each of the structures selected in this way forms the center of a cluster, and the remaining compounds were grouped together with the center that is most similar to them. For our purpose of diversity selection we retained simply the cluster centers as they were yielded by the clustering component in Pipeline Pilot. Pipeline Pilot has its own inbuilt circular substructure fingerprints[24] called FCFP_4, which are based on circular substructural fragments having a maximum diameter of 4 bonds. As Pipeline Pilot does not allow the same degree of control over the process and requires that one provides a desired number of clusters, the subset sizes obtained with dbdiverse as listed in Table 1 were used to request the generation of subsets of equal size. This method is herein referred to as OptiSim/FCFP_4.

3. The third diversity selection method is based on the hierarchical divisive *K*-means[25,26] clustering as implemented by BCI (Barnard Chemical Information).[27] Hierarchical divisive *K*-means clustering is a hierarchical clustering method that generates a complete hierarchy tree. From this tree, a number of horizontal cuts were made such that the numbers of clusters in each of the slices obtained was equivalent to one of the subset sizes in Table 1. The clustering was executed using the Pipeline Pilot FCFP_4 fingerprints, which were folded to a length of 2048 bits by simple modulo division. As BCI does not output designated cluster centers we have chosen from each cluster the member with a median number of fingerprint bits-set in order to avoid any size bias by the center selection procedure. This method is herein referred to as DivKM/FCFP_4.

In addition to the choice of the diversity selection method, it was also necessary to decide on which complexity measures to utilize in this study. Chemists often regard synthetic accessibility as the most important measure of molecular complexity. Although attempts have been made to quantify synthetic complexity, these are usually based on some empirical scoring of structural features[28] or statistical models based on individual assessments by chemists.[29] Here, we wished to use measures based on more general structural features that are independent from empirical parameters or individual assessments, which can be highly variable depending on the individual synthetic expertise of the chemists. The following measures were chosen:

1. Number of non-hydrogen atoms present in the structure as a very simple and intuitive measure.

2. Number of fingerprint bits-set in the fingerprint used for the diversity selection. This measure was chosen for the reason that it is the only complexity measure that is influenced directly by the diversity selection method.

3. Number of unique Similog[30] pharmacophore triplets present in the structure. The Similog keys represent pharmacophoric atom triplets, which are characterized by the bond count of the shortest path between the three atoms, and the properties of those three atoms. Four atom properties are recognized independently from each other: H-bond donor, H-bond acceptor, lipophilicity (recognized as the absence of electronegativity), and the bulkiness of the substituents. The Similog keys, as with pharmacophore keys, are assumed to describe those aspects of structural complexity which can be expected to be related with protein affinity.

4. The number of unique Similog pharmacophore triplets as described above is normalized by the theoretically possible number of atom triplets, which is $N(N-1)(N-2)$, where N is the number of non-H atoms in the molecule and a constant factor of 1/6 being ignored. This measure is herein referred to as the Similog density.

5. The number of bits-set in the FCFP_4 fingerprint normalized by the number of non-H atoms. This is the appropriate normalization since circular substructures are centered at each non-H-atom. This measure is herein referred to as FCFP_4 density.

For the subsets generated with each of the diversity selection methods, the distribution of the complexity measures is computed. The results are shown in Figures 1−4. The figures show clearly that OptiSim/UNITY and OptiSim/FCFP_4 result in a clear complexity bias toward lower complexity when considering increasingly tight dissimilarity criteria for all complexity measures, except the size normalized Similog and FCFP_4 density. For the Similog density OptiSim/UNITY, and to a lesser degree also OptiSim/

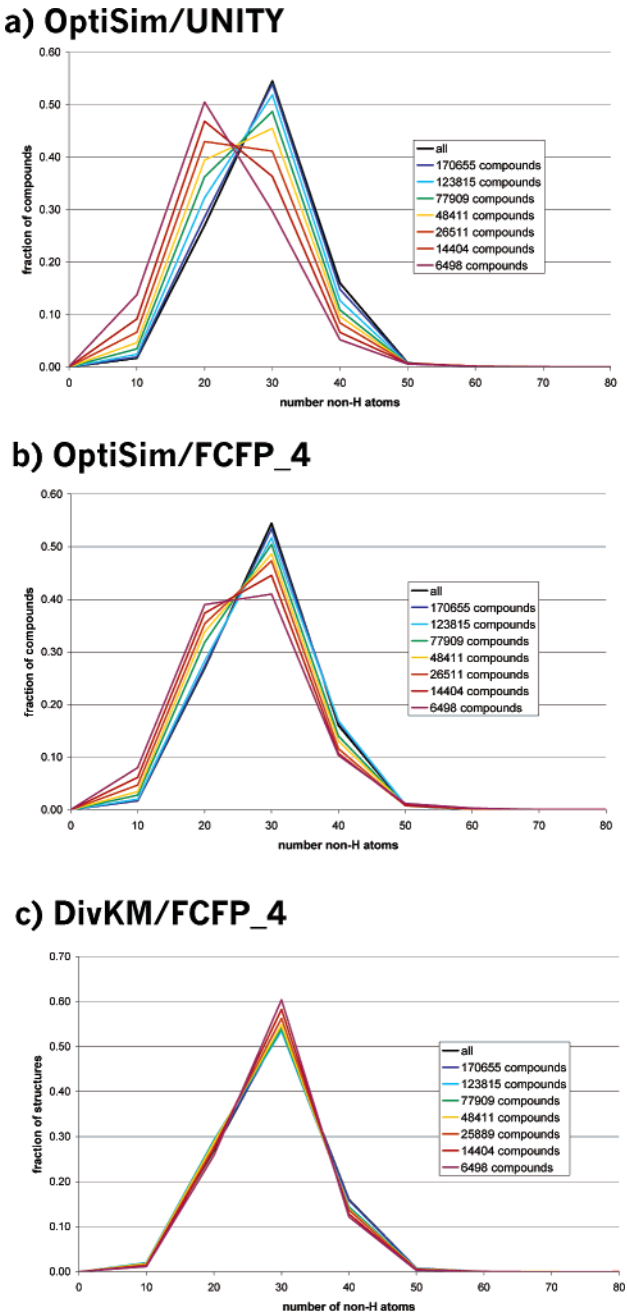

Figure 1. Distribution of the number of non-H atoms in the subsets generated with the different diversity selection methods. In the color coding chosen, the diversity selection criterion becomes tighter, and hence the subset sizes decrease from blue over yellow to red. The black curve shows the distribution in the original set.

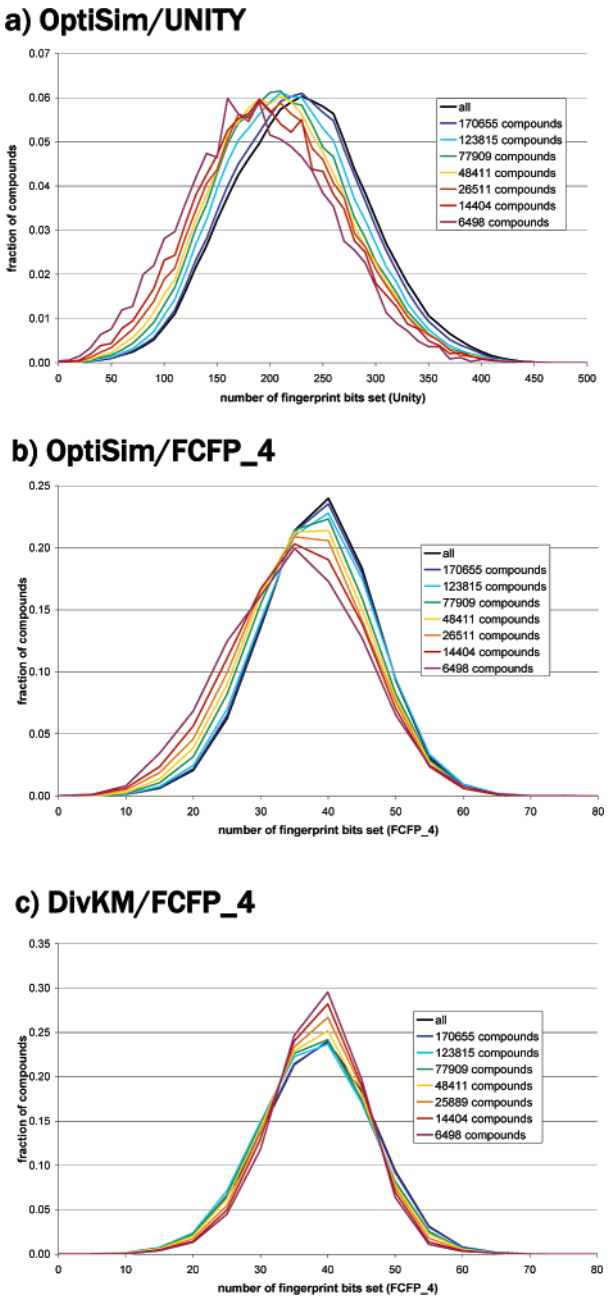

Figure 2. Distribution of the number of fingerprints set in the subsets generated with the different diversity selection methods. The distributions are shown for that fingerprint which was used for diversity selection. Color coding is as in Figure 1.

FCFP_4, a Gaussian distribution can be observed with a higher variance compared to the whole set. For the FCFP_4 density (not shown in the figures) the behavior is similar. The DivKM/FCFP_4 method differs from the OptiSim based methods as it does not exhibit these effects, and the distribution of molecular complexity remains without any significant changes.

RELATION BETWEEN MOLECULAR COMPLEXITY AND BIOLOGICAL ACTIVITY DETECTION IN HIGH-THROUGHPUT SCREENING

As experimentally shown by Kuntz[31] and also suggested by the statistical model of Hann[15] there is a relationship between biological activity and molecular complexity. This means that a complexity bias introduced by diversity selections can have an impact on the outcome of the screening of the selected subsets. To study the relationship between molecular complexity and biological activity it is necessary to use biological activity data obtained by high-throughput processes typically used to screen diversity-selected subsets. For this study we used the IC_{50} summary data of the Novartis screening history. We selected from this data set only such assays in which at least 1000 IC_{50} values have been submitted for determination, which means that most of the assays are HTS assays. This data set covers 160 assays representing a broad range of targets, including GPCR, ion channels, kinases, proteases, and also whole-cell phenotypic assays.

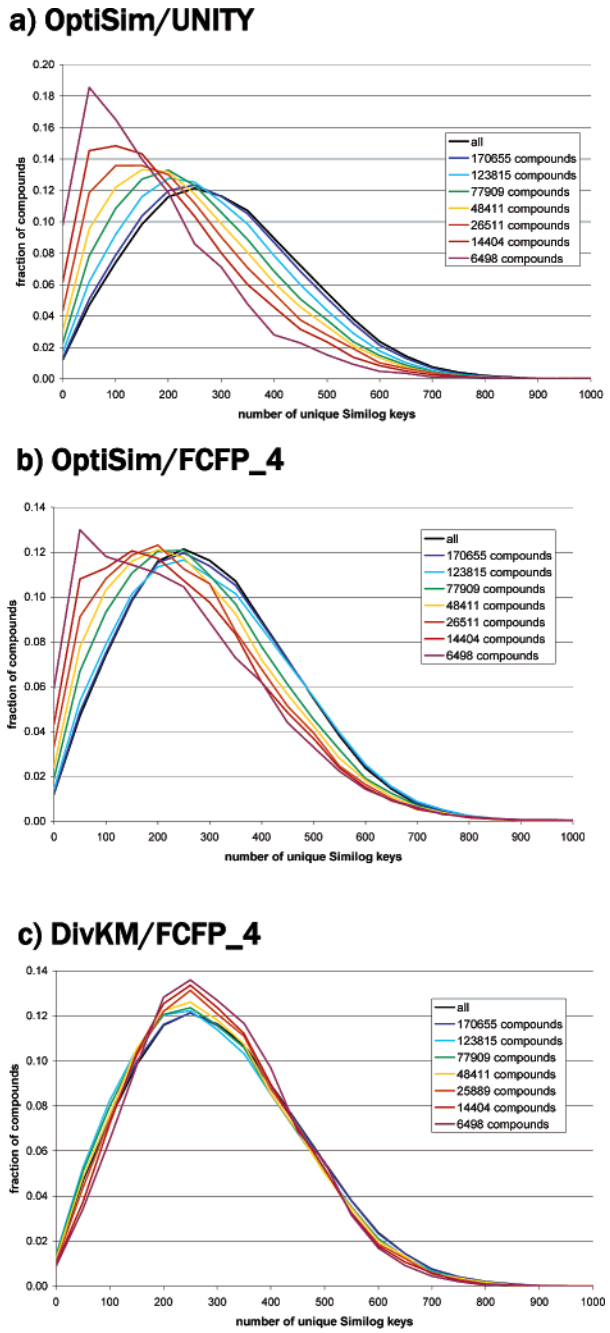

Figure 3. Distribution of the number of unique Similog keys set in the subsets generated with the different diversity selection methods. Color coding is as in Figure 1.

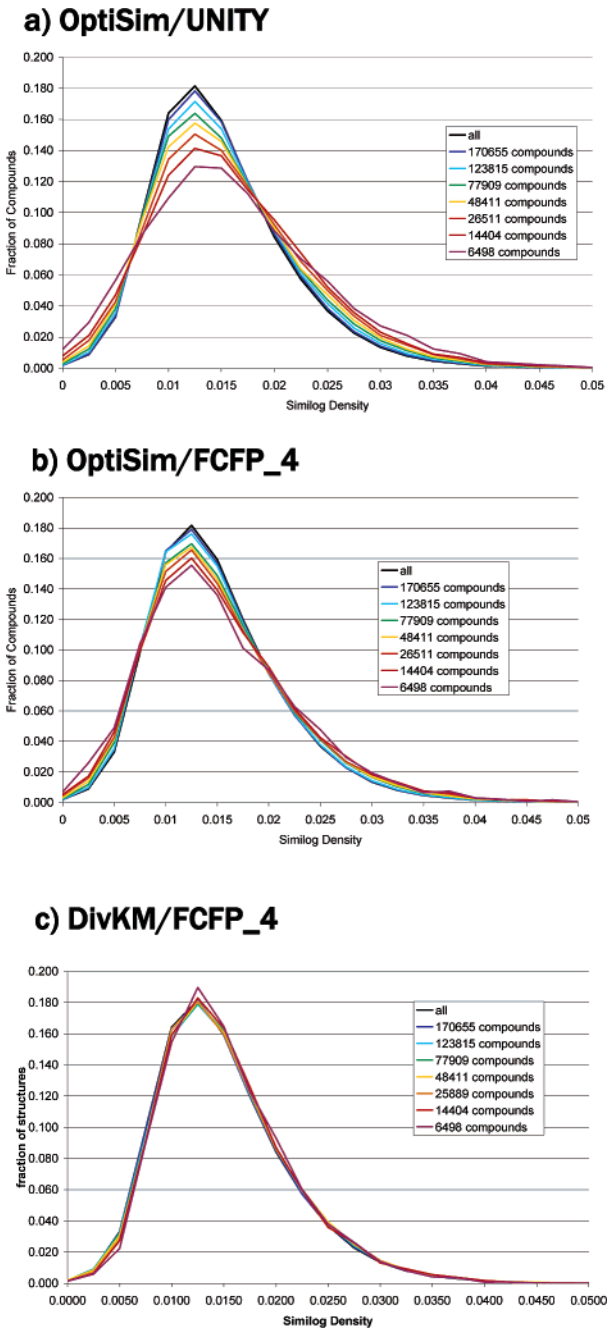

Figure 4. Distribution of the density of unique Similog keys (as the number of unique Similog keys set normalized by the possible number of atom triplets) in the subsets generated with the different diversity selection methods. Color coding is as in Figure 1.

This data matrix is clearly sparse, which means not all compounds have been measured on all assays. However, the majority of the assays are from HTS. HTS typically proceeds in two stages. The first one is the primary screening where the whole screening collection is assayed at a single concentration. The IC_{50} (or EC_{50} in case of receptor agonists) values are determined at a second stage for compounds that exhibit activity exceeding a predefined threshold. One could see the whole procedure as one logical assay which returns, with some error, whether a molecule is active or inactive with a quantified activity value. This gives some basis to assume that all compounds without a reported IC_{50} can be assumed to be inactive. It is acknowledged that HTS is likely to miss some false negative compounds; however, the outcome of an HTS is determined by those active compounds which can be detected by the screening procedure and should there be any impact of the molecular complexity on the HTS outcome, it will materialize itself in a change of the activity distribution of the detectable compounds.

Since the screening collection has changed over time, and also some assays are non-HTS, the general assumption that compounds are inactive if no IC_{50} value is reported is more questionable but may still be justified on the basis of the generally low hit-rates in screens of nonfocused collections. Ideally, any quantitative study of biological activity would be based on the binding energy which can be determined from the dissociation constant K_D. With the approximation of $IC_{50} \sim K_D$ we can use $-\log(IC_{50})$ as a quantifier for

activity. For molecules set as inactive, $-\log(IC_{50})$ was set to zero.

To study the dependence of overall biological activity on molecular complexity, the overall activity of the molecule was set as the maximum of $-\log(IC_{50})$ over all assays, ignoring any lower activity in other assays. The IC_{50} summary data contains only molecules that have at least exhibited some activity in one assay, which justified their submission for IC_{50} determination. While it is still possible that there is no activity found in IC_{50} determination and the molecules have been false-positives in primary screening, a set of these molecules would not be a suitable comparison set of inactive molecules. Therefore, we collected all molecules that have been in the screening collection for at least for 2 years and never found to be active in the primary screen and used this set as the set of inactives. For this set, and for the sets of molecules with $IC_{50} \leq 10\ \mu M$, $IC_{50} \leq 100$ nM, $IC_{50} \leq 1$ nM, the distributions for different complexity descriptors were calculated and are shown in Figure 5. The same descriptors as in section 1 were used, together with the numbers of non-H atoms, fingerprint bits-set, and unique Similog keys. Although additional complexity measures have been studied as we have reported previously,[18] these exhibited less difference in complexity between biologically active and inactive molecules.

It can clearly be seen that higher active molecules are also more complex than the average inactive molecule in terms of the number of non-H atoms, count of unique Similog keys, and number of FCFP_4 fingerprints bits-set. The complexity distribution of the inactive molecules is very similar to that of the whole screening collection (not shown in the figures). This suggests that there is nothing special about the inactive molecules, but the molecules found to be active are more complex than average. The three complexity measures are however highly correlated (Table 2). The absolute molecular size, measured as the number of atoms, appears to be the main parameter influencing the affinity of the molecules to their targets. Therefore, it is of interest to ascertain whether the use of the numbers of FCFP_4 fingerprint bits-set and unique Similog keys have any added value compared to the simple atom count. To investigate this, both biological activity and the complexity measure have been normalized by the molecular size. The ligand efficiency as a size-normalized measure for biological activity can be defined according to Hopkins[32] as $-\log(IC_{50})/N_{\text{non-H-atoms}}$. The normalized form of the Similog and FCFP_4 complexity are Similog and FCFP_4 density as described above. In Figure 6 the relationship between the ligand efficiency and the FCFP_4 and Similog densities, respectively, is shown. In both cases the ligand efficiency increases with increasing complexity density. The structure examples in Figure 6 illustrate that ligands that contain the same structural element repetitively are likely to have a low complexity density.

EFFECT OF DIVERSITY SELECTIONS ON BIOLOGICAL ACTIVITY

Having demonstrated that diversity selection can affect the complexity of the molecules in subsets obtained by their application, and also that biological activity is related to molecular complexity, the direct relationship between diversity selection and the biological activity of the selected

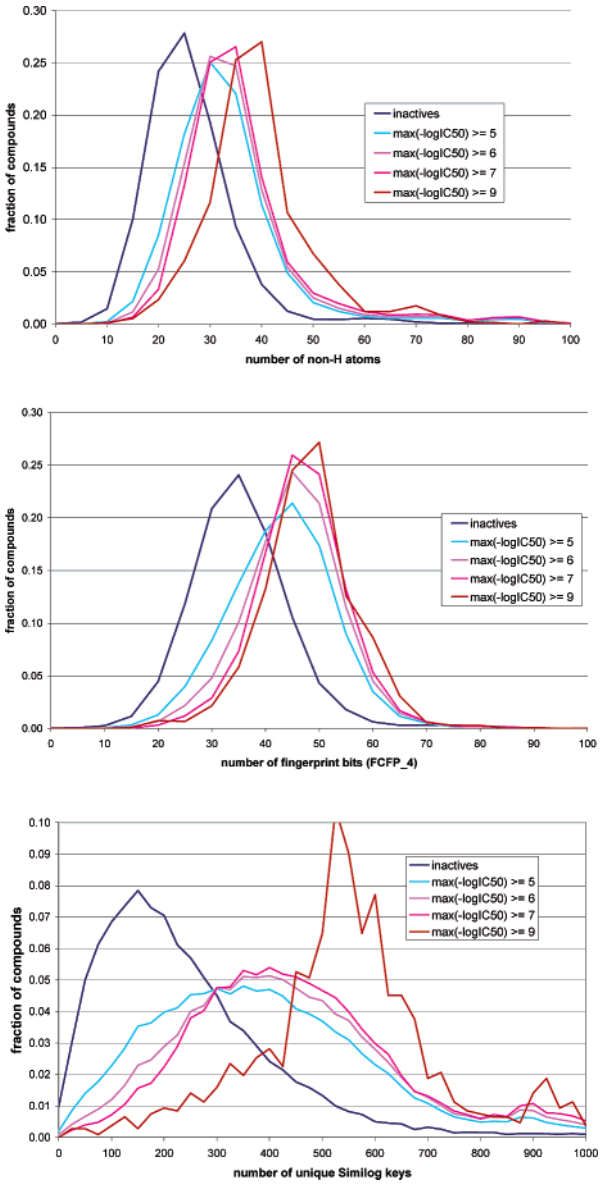

Figure 5. Distribution of molecular complexity measures in sets of molecules with increasing biological activity. Biological activity is measured as the maximal value of $-\log(IC_{50})$ measured found in the corporate IC_{50} summary data.

compounds is also of interest. To investigate this we applied each of the diversity selection methods described earlier (OptiSim/UNITY, OptiSim/FCFP_4, and DivKM/FCFP_4) to the same IC_{50} summary data set. In addition to these three methods which had been chosen for their capability to handle very large data sets, we also decided to study Self-Organizing Maps (SOMs), which are available in our company as an intranet tool for the analysis for medium sized data sets. Our SOM application used Radial Distribution Functions[33] (RDF) as a descriptor, which were derived from conformations obtained with CORINA.[34,35] A molecule's RDF code (eq 1) is a smoothed histogram of all occurring intramolecular atom distances which can be interpreted as the probability distribution of finding atom pairs of distance R

$$g(R) = \sum_{i>j}^{n} \sum_{j=1}^{i-1} a_i a_j e^{-B(R-r_{ij})^2}$$

where n is the total number of atoms, a_i and a_j are atomic

Table 2. Pearson Correlation Coefficients of Molecular Complexity Measures Determined from the Novartis IC_{50} Summary Data

	Num_atoms[a]	Num_FCFP_4[b]	Num_Similog[c]	Similog density[d]	FCFP_4density[e]
Num_atoms[a]	1	0.72	0.85	−0.56	−0.63
Num_FCFP_4[b]		1	0.77	−0.35	−0.04
Num_Similog[c]			1	−0.26	−0.42
Similog density[d]				1	0.58
FCFP_4 density[e]					1

[a] Number of non-H atoms. [b] Number of SciTegic FCFP_4 fingerprint bits-set. [c] Number of unique Similog keys present in the molecule. [d] Number of unique Similog keys present in the molecule normalized by number of atom triplets. [e] Number of bits-set in the SciTegic FCFP_4 fingerprint divided by the number of non-H atoms.

properties of atoms *i* and *j* (e.g. partial charges q_{tot}) and *B* is the smoothing factor which can be interpreted as the temperature parameter defining the fuzziness of atom positions due to thermal movement. In our experiment this fuzziness is very important since it allows us to compare RDF codes by calculating the Euclidean distance. In our studies we applied the partial charges q_{tot} as atomic properties a_{ij}. The RDF code *g(R)* was calculated in the distance range from 1 to 8.5 Å with a resolution of 0.3 Å (giving 25 real numbers). The codes were calculated three times for each molecule. In each run only atom pairs where both atoms have negative charges, one atom has a positive and one atom has a negative charge, and both atoms having negative charges were considered. This led to three structure codes that were concatenated to give a 75 (3 * 25) dimensional structure representation.

Training of the networks proceeded as described by Gasteiger et al.[36] SOMs consist of a two-dimensional arrangement of *x* * *y* connected neurons. Each neuron consists of *z* weights. The number of weights *z* corresponds to the dimensionality of the molecule representation (in our experiment 75). During training the set of molecules is presented to the network several times. In each iteration for each molecule the most similar neuron, the so-called winning neuron, is determined by calculating the Euclidean distance between the structure descriptor and the neuron weights. Then the neuron weights are adjusted to become more similar to the training data. This causes similar compounds being mapped to adjacent neurons. In case compounds are very similar they might be assigned even to the same neuron. This depends on the diversity of the data set, the number of compounds, and the size of the network.

SOMs can be applied for diversity selection and have been used in this aspect to split a data set in training and test set for statistical models.[37] A diversity selection with SOM starts with training a network having a number of neurons corresponding to the number of compounds that have to be selected. After training is finished the representative compound from each neuron, meaning the compound having the descriptor most similar to the neuron weights, is collected to create a representative subset (preserving the diversity of the whole data set). It is not guaranteed that every cell in the network will contain at least one compound, and therefore the size of the subset cannot be controlled exactly. Since the computational effort involved in training of the SOMs increases with the number of neurons, this method is especially suitable for creating small subsets. We used toroidal networks of 100, 70, and 30 neurons square, creating subsets of 7%, 3%, and 1% of the whole data set, respectively. This method is referred to herein as SOM/RDF.

The distribution of the different IC_{50} ranges in subsets of decreasing size obtained with increasingly tight diversity threshold is shown in Figure 7. It can clearly be seen that the OptiSim-based selection methods, which had shown a bias toward selecting less complex molecules, also deplete the fraction of highly active molecules in the selected subsets. However, the DivKM selection method has no discernible effect on the distribution of biological activity in the selected subsets. It can be seen that using SOM/RDF with decreasing subset size the number of highly active compounds increases slightly. Since the SOM/RDF method has not been included in the size bias study reported in the beginning of this paper, we wanted to assess whether bias toward high activity values can be the result of an underlying size bias. It was found that the full subset contains on average 31 non-H atoms per molecule, whereas the 7% subset has on average 35, the 3% subset on average 36, and the 1% subset on average 41 non-H atoms.

While this study does describe the effect of diversity selection on the average activity, it does not give an answer to the question as to whether a diversity selection has an effect on the chance to find ligands that bind to a specific target of interest. In an approach to answer this question we analyzed how well the diversity selections of the IC_{50} summary data cover the hit lists of each individual target. As the hit lists of the different targets were of very different sizes, the potential results from random selections was nonobvious. Therefore, in addition to the diversity selections as described above (OptiSim/UNITY, OptiSim/FCFP_4, and DivKM/FCFP_4) we used also selections of increasingly smaller size obtained by a random number generator for comparison. To ensure that sampling artifacts were minimized from random selection, each selection was conducted randomly five times and the results were averaged; however, the five selections were found to have very little variance.

We used two coverage criteria. The first, very simple criterion defines a target as covered if at least one molecule active on the target with a given activity threshold was contained in the diversity selection. Two activity thresholds were used, $IC_{50} \leq 1\ \mu M$ and $IC_{50} \leq 10\ \mu M$. The results were not found to be dependent largely on the chosen threshold, and thus we will describe only the results obtained for the 10 μM in the following. According to this criterion one could reduce the data set to 4% of its initial size without losing coverage for more than 5% of the targets (Figure 8a). This holds for any diversity selection method, and none of these methods yielded in this aspect better results than random selection. The second coverage criterion was applied to see how well the whole ligand space of the targets was sampled by the diversity selections. According to this second criterion a target was covered if for each of its hit series at least one representative was contained in the subsets. A hit series was defined as a set of molecules being active on the

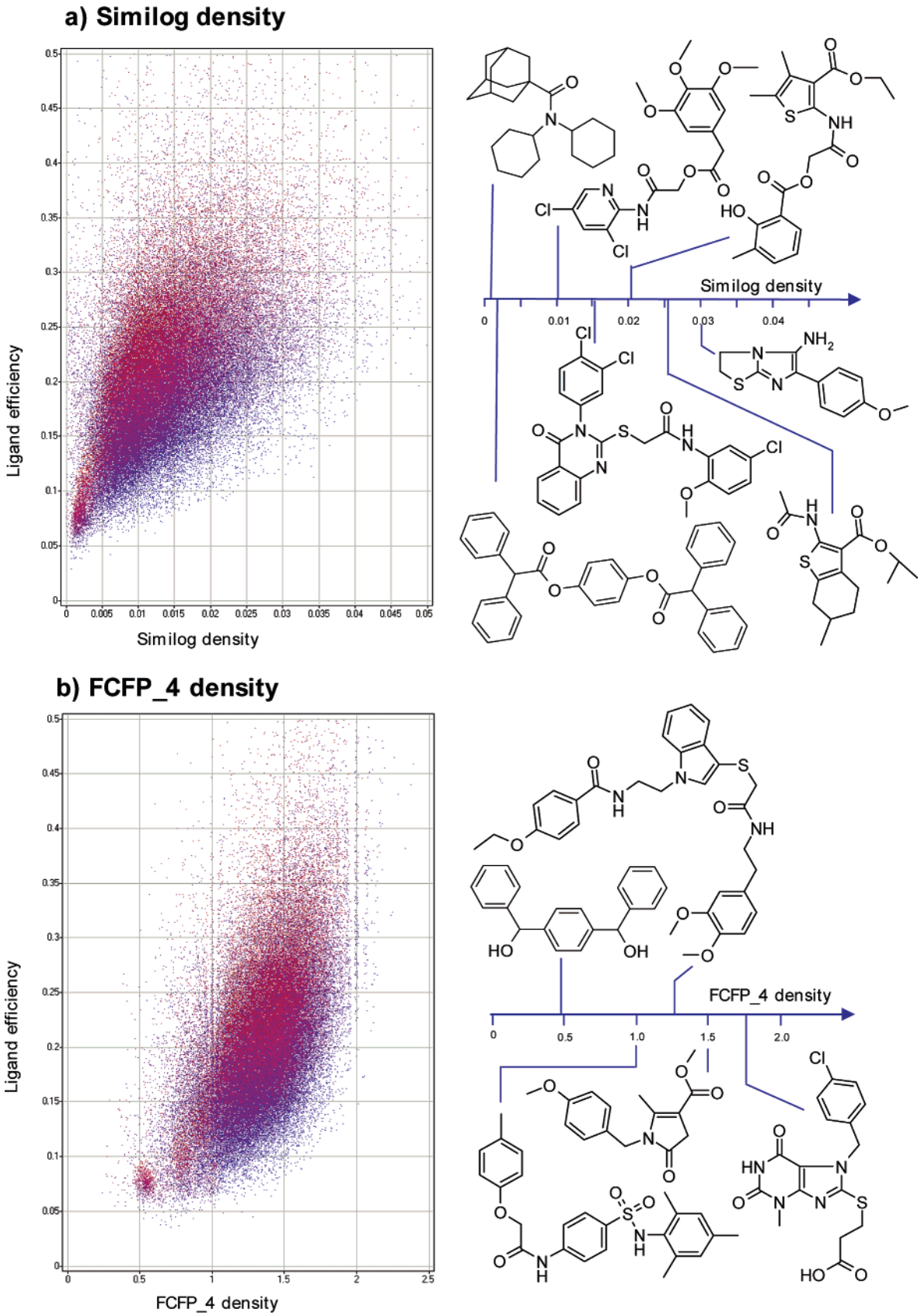

Figure 6. Relationship between maximal ligand efficiency ($-\log(IC_{50})/N_{non-H-atoms}$) and two density-based complexity measures. (a) Similog density (number of Similog keys normalized by the theoretical number of atom triplets). (b) FCFP_4 density (number of bits-set in FCFP_4 fingerprints normalized by the number of non-H atoms). From blue to red the absolute activity of the ligand increases. For each measure, some example molecules taken from commercial vendor catalogues are given, which illustrate the chemical meaning of the complexity measures.

target according to the activity threshold and sharing one common Murcko scaffold, which is defined as the core remaining after pruning all terminal side chains from the molecule.[38] Despite its widespread use, the Murcko scaffold

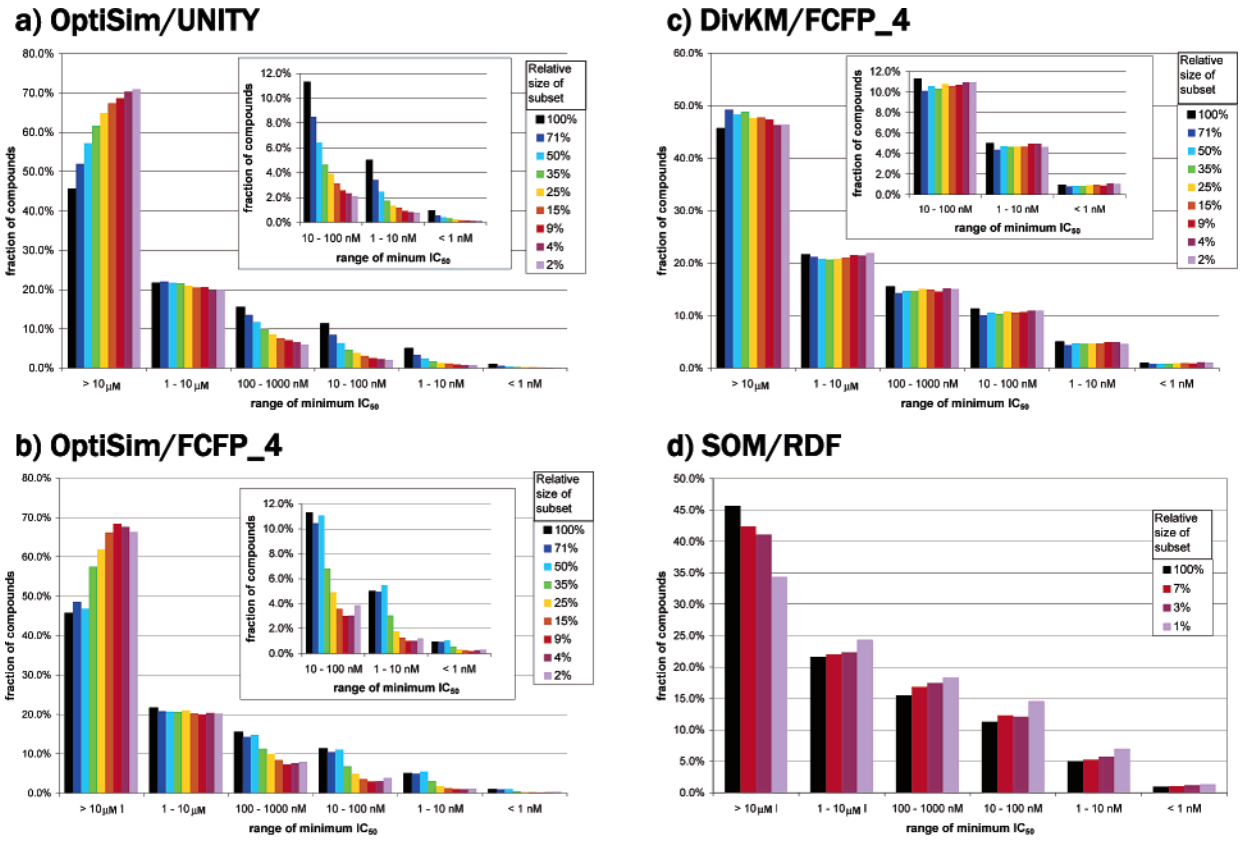

Figure 7. Influence of the diversity selection methods on the distribution of biological activity. Depicted are the distributions of the binned maximal $-\log(IC_{50})$ values of the molecules in subsets selected from the Novartis IC_{50} summary data file with decreasing size obtained by using increasingly tight diversity selection criteria. The insets show the distribution of the high affinity ligands with an enlarged *y*-axis.

definition works well only if the structures have a meaningful cyclic core; e.g. the Murcko scaffold from molecules like oligopeptides and their close analogues having an acyclic core scaffold containing acyclic or cyclic side chains is often meaningless and would lead to a high number of singleton scaffolds. To address problems that arise from this, we decided only to look at hit series that contained at least five members in the original data set. In this case one sees a rather rapid decrease in the fractions of targets covered, which is consistently observed with an increasingly tight diversity threshold and smaller subset size. Moreover, the fraction of targets covered decreased faster compared to random selection when OptiSim-based selection was used. The DivKM selection was only found to be slightly better than random (Figure 8b). As only series of more than five were taken into account, of which only one representative was required to be found in the subsets, a reduction to the size of 20% of the original set, without losing the coverage of any target, is theoretically possible. This benchmark was not reached by far. Reduction to 20% of the data set resulted in a coverage of only 8% to 20% of the targets.

DISCUSSION

The most prominent finding of this article is that there are diversity selection methods, namely those based on OptiSim and similar algorithms, which bias the subsets they generate toward lower molecular complexity. Interestingly, this is more dependent on the selection algorithm, since both of the OptiSim-like algorithms show this bias, than on the descriptor or similarity coefficient. With the same FCFP_4 descriptor used and the same Tanimoto similarity coefficient the divisive *K*-means selection does not introduce a size bias, whereas the OptiSim-like Pipeline Pilot clustering does. One reason for this may be the criterion used in divisive *K*-means of which cluster should be bipartitioned in the next iteration. Here, a decision is made solely on the size of the cluster and not on its homogeneity. This avoids the splitting clusters of rather dissimilar small molecules, which would then lead to the inclusion of more of these small molecules in the subsets, since from each cluster one sample would be taken. Another possible explanation could be the way each *K*-means clustering step works. Rather than comparing individual molecules by Tanimoto similarity, as with OptiSim-like algorithms, in the *K*-means step each molecule is compared with a consensus fingerprint representing the centroid of the cluster, thus removing the Tanimoto Coefficient's inherent bias toward small molecules in dissimilarity-based selection.[20] Coupled with the fact that highly active molecules are also likely to be more complex than the average inactive molecules, the decrease of complexity during diversity selections with OptiSim-like methods results in a depletion of active molecules in the subsets generated with such diversity selection procedures.

The small bias of the subsets generated with SOMs, toward higher biological activity is more difficult to explain. It is likely to be the case that the size bias of the selection method toward higher complex molecules is the cause of the bias in the activity range. However, since we have no comparison to a selection method without size bias using the RDF descriptor, we cannot conclude this with certainty. The size bias of the SOM toward more complex molecules can be explained by the different distance metric used, which is the

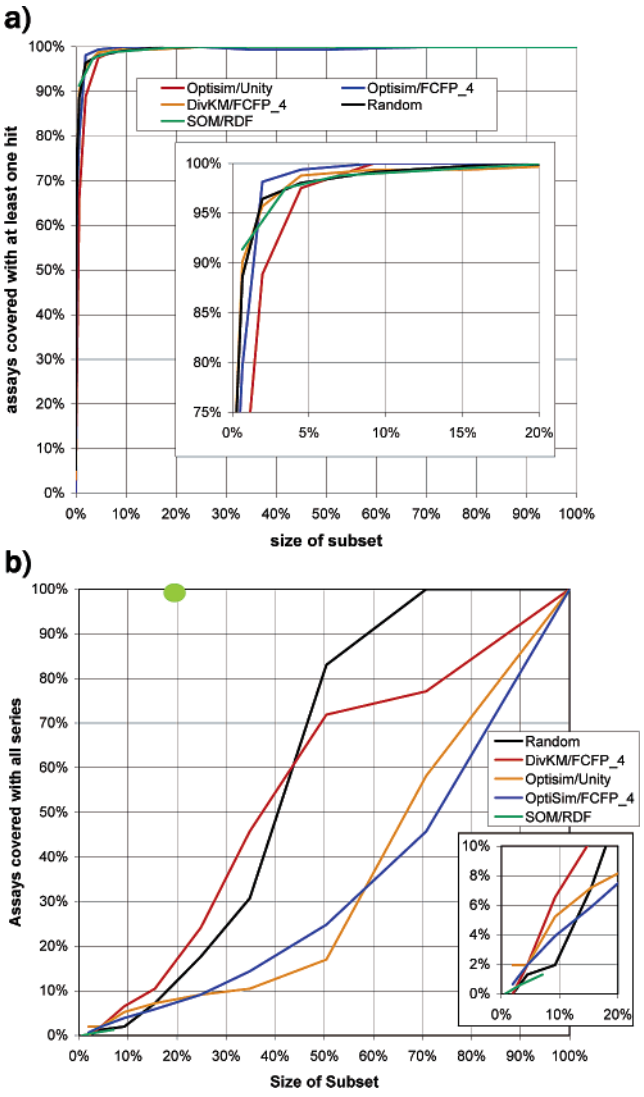

Figure 8. Coverage of the different activity types (assays) in the Novartis IC_{50} summary file in found in subsets obtained by diversity selection. Two coverage criteria were used. (a) An assay is regarded as covered when at least one hit with $IC_{50} \leq 10$ μM is contained in the subset. (b) An assay is regarded as covered if of all hit series sharing a common Murcko scaffold having at least five members with $IC_{50} \leq 10$ μM in this assay at least one representative is contained in the subset. This implies that a reduction of the whole set to 20% is theoretically possible without losing coverage on any assay (green dot).

Euclidean distance instead of Tanimoto dissimilarity. For binary fingerprint descriptors it is known that the Euclidean distance exhibits a higher apparent dissimilarity as the number of bits-set increases, which is a bias in the opposite direction to that of the Tanimoto dissimilarity.[20]

The bias toward lower complexity and biological activity of the OptiSim-based methods is an unfavorable side effect since the fraction of molecules having no IC_{50} below or equal to 10 μM can rise from 45% of the whole set to 70% in the selected subset. Clearly an unwanted bias on the molecular complexity by diversity selection is something one wants to avoid for this reason. There remains however the question, whether actively controlling the molecular complexity in screening sets can be useful, and at what complexity one should aim. The data shown here suggest at first glance that more complex molecules may be preferable, since highly active molecules are likely to be more complex than the average inactive molecule. Unfortunately this does not mean that the likelihood of a molecule being active in the first place increases in proportion with the molecular complexity. On the contrary, the likelihood that a molecule fits into the receptor binding site without any feature mismatches is likely to decrease with its complexity as pointed out by Hann,[14] and the difficulty to sample chemical ligand space completely increases with the maximum molecular size that one wishes to include.[15]

In combination, these results constitute a serious challenge: in order to identify ligands in an activity range that is suitable for the anticipated drug, we need to screen molecules in a complexity range where it is very unlikely that a molecule matches the binding site of a target protein with all its features. As shown in Figure 8, diversity selection is not a solution to this problem since the diverse subsets cover only very few assays with all major hit series. This is true for each of the diversity selection methods studied; including the divisive *K*-means, a selection method without complexity bias, and the selection by SOMs that tend to have a small bias toward higher activity compounds. There is hardly any improvement compared to a random selection of a subset. It can be argued that as long as one recovers at least one hit that this will be sufficient. In this aspect all the methods seem to perform well. However, typically the attrition rate from a screening hit to a lead is high, and therefore it is desirable to cover as many hit series as possible for each assay. It should also be noted that the IC_{50} summary data used here contains only assays where at least 1000 compounds have been validated, meaning that difficult assays yielding only very few hits in the primary screening stage have not been included. For such assays the probability that one hit is included in a diverse subset is much smaller. Even a moderate reduction of the compound set by the OptiSim/UNITY diversity methods using a maximum Tanimoto similarity constraint of only 0.80 leads only to subsets that still contain 40% of the original set. And yet the inherent assumption underlying all diversity selections, that all molecules being more similar to a selected molecule than the selected similarity constraint are equally active is, as Martin et al. have demonstrated,[2] only true for 15−50% of the molecules having a Tanimoto similarity ≥ 0.8 based on the Daylight fingerprint, which is very similar to the UNITY fingerprint used here.

This suggests that attempting to obtain a highly active ligand in one screening run with no prior knowledge regarding the target structure or known ligands might not always be the optimal strategy. According to Hann's model the chance of finding a binding ligand are higher when one initially aims to find a weakly binding ligand with low complexity. There are however several examples that such ligands have been discovered by the biophysical screening of small molecular fragments that could then be optimized to obtain ligands with higher activity.[39−41] A further approach is the so-called iterative or "smart" screening, where in a first step only a subset of the compound collection is screened. The outcome of this screen is then used to predict, either by similarity searching or statistical models, which of the remaining compounds are most likely to be hits.[42,43] Since statistical models for activity prediction such as Bayesian models[44] are often based on molecular fragment descriptors such as circular substructures, one can expect that a fragment

based screen as a first iteration should be especially suitable to predict further ligands since the contribution of a rather isolated fragment to the target protein binding can be observed. When building statistical models or running similarity searches on fragment descriptors, one generally assumes, in the absence of better knowledge, that all fragments of the molecules are contributing to its activity. This is however not necessarily the case, and inactive fragments that are attributed incorrectly as active can introduce statistical noise into the models. In molecules with high ligand efficiency one can expect that there are less of these noninteracting fragments present, and therefore such molecules should be especially suitable to train statistical models or run similarity searches. Complexity density measures such as the Similog density or the FCFP_4 density can help to select ligands which in the case they bind to the target at all are likely to have high ligand efficiency.

However, realizing the potential benefits from iterative screening and the use of individually focused compound sets for each target, as obtained by virtual screening, poses a challenge to current industrial HTS processes since they require higher flexibility in the screening process, more cherry-picking capacity to allow for a true random access to the compound collection, and also a higher integration of cheminformatics into the screening process.

ACKNOWLEDGMENT

The authors thank Peter Willett (University of Sheffield), Peter Fürst, Kamal Azzaoui, Christian Parker, Meir Glick, Phillip Floersheim, and Hans-Jörg Roth (all Novartis Institutes of BioMedical Research) for insightful discussions.

Supporting Information Available: List of vendor collections used in the section "Effect of Diversity Selection on Molecular Complexity". This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Maggiora, G. M.; Johnson, M. A. *Concepts and applications of molecular similarity*; John Wiley & Sons: New York, 1990; pp 99−117.
(2) Martin Y. C.; Kofron, J. L.; Traphagen L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(3) Jacoby, E.; Schuffenhauer, A.; Acklin, P. The contribution of molecular informatics to chemogenomics. Knowledge-based discovery of biological targets and chemical lead compounds. *Methods and Principles in Medicinal Chemistry;* 2004; Vol. 22 (Chemogenomics in drug discovery), pp 139−166.
(4) Van Drie, J. H. Pharmacophore discovery—lessons learned. *Curr. Pharm. Des.* **2003**, *9*, 1649−1664.
(5) 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 Discovery* **2004**, *3*, 935−949.
(6) Gillet, V. J. De novo molecular design. *Methods and Principles in Medicinal Chemistry*; 2000; Vol. 8 (Evolutionary Algorithms in Molecular Design), pp 49−69.
(7) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−63.
(8) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.
(9) Villar, H. O.; Koehler R. T. Comments on the design of chemical libraries for screening. *Mol. Divers.* **2000**, *5*, 13−24.
(10) Schuffenhauer, A.; Popov, M.; Schopfer, U.; Acklin, P.; Stanek, J.; Jacoby, E. Molecular diversity management strategies for building and enhancement of diverse and focused lead discovery compound screening collections. *Comb. Chem. High Throughput Screening* **2004**, *7*, 771−781.
(11) Lewis, R. A.; Pickett, S. D.; Clark, D. E. Computer-aided molecular diversity analysis and combinatorial library design. *Rev. Comput. Chem.* **2000**, *16*, 1−51.
(12) Martin, Y. C. Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.* **2001**, *3*, 231−250.
(13) Perez J. J. Managing molecular diversity. *Chem. Soc. Rev.* **2005**, *34*, 143−152.
(14) Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856−864.
(15) Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 225−263.
(16) Hann, M. M.; Leach, A. R.; Green, D. V. S. Computational chemistry, molecular complexity and screening set design. *Methods and Principles in Medicinal Chemistry*; 2005; Vol. 23 (Chemoinformatics in Drug Discovery), pp 43−57.
(17) Goodnow, R. A., Jr.; Gillespie, P.; Bleicher, K. Chemoinformatic tools for library design and the hit-to-lead process: a user's perspective. *Methods and Principles in Medicinal Chemistry*; 2005; Vol. 23 (Chemoinformatics in Drug Discovery), pp 381−435.
(18) Selzer, P.; Roth H. J.; Ertl, P.; Schuffenhauer, A. Complex molecules − do they add value? *Curr. Opin. Chem. Biol.* **2005**, *9*, 1−7.
(19) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(20) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819−828.
(21) Clark, R. D. OptiSim: An extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(22) Tripos, Inc. is at http://www.tripos.com.
(23) Pipeline Pilot is a product of SciTegic, Inc. is at http://www.scitegic-.com.
(24) 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.
(25) Steinbach, M.; Karypis, G.; Kumar, V. A. *Comparison of Document Clustering Techniques*; Technical Report 00-034; Department of Computer Science & Engineering: University of Minnesota, 2000.
(26) Bocker, A.; Derksen, S.; Schmidt, E.; Teckentrup, A.; Schneider, G. A Hierarchical Clustering Approach for Large Compound Libraries. *J. Chem. Inf. Model.* **2005**, *45*, 807−815.
(27) BCI (Barnard Chemical Information Ltd.) is at http://www.bci.gb.com.
(28) Allu, T. K.; Oprea, T. I. Rapid Evaluation of Synthetic and Molecular Complexity for in Silico Chemistry. *J. Chem. Inf. Model* **2005**, in press. doi 10.1021/ci0501387.
(29) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269−1275.
(30) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.
(31) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997−10002.
(32) Hopkins, A. L.; Groom, C. R. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430−431.
(33) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectrosc.* **1999**, *19*, 151−164.
(34) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Method.* **1990**, *3*, 537−547.
(35) CORINA is a product of Molecular Networks, GmbH at www.mol-net.de.
(36) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design;* Wiley-VCH: Weinheim, 1999.
(37) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429−434.
(38) Bemis, G. W.; Murcko, M. A. The Properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(39) Rees, D. C.; Congreve, M.; Murray C. W.; Carr, R. Fragment based lead discovery. *Nat. Rev. Drug. Discovery* **2004**, *3*, 550−672.
(40) Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-based drug discovery. *J. Med. Chem.* **2004**, *47*, 3463−3482.
(41) Schuffenhauer, A.; Ruedisser, S.; Marzinzik, A. L.; Jahnke, W.; Blommers, M.; Selzer, P.; Jacoby, E. Library Design for Fragment Based Screening, *Curr. Top. Med. Chem.* **2005**, *5*, 751−762.

(42) Engels, F. M.; Venkatarangan, P. Smart screening: Approaches to efficient HTS. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 275−283.
(43) Baringhaus, K.-H.; Hessler, G. Fast similarity searching and screening hit analysis. *Drug Discovery Today: Technol.* **2004**, *1*, 197−202.
(44) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of Extremely Noisy High-Throughput Screening Data Using a Naïve Bayes Classifier. *J. Biomol. Screen.* **2004**, *9*, 32−36.

CI0503558