# Critical Comparison of Virtual Screening Methods against the MUV Data Set

Pekka Tiikkainen,*[,†,‡,#] Patrick Markt,[§,#] Gerhard Wolber,[§] Johannes Kirchmair,[§] Simona Distinto,[§] Antti Poso,[||] and Olli Kallioniemi[†,‡]

University of Turku and VTT Medical Biotechnology, Itäinen Pitkäkatu 4 C, FI-20521 Turku, Finland, FIMM Institute for Molecular Medicine, Tukholmankatu 8, FI-00290 Helsinki, Finland, Department of Pharmaceutical Chemistry, Faculty of Chemistry and Pharmacy and Center for Molecular Biosciences (CMBI), University of Innsbruck, Innrain 52, A-6020 Innsbruck, Austria, and Department of Pharmaceutical Chemistry, University of Kuopio, Yliopistonranta 1 C, FI-70211 Kuopio, Finland

Received July 14, 2009

In the current work, we measure the performance of seven ligand-based virtual screening tools - five similarity search methods and two pharmacophore elucidators - against the MUV data set. For the similarity search tools, single active molecules as well as active compound sets clustered in terms of their chemical diversity were used as templates. Their score was calculated against all inactive and active compounds in their target class. Subsequently, the scores were used to calculate different performance metrics including enrichment factors and AUC values. We also studied the effect of data fusion on the results. To measure the performance of the pharmacophore tools, a set of active molecules was picked either random- or chemical diversity-based from each target class to build a pharmacophore model which was then used to screen the remaining compounds in the set. Our results indicate that template sets selected by their chemical diversity are the best choice for similarity search tools, whereas the optimal training sets for pharmacophore elucidators are based on random selection underscoring that pharmacophore modeling cannot be easily automated. We also suggest a number of improvements for future benchmark sets and discuss activity cliffs as a potential problem in ligand-based virtual screening.

## INTRODUCTION

Although the amount of crystallographic data for 3D structures of proteins increases every year, only a few structures of membrane-bound proteins which represent important targets for drug discovery have been identified yet.[1] Thus, the starting point for many virtual screening campaigns is only a set of known ligands for a specific target and not the 3D structure of the protein. To cope with this virtual screening problem, several ligand-based approaches that do not depend on the availability of 3D protein structures have been employed successfully to enrich active compounds within chemical databases for subsequent high-throughput screening (HTS) of a focused compound library or for biological testing of a small set of compounds with promising lead potential. In a HTS campaign the only virtual screening requirement is to discriminate between active compounds and decoys (inactive compounds which are structurally similar to active molecules) better than a random selection, whereas for lead discovery, where only the top-scored 1−10% of the whole chemical database is biologically evaluated, ligand-based approaches should be able to enrich most of the active compounds within the top of a rank-ordered hit list. For the latter virtual screening problem, it is of utmost importance to know if the applied ligand-based method is able to assign top ranks to active compounds during virtual screening or if the method performs only comparable to random compound selection.[2−6]

Several studies have been performed to determine which ligand-based approach is the best performing method for which virtual screening problem. Most of them are retrospective studies which measure the performance of different virtual screening methods in proprietary data, commercial (MDDR[7]), or public (DUD[8]) compound databases.[2,9,10] However, such data sets often suffer from a low compound diversity. This so-called "analogue bias" leads to artificially high enrichments if similarity-based virtual screening techniques such as 2D fingerprints are validated. In general this is not considered to be a problem for molecular docking approaches,[11] but this notion has been recently challenged.[12] Thus, a good validation data set for ligand-based virtual screening techniques should consist of structurally diverse actives which are well embedded in sets of structurally diverse decoys. Rohrer and Baumann[13] claim that their recently published Maximum Unbiased Validation (MUV) Data Sets derived from PubChem[14] bioactivity data fulfills this criterion and consequently should be usable for the validation of ligand-based virtual screening methods.[11,15]

We used the MUV data sets to evaluate the discriminatory power of seven virtual screening methods, including five similarity search tools and two pharmacophore elucidators. Single and multiple queries were used as templates for screening the MUV. Consecutively, the enrichment within the top 5% of the retrieved hit list was determined. We also combined results from the five similarity search tools to see

* Corresponding author phone: +358 400 399 645; e-mail: Pekka.tiikkainen@utu.fi.
† University of Turku and VTT Medical Biotechnology.
‡ FIMM Institute for Molecular Medicine.
§ University of Innsbruck.
|| University of Kuopio.
# These authors contributed equally.

10.1021/ci900249b CCC: $40.75 
Published on Web 10/02/2009

**Table 1.** MUV Ligand Sets Used in the Study[a]

| target | mode of interaction | target class | primary assay (AID) | confirm. assay (AID) | assay-type | scaffolds |
|---|---|---|---|---|---|---|
| S1P1 rec. | agonists | GPCR | 449 | 466 | reporter gene | 28 |
| PKA | inhibitors | kinase | 524 | 548 | enzyme | 27 |
| SF1 | inhibitors | nuclear receptor | 525 | 600 | reporter gene | 24 |
| Rho-Kinase2 | inhibitors | kinase | 604 | 644 | enzyme | 27 |
| HIV RT-RNase | inhibitors | RNase | 565 | 652 | enzyme | 27 |
| Eph rec. A4 | inhibitors | rec. tyr. kinase | 689 | 689 | enzyme | 29 |
| SF1 | agonists | nuclear receptor | 522 | 692 | reporter gene | 30 |
| HSP 90 | inhibitors | chaperone | 429 | 712 | enzyme | 27 |
| ER-a-Coact. Bind. | inhibitors | PPI | 629 | 713 | enzyme | 26 |
| ER-$\beta$-Coact. Bind. | inhibitors | PPI | 633 | 733 | enzyme | 28 |
| ER-a-Coact. Bind. | potentiators | PPI | 639 | 737 | enzyme | 28 |
| FAK | inhibitors | kinase | 727 | 810 | enzyme | 28 |
| Cathepsin G | inhibitors | protease | 581 | 832 | enzyme | 24 |
| FXIa | inhibitors | protease | 798 | 846 | enzyme | 21 |
| FXIIa | inhibitors | protease | 800 | 852 | enzyme | 24 |
| D1 rec. | allosteric modulators | GPCR | 641 | 858 | reporter gene | 24 |
| M1 rec. | allosteric inhibitors | GPCR | 628 | 859 | reporter gene | 29 |

[a] Primary Assay identifier is the PubChem assay ID for the primary high throughput screen where both active and decoy molecules had been tested. Confirmatory Assay identifier refers to PubChem assay record for the confirmatory assay of the active molecules found in the primary screen. The column entitled "scaffolds" gives the number of distinct Murcko scaffolds in the set of active molecules.

if similarity fusion[16,17] improves results. This early enrichment should give an answer to the immanent question, which is the optimal virtual screening solution for lead discovery if only the structure of a few active compounds and no published decoys are available.

## MATERIALS AND METHODS

**MUV Data Sets.** The active and decoy molecules of the MUV (Maximum Unbiased Validation) data set[13] were downloaded from the project's home page at http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html (accessed October 2, 2008). The data set comprises confirmed active molecules and decoys for 17 target classes (Table 1) taken from NIH's PubChem assay database.[14] For each target class there are 30 active molecules and 15,000 decoys. Authors of the MUV data set had chosen the active molecules so that they occupy different areas of chemical space as defined with simple chemical properties such as heavy atom count and hydrogen bond donors and acceptors. In contrast, decoys that resemble the active molecules with respect to these simple properties were chosen. This process led to data sets where active molecules cannot be separated from decoys using simple properties. An advanced virtual screening tool is expected to succeed in this - otherwise the tool would be of little use taking into account its added complexity and strain on computational resources compared to using a vector of simple properties.

*A) Similarity Search Methods.* The performance of five similarity search methods Brutus, ROCS, FCFP4, ECFP4, and EON was investigated using single templates as well as multitemplate sets derived from the active compounds of the MUV data sets.

*3D Shape and Electrostatic Similarity Tools.* Brutus version 0.8.7[18,19] performs an exhaustive overlay of two 3D conformers by scoring complementarity of the grid-based electrostatic fields around the molecules. First a rectilinear grid is built around both molecules being compared. Electrostatic potential is calculated at each node for both grids. Overlay is done by keeping template molecule's grid static and rotating database molecule's grid. After each rotation step the two grids are compared by mapping a node of a grid to the closest node in the other grid. Shape complementarity is taken into account by assigning positive charge to points inside van der Waals radii of atoms while using the actual electrostatic values outside.

ROCS[20] performs shape-based overlay with atom-centered Gaussian functions to represent molecule shape. In addition to the pure shape similarity, a color score is given to represent complementarity of functional groups. Using the sum of the two scores represented by the comboscore as implemented in ROCS has been shown to lead to improved results compared to using shape complementarity alone,[21] and therefore it was used in the current study. One more similarity measure, EON,[22] was applied to quantify electrostatic field similarity of the molecule pairs first overlaid with ROCS. For scoring, ET_combo which is the default score of EON and sums up the Poisson−Boltzmann electrostatic Tanimoto coefficient (ET_pb) and the shape Tanimoto (ST) was used in the current work.

3D overlay tools described above require good quality conformers of the molecules. Thus, we applied the default settings of the conformer generator OMEGA2[23] to create the multiconformer ensembles for each active and decoy molecule using the 2D atom connectivity data downloaded from the MUV Web site as input. For each active compound, the conformer with the lowest energy was used as template.

*2D Fingerprints.* Chemical similarities were also calculated using the extended connectivity fingerprints ECFP4 and the functional class fingerprints FCFP4. The two fingerprints are generated by iteratively inspecting the atom neighborhood of each non-hydrogen atom up to a predetermined diameter (four atoms in this case). Each individual atom neighborhood found leads to a specific fingerprint bit turned on. Tanimoto score was used to quantify similarity of the fingerprints. Given the scaffold diversity of the MUV data sets, it was interesting to see how well these fast fingerprints compare to the computationally more demanding 3D overlay methods. Both the FCFP4 and ECFP4 calculations were performed with Pipeline Pilot Student Edition.[24]

Additionally, combined performance of these similarity search methods was assessed with data fusion (similarity fusion). Two commonly used and simple data fusion rules were employed, min_rank (analogous to MAX rule) and avg_rank (analogous to the SUM rule). Using the former rule, each molecule is given the smallest (minimum) rank it has on any of the lists being merged. With the latter rule, the compound is given the average of its ranks in the lists. The molecules are then reranked according to the new rank to generate the fused list. In earlier work[21,25] the two data fusion strategies have been shown to lead to improved and more consistent results.

*Single Templates for the Similarity Search Tools.* Using each active molecule as template, the chemical similarity between this single template and the other active molecules and all decoys in its ligand group was calculated applying the five similarity search tools.

*Multitemplate Sets for the Similarity Search Tools.* Two strategies were employed for picking the template sets for multitemplate data fusion (group fusion): random and diversity-based. In the case of random selection, for each combination of target class and template set size (ranging from two to ten), templates were picked 100 times by random.

The PAM (Partitioning Around Medoids)[26] algorithm was used for diversity-based set selection. Within each target class a set of distance matrices of active compounds were generated, one for each similarity method (min_rank and avg_rank were considered as independent similarity methods). To calculate the distances, similarity values were subtracted from the maximum score of the corresponding method (with the exception of min_rank and avg_rank for which no modification was needed). Each distance matrix was imported into statistical software R where it was given as input for the PAM algorithm. The algorithm was run systematically to pick from one to ten cluster centers (centroids) which constitute the template set.

For both random and diversity-based template selections, the MAX rule was used to merge the hit lists obtained for the individual templates of one set.

*Performance Metrics.* The popular performance metric enrichment factor (EF) was used in the present study. The number of active molecules found in a given top % fraction of the hitlist is related to the number of active molecules expected by random (eq 1)

$$EF_{x\%} = \frac{Act\_found}{Act\_total * \left(\frac{x}{100}\right)} \quad (1)$$

where *Act_found* is the number of active molecules found in the top *x*% of the ranked result list, and *Act_total* is the total number of active molecules in the ligand set. In this paper, we report the enrichment results for the top 5% of a virtual screening hitlist ranked by the score of the similarity search tools (EF5 value).

*Pubchem Assay Data Sets.* One possible explanation for failure of similarity search methods to separate decoys from actives (see Results and Discussion) are so-called activity cliffs.[27,28] These are small differences in ligand structure that lead to major differences in activity. To estimate the frequency of activity cliffs in our results, we calculated the frequency that a template active molecule and a high ranking decoy molecule are both active against other targets.

For this purpose, we downloaded results for 391 bioassays from NIH's Pubchem service[14] (accessed April 15, 2009). Only confirmatory assays with a defined protein target were considered, as primary screens contain too many false positives and therefore would add noise to the analysis. We also ignored growth inhibitory cell assays as there the exact molecular target is not defined. The list of bioassays analyzed here can be found in Table S1 in the Supporting Information.

Afterward, all active template - decoy pairs were retrieved from the top 1% of ranked hitlists of any method. For each compound pair, we calculated the number of assays where both compounds had been tested and also the number of assays where both were deemed active. The ratio of these gives us an idea how common activity cliffs are.

*B) Pharmacophore Elucidators.* Training sets for pharmacophore generation comprised active molecules selected from the MUV data sets. As for the five similarity search methods, conformational ensembles for the MUV active compounds and decoys were generated applying the default settings of OMEGA2.[23] The HipHop algorithm[29] as implemented in the software package Catalyst 4.11[30] and the Pharmacophore Elucidator algorithm of MOE 2007.09[31] were used for pharmacophore elucidation. HipHop identifies common feature configurations by aligning the conformations of the training set compounds with respect to their chemical features. The algorithm starts with the generation of pharmacophore models containing two features and increases the number of chemical features until no common pharmacophore can be found for the training set compounds. This pharmacophore generation process results in models that match at least one conformation of each of the training set compounds.[30] The MOE Pharmacophore Elucidator creates common feature pharmacophore models in a similar way. It initially generates pharmacophore models based on two features and uses the most popular ones to add an additional feature, thereby increasing the number of features step by step. In contrast to HipHop, the resulting pharmacophore models do not have to match all training set compounds but only a specified percentage. By default this coverage of active compounds is set to equal or more than 90% of all compounds.[31] Using the default chemical feature sets, both pharmacophore modeling algorithms contain features for hydrogen bond acceptors, hydrogen bond donors, positive and negative charges, positive and negative ionizable moieties, hydrophobic and aromatic interactions.

The generated ligand-based models were virtually screened against the remaining active molecules and the decoys of the MUV ligand group containing the training set compounds. The quality of the alignment between the virtual screening hit and the pharmacophore model was scored using the fit value in Catalyst and the rmsd score in MOE.[30,31] The Catalyst fit value is the sum of single alignment values calculated for each feature of the pharmacophore model. Each single value is composed of the feature weight and the distance between the feature center and the corresponding chemical moiety of the mapping compound. Since the feature weight is one by default, a fit value of five for a five-feature pharmacophore model means that all five features are perfectly aligned to the chemical moieties of the compound. Whereas fit values lower than five indicate that at least one

feature center does not optimally map the compound.[30] MOE uses a similar score which calculates the root of the mean square distance between the pharmacophore features and the corresponding moieties of the compound.[31]

Ligand-based pharmacophore modeling was performed in a fully automated workflow. A sequence of perl, shell, and MOE SVL scripts was responsible for the generation of conformational models for the molecules of the MUV data sets as well as for pharmacophore elucidation and virtual database screening.

Default settings were used for pharmacophore elucidation and virtual database screening. However, to obtain comparable results for both ligand-based pharmacophore modeling algorithms, we changed some of the default parameters.

First, Catalyst stores only the fit value of the best matching conformation of a compound in a hit list, whereas MOE exports the rmsd score for all matching conformations by default. For competitive conditions, we added a Perl script to our MOE modeling workflow which, like Catalyst, just reports the rmsd score of the best matching conformation for each compound.

Second, the default Catalyst database search hit list is limited to 300 matching compounds. Since MOE does not restrict the number of virtual screening hits by default, we set the Catalyst parameter 'ViewDatabase.maxHits' to 15030, which allows the retrieval of all compounds if one of the 17 MUV ligand groups is virtually screened.

*Selection of Training Sets for Pharmacophore Modeling.* In accordance with the experimental setup for the enrichment assessment of the similarity search tools, we analyzed the correlation between the use of different training sets for model generation and the enrichment of active compounds in the MUV data sets. In contrast to the similarity search tools mentioned above, the use of a single active molecule for the generation of HipHop or MOE pharmacophore models resulted in no enrichment. This can be explained by the fact that the structural information of only one active molecule leads to the generation of highly restrictive models describing a multitude of chemical features of the molecule. Thus, the pharmacophore models were too restrictive to match any compound of the MUV data sets except for the active molecule used for their generation. On that account, we focused our enrichment study on common feature-based models derived from training sets containing two to ten active molecules. Analogous to the study design for the similarity search methods, training sets were either built up by random- or diversity-based selection.

For each combination of target class and training set size we first generated 100 pharmacophore models based on randomly selected training set compounds.

Second, diversity-based training sets were created. For this purpose, we applied distance matrices which describe the accuracy of the pharmacophore-based pairwise alignment of the 30 active molecules of each MUV target class. Based on these distance matrices two types of diversity-based training sets were created. The first training set type contained the active molecules with the maximum mutual similarity within one MUV target class. This was done by a series of SQL commands against a table with pharmacophore similarity values. The second training set type included active molecules comprising the most diverse pharmacophoric patterns. This set type was generated by employing the PAM algorithm[26] as described for the five similarity search tools.

The splitting of the training sets in similar and diverse sets of active molecules should answer the question if a good performing 3D pharmacophore model should be based either on the alignment of ligands with mutual similarity in terms of their chemical feature patterns or on compounds comprising diverse pharmacophoric patterns. In theory, both approaches to select training sets imply the potential for a successful virtual screening campaign or for a complete failure. For example, a training set comprising the most similar active compounds could lead to the generation of a highly restrictive model, which could accurately discriminate between active compounds and decoys. However, such a model could match only a few active molecules of the compound database and omit the majority of promising active molecules. In contrast to that, a more general model created by using the most dissimilar active molecules (in terms of their pharmacophoric patterns) could contain only a limited number of common chemical features. Hence, it would match most of the active molecules of a database but could also retrieve a large number of decoys and thereby decreasing the enrichment of active compounds.

As for the performance assessment of the similarity search methods, the EF5 value was applied to measure the enrichment of active compounds obtained for the two pharmacophore modeling methods.

## RESULTS AND DISCUSSION

**A) Enrichment Assessment of Similarity Search Methods.** *Similarity Search Tools with a Single Template Molecule.* The heatmap in Figure 1 visualizes the EF5 values for using single templates. The 510 active molecules in the MUV set are clustered on rows, while the five virtual screening methods and the two data fusion methods are on the columns. At a quick glance, it is apparent that in most cases, the enrichment of active molecules is very low.

The heatmap can be divided vertically in five large segments. Starting from the bottom, segment 1 contains a group of templates where all methods give moderate to high enrichment of active compounds. Moving up, segment 2 comprises templates where the two fingerprint methods FCFP4 and ECFP4 perform well, while the performance of the 3D overlay tools drops. The majority of the templates falls in the central region where the performance of all methods is mediocre at best (segment 3). Top of the heatmap is divided into two segments. In segment 4, the three overlay methods give some enrichment, while the fingerprint tools perform worse. In segment 5, the only 3D overlay tool still performing somewhat is ROCS, while the two fingerprint tools both lead to some enrichment.

The heatmap also gives interesting information on the overlap of the virtual screening results of the five similarity search tools. To no surprise the two fingerprint tools cocluster. More interestingly, the closest method to EON is BRUTUS and not ROCS as one might initially expect as ROCS is used to derive the overlays for EON. This can be explained by the fact that the pharmacophoric complementarity is defined in both tools with electrostatic fields rather than the color field employed in ROCS.

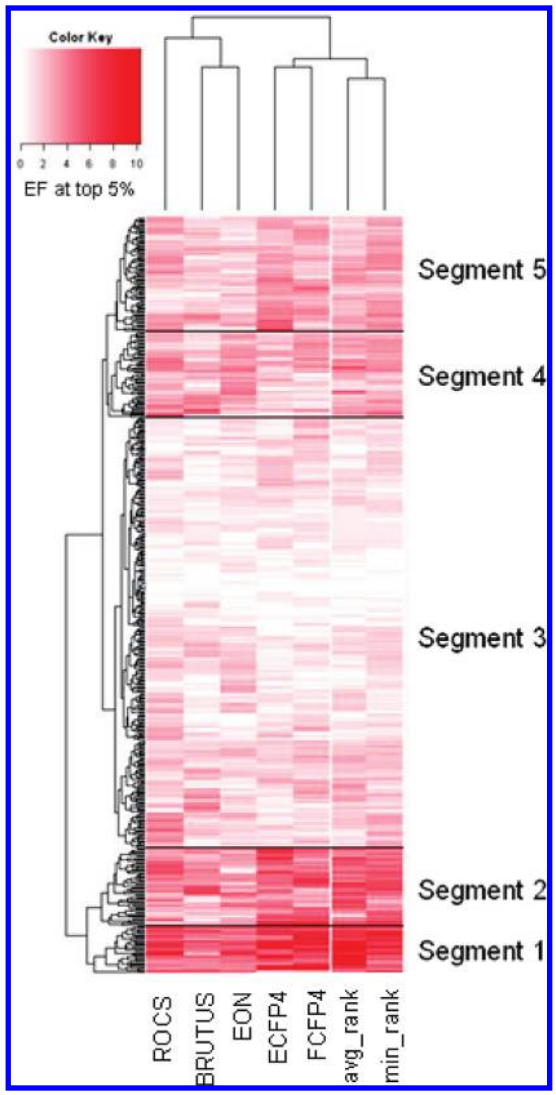


**Figure 1.** All 510 MUV active molecules used as templates are clustered on the vertical dendrogram according to their enrichment factor (top 5%) across the similarity search methods. The five similarity search and two data fusion methods are clustered horizontally according to the enrichment factor (top 5%) across the templates.

Another important note in Figure 1 is the consistent performance of the two similarity fusion methods. Whereas performance of individual methods varies, both min_rank and avg_rank rules produce moderate to high enrichments virtually everywhere except segment 3. This supports the case for data fusion, as it is largely independent of shortcomings with individual methods in various cases.

Column diagram in Figure 2 gives the average enrichment of the methods across the 17 ligand classes. Here, it is apparent that there are great differences in performance between the classes. The methods perform well in only four classes (aid548, aid832, aid846, and aid852) with the two data fusion methods usually giving the best results in most of the classes. Performance seems to be inversely proportional to the number of scaffolds in the ligand class as the three latter classes all have the lowest count of distinct Murcko scaffolds[32] in the whole data set (Table 1). This means that none of the methods is able to perform scaffold hopping to a large extent within the more diverse data sets which is rather surprising. After all, the main reason for using

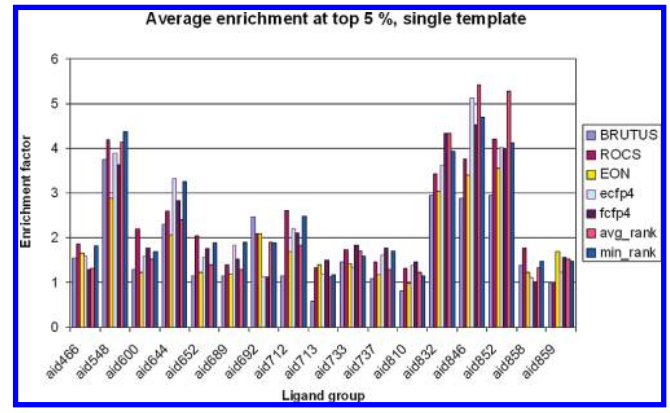


**Figure 2.** Average single template enrichment factors at the top 5% of the ranked list across ligand groups.

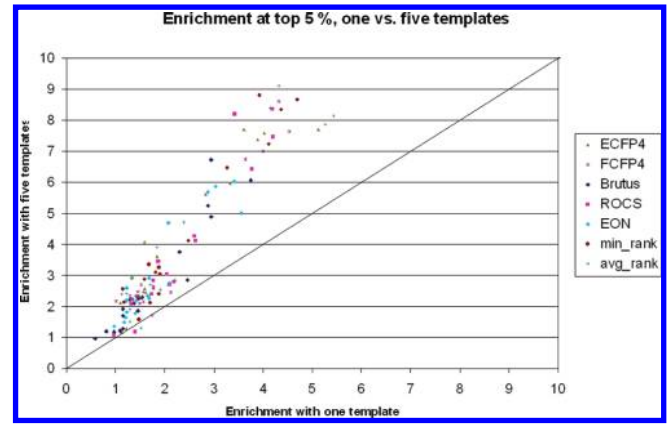


**Figure 3.** Combinations of a ligand group and similarity search method plotted according to their average enrichment factor (top 5%) with single templates (*x*-axis) and the average enrichment factor (top 5%) across the 100 random five-membered template sets.

3D overlay tools such as ROCS is their alleged capability to find novel chemotypes.[33]

The observations reported above do not depend on the performance metric. The data for other enrichment factor cut off points and ROC AUC are given in Table S2 in the Supporting Information.

*Use of Multiple Templates.* Using not only one but several template molecules (also called group fusion) has been shown to improve retrieval of actives.[34] Therefore, we wanted to find out if using several templates results in any improvement compared to the single template case presented above.

Figure 3 clearly exemplifies how enrichment measured at the top 5% improves when five randomly chosen template molecules are used instead of just one. Data point values are averages of 100 random template selections for each combination of method and ligand class. However, using several templates is no universal panacea: if no enrichment is observed for a single template, the corresponding multitemplate value is not significantly better as demonstrated by the data points lying close to the diagonal in the bottom left of Figure 3.

In order to assess how template set picking strategy affects results, we also used PAM (Partioning Around Medoids) clustering to pick a diverse subset of actives. Figure 4 illustrates the effect the strategy has on performance for the shape overlay tool ROCS with template set size of five. With almost all the ligand classes, the PAM template set outperforms the average random case and in one case (aid846) gives superior performance to the best of the 100 random sets.

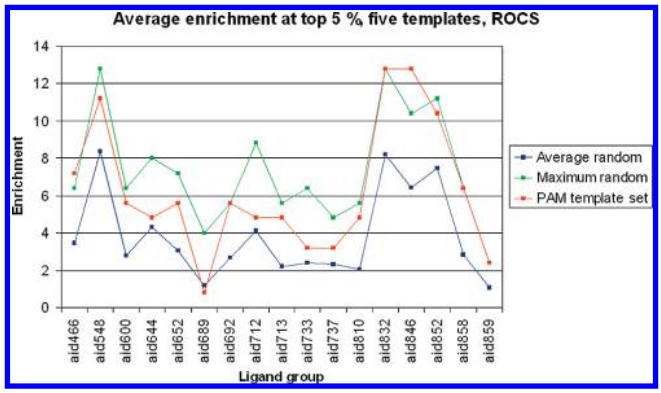


**Figure 4.** Comparison of template set picking strategies for ROCS. The dark blue line gives the average enrichment factor (top 5%) for the 100 randomly picked five-member template sets. The green line gives the maximum enrichment achieved with any of the random sets. The orange line depicts the enrichment with the template set selected by diversity (using PAM clustering).

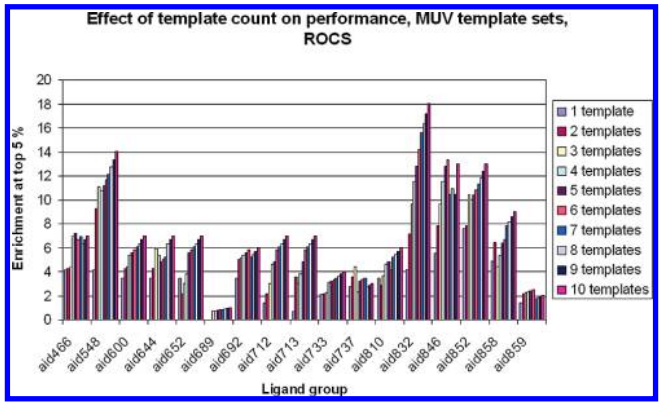


**Figure 5.** ROCS enrichment factors (top 5%) for diversity picked template sets with one to ten members.

Improved results with the diverse template set are hardly surprising when one considers that a larger portion of chemical space is represented by the templates than in the average random case. This was already noticed by Hert et al.[35] who showed that group fusion's relative performance to a simple search (one template) grows as the mean pairwise similarity of the active set decreases. It should be noted however that picking the templates diversely with the PAM algorithm is not an absolutely optimal approach since there are still randomly chosen subsets that perform better as seen in Figure 4. It is also generally true that performance improves as more templates are used. This is shown in Figure 5 for ROCS where enrichment at the top 5% is plotted as a function of the diverse template count for each ligand group. These observations hold also for the other four similarity search tools and the two similarity fusion methods. Plots for them are given in Figures S5−S10 of the Supporting Information.

*Pitfalls of Similarity Search Methods.* It is rather surprising to see that the similarity search methods fail in so many target classes. There are at least three possible reasons for this: false negatives, different binding (sub)cavities, and activity cliffs. In the following each of these issues is discussed.

The first of these issues was already discussed in the original MUV article.[13] They had done a small-scale literature review for the very top decoys in each target group and had not found any evidence for the decoys being active against the target in question. For a more quantitative and

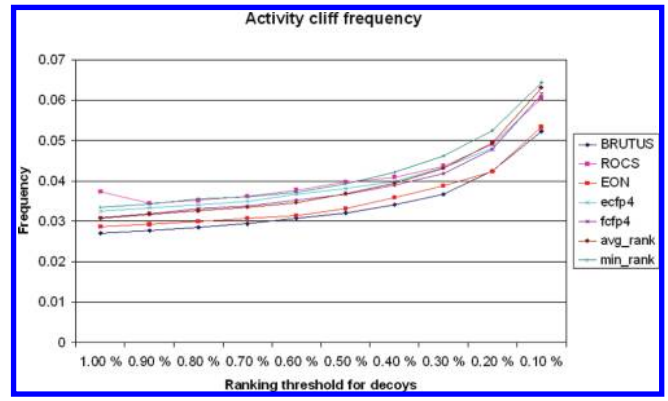


**Figure 6.** Frequency across similarity search methods by which the active compound used as template and a decoy highly ranked in the template's hit list are both reported as active molecules in at least one bioassay in Pubchem. Going from left to right, more highly ranked decoys are considered.

reliable assessment of the false negative rate, a confirmatory screen for the top scoring decoys would be needed. However, as mentioned in the original article, the situation is still better with MUV than many other benchmark sets where the decoys are just *assumed* to be inactive.

Different ligands may bind the same protein target in different ways, occupying only partially overlapping parts of the same binding cavity or by acting allosterically binding a completely separate cavity. All similarity search tools studied here assume the two molecules being compared to occupy the same space in the cavity. If this is not the case, poor performance should not be surprising. It can also be argued that the improved results we observed with the use of multiple templates is because each template molecule represents a group of ligands binding a given (sub)cavity. One more confusing factor is the assay type. MUV contains five target classes which were assayed with a reporter gene assay (Table 1). As these are cell-based screens, it cannot be ruled out that the observed activity would not be due to the ligand binding a different target than the one intended.[36] Interestingly, none of the reporter gene assay-based target groups were among those where similarity search methods were successful. This raises the concern that the poor performance is due to off-target binding.

The third Achilles' heel for similarity search methods are the so-called activity cliffs as mentioned above in the materials and methods paragraph.[27,28] Such activity cliffs can arise when a bulky side group in the decoy clashes sterically with the protein. As these small structural differences do not dramatically change other properties of the ligand, such as steric volume and overall atom connectivity, a similarity search method still scores the two molecules as highly similar leading to high ranks for the decoys.

To analyze effect of activity cliffs, we counted how often both the active template and a high-ranked decoy had been found to be active in other assays available at the Pubchem Web site (see Materials and Methods for details). Although a particular structural modification might render the decoy molecule inactive against one target, this is not necessarily the case with other targets where the protein can better accommodate the additional fragment.

Results of the analysis are given in Figures 6 and 7. The higher the rank of the decoy is, the greater the probability that the active molecule used as template and the decoy

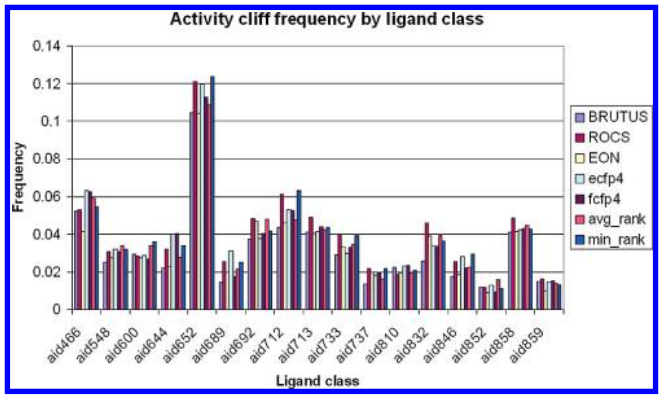


**Figure 7.** Frequency of putative activity cliffs (template-decoy pairs active in at least one Pubchem bioassay) by ligand group. Only pairs with the decoy in the top 0.5% of its respective hit list are considered.

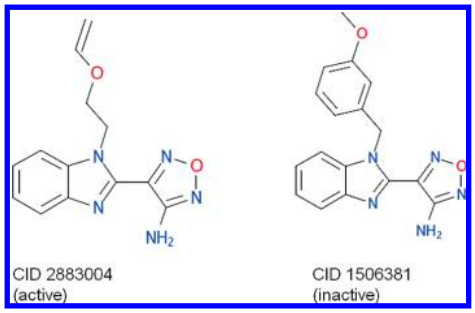


**Figure 8.** A putative activity cliff pair. The compound on the left (CID 2883004) is active against the Rho-Kinase 2 (Pubchem Assay ID 644), while the structural analog on the right (CID 1506381) had been found inactive against the kinase. However, the two molecules are both active in three other Pubchem assays.

molecule have been deemed active in at least one of the 391 confirmatory bioassays (Figure 6). This effect is the same for all methods studied here.

Figure 7 in turn gives us the same information across the 17 MUV target classes. Decoys considered here are in the top 0.5% of the hit list. The bars show the share of active-decoy pairs where both molecules had been found active in the same assay. Differences between target classes are considerable, meaning that there are targets where activity cliffs are more common than in others. The tallest bars are for aid652 (HIV RT-RNase inhibitors) which also is one of the groups where all similarity search tools perform poorly. It appears that this particular target is especially sensitive to small changes in ligand structure.

An example of a putative activity cliff is given in Figure 8. The figure illustrates a validated Rho-Kinase 2 inhibitor (CID 2883004) and a structurally very similar decoy molecule (CID 1506381). The decoy is the top ranking molecule for three methods (ROCS, ECFP4, EON) and the second ranking with BRUTUS and FCFP4 when CID 2883004 is used as template. Apparently, the more bulky side group of the decoy molecule leads to a steric clash with Rho-Kinase 2. This does not seem to be a problem with three other targets against which both molecules are active according to Pubchem: Cathepsin L (Assay ID 825, inhibitor assay), Cathepsin B (Assay ID 830, inhibitor assay), and Thyroid Stimulating Hormone Receptor (Assay ID 938, agonist assay). A confirmatory assay would be needed for the decoy to make sure it is not a false negative.

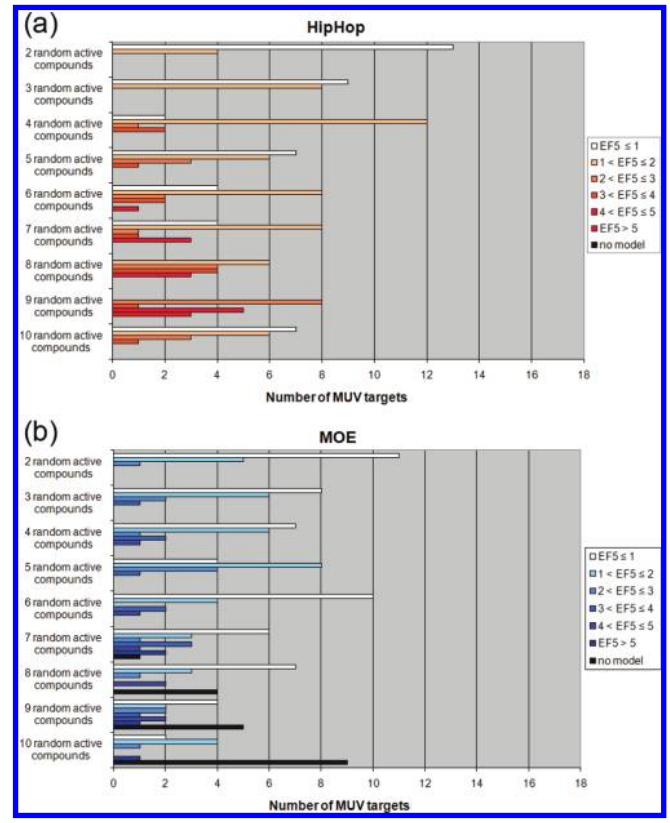


**Figure 9.** Mean enrichment of active molecules within the MUV data sets determined for 100 HipHop (a) and MOE models (b) based on randomly selected training sets. The best overall performance for HipHop and MOE models in the MUV data sets is obtained by using training sets comprising nine and seven randomly selected active compounds, respectively.

**B) Enrichment Assessment of Pharmacophore Elucidators.** *Effect of Training Set Selection on Pharmacophore Model Performance.* In order to assess the virtual screening performance of models derived from different types of training sets, we determined the number of MUV target classes for which HipHop or MOE pharmacophore-based virtual screening resulted in an EF5 value $\leq 1 > 1$ and $\leq 2$, $> 2$ and $\leq 3$, $> 3$ and $\leq 4$, $> 4$ and $\leq 5$, and $>5$, respectively. Moreover, we analyzed the number of MUV target classes for which no model could be elucidated based on the investigated training set.

Figure 9 and Figure S1 (see the Supporting Information) show the mean enrichment results for the 100 HipHop and MOE models based on nine types of random training sets comprising two to ten active molecules. Nine active molecules represented the optimal size for random HipHop training sets (Figure 9a). Random training sets of this size caused the generation of HipHop models which obtained for all 17 classes a mean EF5 value >2, respectively. For MOE modeling, seven randomly selected active molecules represent the most promising training set (Figure 9b). Using this training set, for seven MUV target classes, MOE models with a mean EF5 value >2 were generated. Only for one target class no model could be created

The enrichments determined for the most similar training sets are displayed in Figure 10 and Figure S2 (see the Supporting Information). As shown in Figure 10a, the selection of the three or four most similar active molecules resulted in the best performing HipHop models. Using these

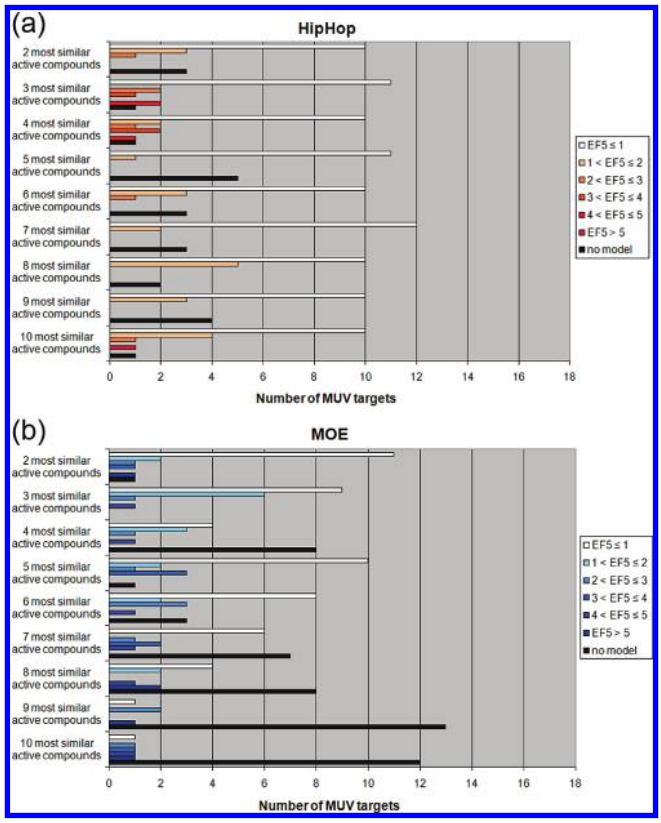


**Figure 10.** Frequency of different EF5 values obtained for HipHop (a) and MOE models (b) derived from the nine most similar pharmacophore-based training sets. For HipHop the best performing models were based on the training sets consisting of the three and four most similar active compounds. In contrast to that, the MOE models with the best virtual screening results were generated using the five most similar active compounds.

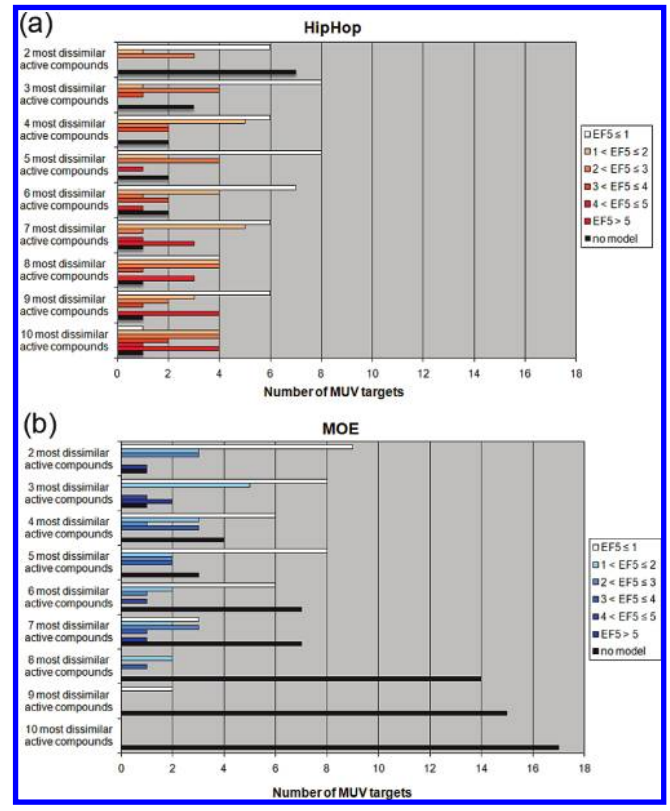


**Figure 11.** Frequency of different EF5 values for HipHop (a) and MOE models (b) generated based on the training sets comprising the most dissimilar active molecules. The ten most dissimilar training set compounds led to the best performing HipHop model, whereas the optimal MOE training set consisted of the three most diverse active compounds.

training sets, only for one MUV target class no model could be generated, and the number of MUV targets with EF5 values >2 was higher than the number determined for any other of the nine HipHop training sets. In contrast to HipHop, for the MOE Pharmacophore Elucidator, the number of successful common feature pharmacophore generations tends to decrease as the training set size grows. Thus, the training sets containing nine and ten similar compounds lead to no MOE model for 13 and 12 MUV targets, respectively, whereas for all 17 target classes a model was generated when the three most similar compounds were used as the training set. The models derived from the five most similar training set compounds obtained for four target classes an EF5 value >2. Since MOE pharmacophore elucidation failed only for one MUV target class, the five-membered set represents the best performing selection of active molecules among the nine most similar training sets.

Figure 11 and Figure S3 (see the Supporting Information) represent the enrichments retrieved for HipHop and MOE models derived from the nine in terms of pharmacophoric patterns most dissimilar training sets. Also these enrichments showed that the number of MOE pharmacophore elucidation failures correlates with the number of compounds used as training set (Figure 11b). However, in contrast to the virtual screening results obtained for the most similar training sets, the best performing training set consisted of only three active molecules. This set not only caused the lowest number of pharmacophore generation failures but also led to the generation of MOE models that caused more often EF5 values >4 than models based on the other eight dissimilar training sets.

For HipHop pharmacophore modeling, the number of virtual screens with EF5 values >2 increased and the number of targets for which no HipHop model was created decreased if training sets with an increasing number of compounds were used for HipHop pharmacophore elucidation (Figure 11a). In agreement with this observation, the training set which led to the generation of HipHop models with the best virtual screening results comprised ten dissimilar active compounds. HipHop models derived from this set produced EF5 values >2 for 11 MUV target classes.

Only for the MUV target focal adhesion kinase (FAK, aid810) no model was retrieved when the ten most dissimilar training set compounds were subjected to HipHop pharmacophore generation. However, FAK was a challenging target for the HipHop and MOE pharmacophore modeling in general. Only the two most dissimilar and the three most similar compounds led to MOE models and the eight most similar compounds to HipHop models with some enrichment of active compounds among the top 5% of the FAK validation database. To determine the reason for this weak performance of MOE and HipHop in creating models for FAK inhibitors, we investigated the pharmacophore similarity between the 30 active compounds for each MUV target. For this purpose, we analyzed again the HipHop and MOE distance matrices, which contain the pairwise alignment scores for the 30 compounds of each MUV active compound set and were used for the selection of the diversity-based training sets (see Materials and Methods). The fit values and

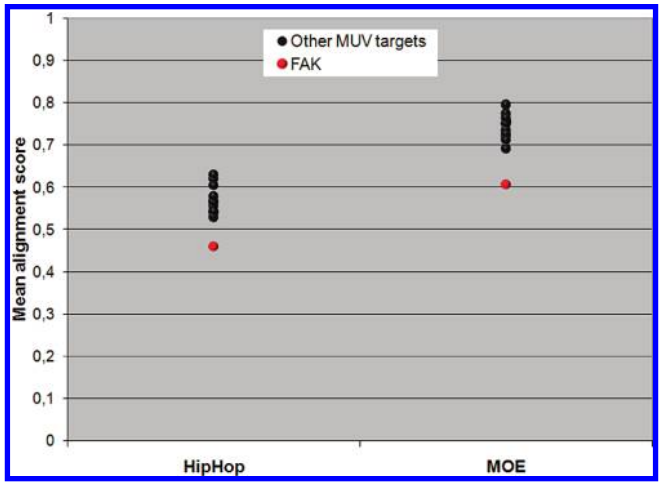


**Figure 12.** Determination of the pharmacophore similarity between the 30 compounds of each of the 17 MUV active compounds set. A high mean alignment score corresponds to a high similarity of the active compounds in terms of chemical feature patterns. The pharmacophore similarity between the 30 FAK inhibitors is significantly lower than the similarity observed for the other MUV active compounds sets. Thus, the lack of success in generating HipHop and MOE models for this target can be explained by the significant higher pharmacophore diversity of the FAK actives.

MOE overlap scores of the distance matrices were rescaled to gain easily to handle alignment scores between 0 and 1. A high alignment score indicated that the two active compounds contain similar pharmacophoric patterns, whereas a low score showed that the chemical feature patterns of the compounds are dissimilar. When we determined the distribution of the mean alignment scores for the 17 MUV targets, the FAK target was identified as a data set outlier (Figure 12).

The 30 FAK inhibitors are far more diverse with respect to their pharmacophore than any other set of active compounds in the MUV. This could be an explanation for the frequent failure of the HipHop and MOE algorithm to generate common feature-based pharmacophore models for this target.

*Random Selection versus Diversity-Based Selection.* The results of our study show that training set selection has a great impact on pharmacophore-based enrichment of active compounds. On that account, a simple rule for the optimal training set selection would be of utmost importance in the field of pharmacophore modeling. Comparing the enrichments between the nine most similar and the nine most dissimilar training sets indicates that an optimal training set selection strategy should be based on the active molecules with the most diverse pharmacophoric patterns. In case of MOE modeling, the set should consist of a few active molecules, whereas HipHop pharmacophore elucidation performs best if a large training set is used as input. However, the question arises if this diversity-based training set selection outperforms randomly selected training sets.

To answer this question we directly compared the enrichments obtained for random-based and diversity-based training sets of optimal size for HipHop and MOE modeling (Figure 13). For both pharmacophore modeling methods, the randomly selected training sets of optimal size led to the generation of models that performed significantly better than the models derived from the best diversity-based training sets.

*HipHop versus MOE.* Finally we wanted to find out if the HipHop algorithm or the MOE Pharmacophore Elucidator leads to a better pharmacophore-based enrichment of active

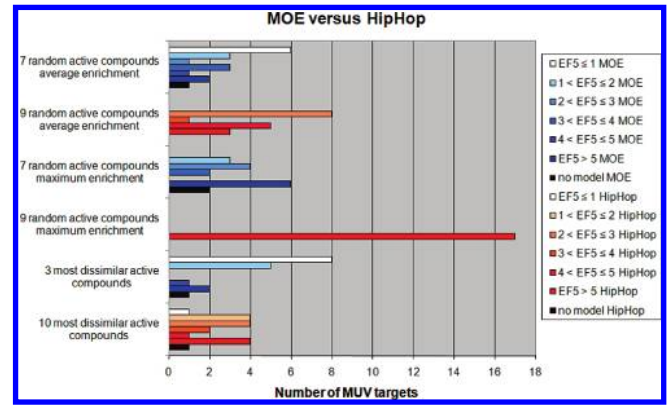


**Figure 13.** Enrichment comparison between randomly selected and diversity-based training sets for HipHop and MOE pharmacophore modeling of optimal size. For both pharmacophore modeling methods, the average enrichment obtained for 100 models based on randomly selected training sets is better than the enrichment determined for the models derived from the two optimal diversity-based training sets. Comparing the enrichments between HipHop and MOE models derived from these automatically selected training sets, HipHop models perform significantly better than MOE models. The maximum single enrichment determined for the 100 random-based HipHop and MOE models is far higher than the average enrichment of these models or the enrichment calculated for the diversity-based models. This indicates that a simple trial and error selection strategy using sets of active compounds and decoys for virtual screening validation will most probably end up in training sets performing better than automatically selected sets.

molecules in the MUV data sets. As displayed in Figure 13, the random and diversity-based training sets of optimal size led to the generation of MOE models that cause for a lower number of MUV target classes EF5 values >2 than the corresponding HipHop models.

The reason for this difference in performance could be that the MOE Pharmacophore Elucidator generates models with two, three, four, and if possible with five features but allows only the export of models including four or five features. In contrast to that, HipHop also exports models that are based on three or more than five features. Although, it is reasonable to restrict the number of features for a model to a specific number to avoid unselective three-feature models as well as too restrictive models because of a chemical feature count larger than five, this strategy can be an obstacle for model generation in the case of large dissimilar training sets. For example, no MOE model was derived from the training set containing the ten most dissimilar active compounds, and only weak performing models with no enrichment at the top 5% were generated when the nine most dissimilar active compounds were used as training set.

Because a large and diverse set of active molecules contains more chemical information about the common pharmacophore of all ligands for one target than a small training set, a three-feature model based on ten active molecules can outperform a four-feature model containing only the information of three active molecules. This was the case for MUV target class aid832, where the simple three-feature HipHop model derived from the ten most dissimilar active molecules (EF5 = 9) dramatically outperformed the four-feature MOE model based on the three most dissimilar compounds (EF5 = 0.7). The HipHop model was also superior to all other MOE models based on the nine similar and the nine dissimilar training sets (EF5 ≤ 4). On that account, less restrictive limits for the feature count of models

generated by the MOE Pharmacophore Elucidator could result in a better overall performance in the MUV.

*Pitfalls of Pharmacophore-Based Virtual Screening.* Even the HipHop and MOE models derived from random- and diversity-based training sets of optimal size produced only for a few MUV target classes enrichments applicable to lead discovery. The reasons for this weak performance are, on the one hand, the pitfalls, such as activity cliffs, discussed above for the five similarity search tools and, on the other hand, the automated selection of random- and diversity-based training sets.

As shown in Figure 13 and Figure S4 (see the Supporting Information), if not the mean enrichment of 100 models based on the optimal-sized set of randomly selected active compounds but only the enrichment of the best of the 100 models is determined, EF5 values >5 are obtained for all and six MUV target classes for the best HipHop and MOE models, respectively. Comparing these results with the mean enrichments of the 100 random models and the enrichments of the models based on the optimal diversity-based training sets, where the best HipHop and MOE models caused for four and two target classes EF5 values >5, respectively, it is obvious that the training set with the maximum performance cannot be generated by diversity-based selection and on average also not by random selection. Thus, a simple trial and error training set selection strategy, where the resulting models are validated by virtually screening sets of known active compounds and decoys, will rather result in good performing models than any random- or diversity-based training set selection. Therefore, pharmacophore modeling using the HipHop and MOE algorithm for elucidation of common chemical features of a set of known active compounds cannot be done in an automated way. This indicates that an expert who validates the performance of several rationally selected training sets has a far higher probability to generate models applicable to lead discovery than an automated pharmacophore modeling workflow.

## CONCLUSIONS

In the current work, performance of five similarity search methods and two pharmacophore tools was assessed against a carefully selected benchmark series. In addition to the individual similarity search methods, fusion of their results (min_rank and avg_rank) was studied. Both single and multiple template scenarios were studied with the similarity search methods. There are a number of lessons to be learned from the current study.

The results further strengthen the case for both similarity and group fusion (use of multiple methods and templates, respectively). Both min_rank and avg_rank similarity fusion techniques produce more consistent results than any of the individual methods. With multiple templates, picking the set with chemical diversity in mind (PAM clustering in this case) almost always leads to a performance superior to the average randomly selected template set of the same size.

The disappointing performance of all the studied similarity search methods in most of the 17 target classes raises the need to explain why this is the case. Three putative reasons - false negatives, activity cliffs, and differing binding modes - were discussed in detail above. To improve the situation, certain guidelines can be given that would be helpful in building future versions of benchmark data sets and computer-aided drug design in general.

First, validation of inactive molecules would be valuable. After the primary screen, the research could identify a set of inactives most similar to the active compounds found by a simple fingerprint method for example. This subset of the inactives would then be retested to reduce the false negative rate.

Second, future benchmark sets should, if possible, contain only active compounds binding the target in similar manner. The similarity search methods cannot be expected to identify active molecules binding different parts of the target. This would require experimental work in the form of competitive assays against a ligand with a known binding mode or cocrystal structures.

Activity cliffs are a serious Achilles' heel for similarity search methods as it is impossible to tell without knowledge of protein structure which small structural differences cause significant changes in activity. This means that the simple approach of using only active molecules as templates is not enough, and information of structurally similar decoys is needed.[37]

We also investigated the performance of two well established pharmacophore modeling tools, Catalyst's HipHop algorithm and MOE's Pharmacophore Elucidator, using different kinds of automatically created training sets. For both pharmacophore elucidators, models derived from training sets consisting of randomly selected active molecules outperform models generated using training set compounds clustered by their pharmacophoric pattern diversity. Comparing the models based on the best training sets, HipHop performs significantly better than the MOE Pharmacophore Elucidator. However, single enrichments of randomly selected training sets show that the maximum performing model is rather generated by a training set selected manually by trial and error than by automatically selected training sets. Thus, HipHop and MOE pharmacophore modeling cannot be easily automated, and 3D pharmacophore-based virtual screening enrichment highly depends on the expert generating the model.

## ACKNOWLEDGMENT

The authors thank Openeye Scientific Software, Inc. for providing the academic license for the Openeye package, Accelrys Software, Inc. for the free academic license of Pipeline Pilot Student Edition, and CSC - IT Center for Science Ltd. for the computing resources used in this work.

**Supporting Information Available:** Figures and tables for additional performance metrics for MOE, HipHop, and the five ligand-based screening tools and the table of confirmatory Pubchem assays analyzed. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Kirchmair, J.; Markt, P.; Distinto, S.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Langer, T.; Wolber, G. The Protein Data Bank (PDB), its related services and software tools as key components for in silico guided drug discovery. *J. Med. Chem.* **2008**, *22*, 7021–7040.

(2) Hristozov, D. P.; Oprea, T. I.; Gasteiger, J. Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. *J. Comput.-Aided Mol. Des.* **2007**, *10−11*, 617–640.

(3) Markt, P.; Feldmann, C.; Rollinger, J. M.; Raduner, S.; Schuster, D.; Kirchmair, J.; Distinto, S.; Spitzer, G. M.; Wolber, G.; Laggner, C.; Altmann, K. H.; Langer, T.; Gertsch, J. Discovery of novel CB2 receptor ligands by a pharmacophore-based virtual screening workflow. *J. Med. Chem.* **2009**, *2*, 369–378.
(4) Markt, P.; Petersen, R. K.; Flindt, E. N.; Kristiansen, K.; Kirchmair, J.; Spitzer, G.; Distinto, S.; Schuster, D.; Wolber, G.; Laggner, C.; Langer, T. Discovery of novel PPAR ligands by a virtual screening approach based on pharmacophore modeling, 3D shape, and electrostatic similarity screening. *J. Med. Chem.* **2008**, *20*, 6303–6317.
(5) Schwarz, O.; Jakupovic, S.; Ambrosi, H. D.; Haustedt, L. O.; Mang, C.; Muller-Kuhrt, L. Natural products in parallel chemistry--novel 5-lipoxygenase inhibitors from BIOS-based libraries starting from alpha-santonin. *J. Comb. Chem.* **2007**, *6*, 1104–1113.
(6) Mochalkin, I.; Miller, J. R.; Narasimhan, L.; Thanabal, V.; Erdman, P.; Cox, P. B.; Prasad, J. V.; Lightle, S.; Huband, M. D.; Stover, C. K. Discovery of antibacterial biotin carboxylase inhibitors by virtual screening and fragment-based approaches. *ACS Chem. Biol.* **2009**, *6*, 473–483.
(7) MDL Drug Data Report. Symyx Technologies. 2009.
(8) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *23*, 6789–6801.
(9) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *4*, 1504–1519.
(10) Evers, A.; Hessler, G.; Matter, H.; Klabunde, T. Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols. *J. Med. Chem.* **2005**, *17*, 5448–5465.
(11) Irwin, J. J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *3*−*4*, 193–199.
(12) Mackey, M. D.; Melville, J. L. Better than random? The chemotype enrichment problem. *J. Chem. Inf. Model.* **2009**, *5*, 1154–1162.
(13) Rohrer, S. G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* **2009**, *2*, 169–184.
(14) The Pubchem Project. http://pubchem.ncbi.nlm.nih.gov/ (accessed month day, year).
(15) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **2008**, *3*−*4*, 169–178.
(16) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *23*−*24*, 1046–1053.
(17) Tiikkainen, P.; Poso, A.; Kallioniemi, O. Comparison of structure fingerprint and molecular interaction field based methods in explaining biological similarity of small molecules in cell-based screens. *J. Comput.-Aided Mol. Des.* **2009**, *4*, 227–239.
(18) Tervo, A. J.; Ronkko, T.; Nyronen, T. H.; Poso, A. BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. 1. Alignment and virtual screening applications. *J. Med. Chem.* **2005**, *12*, 4076–4086.
(19) Ronkko, T.; Tervo, A. J.; Parkkinen, J.; Poso, A. BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. II. Description and characterization. *J. Comput.-Aided Mol. Des.* **2006**, *4*, 227–236.
(20) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *5*, 1489–1495.
(21) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How to optimize shape-based virtual screening: choosing the right query and including chemical information. *J. Chem. Inf. Model.* **2009**, *3*, 678–692.
(22) Openeye Scientific Software, Inc. *EON*; 2007.
(23) Openeye Scientific Software, Inc. *OMEGA2*; 2007.
(24) Accelrys, Inc. *Pipeline Pilot Student version 6.1.5.0*; 2007.
(25) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of data fusion methods in virtual screening: similarity and group fusion. *J. Chem. Inf. Model.* **2006**, *6*, 2206–2219.
(26) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley: New York, 1990.
(27) Guha, R.; Van Drie, J. H. Assessing how well a modeling protocol captures a structure-activity landscape. *J. Chem. Inf. Model.* **2008**, *8*, 1716–1728.
(28) Guha, R.; Van Drie, J. H. Structure--activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *3*, 646–658.
(29) Clement, O. O.; Mehl, A. T. HipHop: Pharmacophores based on multiple common-feature alignments. In *Pharmacophore perception, development, and use in drug design*; International University Line: La Jolla, CA, U.S.A., 2000; pp 71−84.
(30) Accelrys, Inc. *Catalyst*; 2005.
(31) Chemical Computing Group. *MOE*; 2007.
(32) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *15*, 2887–2893.
(33) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump. *QSAR Comb. Sci.* **2006**, *12*, 1162–1171.
(34) 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**, *22*, 3256–3266.
(35) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* **2006**, *2*, 462–470.
(36) Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Kang, Z. B.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M. Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data. *J. Chem. Inf. Model.* **2007**, *4*, 1319–1327.
(37) Kalliokoski, T.; Ronkko, T.; Poso, A. FieldChopper, a new tool for automatic model generation and virtual screening based on molecular fields. *J. Chem. Inf. Model.* **2008**, *6*, 1131–1137.

CI900249B