JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Improving Classical Substructure-Based Virtual Screening to Handle Extrapolation Challenges

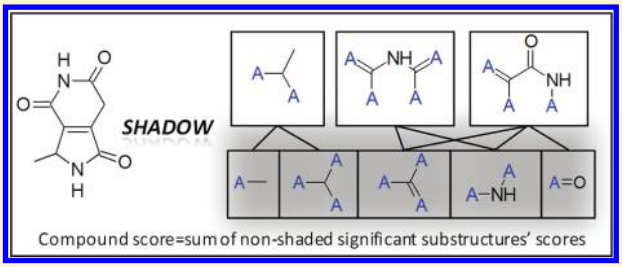
Tammy Biniashvili,[†,‡] Ehud Schreiber,[†] and Yossef Kliger[†,*]

[†]Compugen LTD, Tel Aviv 69512, Israel

[‡]The Mina and Everard Goodman Faculty of Life Sciences, Bar Ilan University, Ramat-Gan 52900, Israel

Supporting Information

ABSTRACT: Target-oriented substructure-based virtual screening (sSBVS) of molecules is a promising approach in drug discovery. Yet, there are doubts whether sSBVS is suitable also for extrapolation, that is, for detecting molecules that are very different from those used for training. Herein, we evaluate the predictive power of classic virtual screening methods, namely, similarity searching using Tanimoto coefficient (MTC) and Naive Bayes (NB). As could be expected, these classic methods perform better in interpolation than in extrapolation tasks. Consequently, to enhance the predictive ability for extrapolation tasks, we introduce the Shadow approach, in which inclusion relations between substructures are considered, as opposed to the classic sSBVS methods that assume independence between substructures. Specifically, we discard contributions from substructures included in ("shaded" by) others which are, in turn, included in the molecule of interest. Indeed, the Shadow classifier significantly outperforms both MTC ($pValue = 3.1 \times 10^{-16}$) and NB ($pValue = 3.5 \times 10^{-9}$) in detecting hits sharing low similarity with the training active molecules.



INTRODUCTION

Drug discovery has been greatly influenced by advances in high throughput screening (HTS),[1] which have led to an unprecedented wealth of experimental data. Yet, today's drug discovery industry is facing financial pressure that demands higher rate of new molecular entities per year. The straightforward solution, that is, to increase further the number of molecules synthesized and screened, is time-, labor-, and money-intensive.[2] Thus, it makes sense to use the vast amount of information already gathered from experiments, mainly HTS efforts, to develop computational methods that may enhance hit rate.[3−6]

Virtual screening (VS) is used for hit discovery and takes many forms.[7−9] A VS procedure takes as input a large amount of potential molecules and provides predictions as to which compounds have a higher chance of being active.[10] This process has been compared to finding a needle in a haystack, where the needle is a promising hit and the haystack represents the complex multidimensional chemical space.[11] One VS approach, first suggested decades ago,[12,13] is motivated by the abundance of certain substructures associated with a desired property (e.g., biological activity) in drugs and drug candidates. In recent years, this substructure based VS (sSBVS) approach was successfully used to predict molecules that target a specific molecule,[11,14−23] and to predict various molecular properties, mostly relevant for the chances that the molecule may become a therapeutic agent.[24−30]

Whereas sSBVS achieves high performance when the screened molecules share high similarity to known active compounds (i.e., when interpolating), much lower performance is expected when trying to predict properties for molecules that are less similar to the active compound used for training (i.e., when extrapolating).[11,31,32] In this study, we quantitatively evaluated this phenomenon for two of the classic sSBVS methods. The first is the Tanimoto coefficient-based similarity search (MTC), which scores the screened molecule as its Tanimoto coefficient value with the most similar training active molecules (Figure 1, left column). The second is the Naive Bayes classifier (NB), which scores the screened molecule as the sum of the activity contribution scores of all the significant substructures within it (Figure 1, middle column). Indeed, our analysis reveals that both MTC and NB methods achieve higher performance for interpolation than for extrapolation. Whereas the ability to interpolate molecular features has practical importance, extrapolation, which is more challenging scientifically, has also practical potential with some important advantages such as molecular diversity and avoidance of intellectual property issues.[33] Next, to improve performance in the extrapolation task, we decided to tackle one of the basic assumptions of the classic methods: the independence between substructures. It is clear that this assumption is not true, but most of the algorithms use this assumption for simplicity. To the best of our knowledge, the consequence of relaxing this assumption was not directly tested before. Our modified classifier (the Shadow classifier, Figure 1, right column) is a

Received: October 4, 2011
Published: February 23, 2012

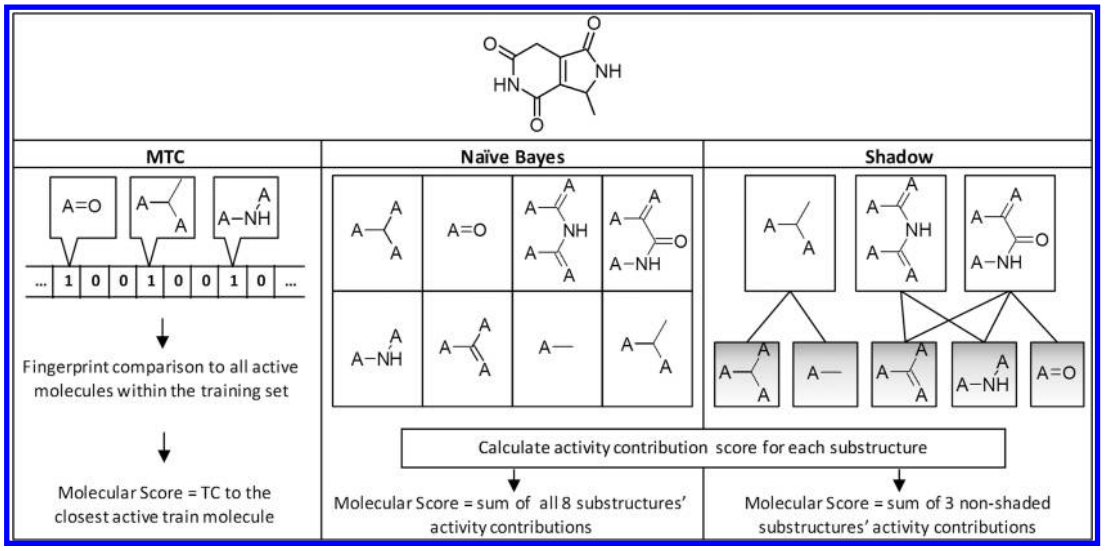


Figure 1. Schematic description of sSBVS methods. A screened compound is presented as an example with a few of its corresponding substructures. MTC (left column) represents the screened molecule as a fingerprint and scores the molecule as the Tanimoto coefficient (TC) value with the most similar training active molecules. The Naive Bayes classifier (middle column) scores the screened molecule as the sum of the activity contribution scores of all the significant substructures within it. The Shadow classifier (right column) scores the screened molecules as the sum of activity contribution scores of substructures that are not included (not-shaded) by any other substructure. Inclusion is denoted by a line, where the upper substructure includes the lower one. More generally, presence of the higher substructure in a compound implies the presence of the lower one. All screened molecules are ranked according to their calculated score where a higher score indicates a better probability for bioactivity against the target of interest. "A" designates any heavy atom (i.e., non-hydrogen; carbon, nitrogen, or oxygen in our example). An unmarked vertex designates a carbon atom together with the hydrogen atoms needed to complete its valence number to 4.

variant of the NB one, which considers perhaps the simplest relation between significant substructures, that is, full inclusion. Specifically, it discards contributions from substructures included in ("shaded" by) others which are, in turn, included in the molecule of interest. Our results show that the Shadow classifier enhances the performance in cases of extrapolation and significantly outperforms the classic sSBVS methods.

RESULTS

Most Known Active Molecules Are Similar to Other Active Molecules. A basic difficulty for sSBVS methods stems from redundancies and similarity issues common in small molecule databases. Thus, a major critique of sSBVS is that while it offers good performance for interpolation tasks, its performance dramatically deteriorates when trying to predict the activity of molecules less similar to known active molecules (i.e., for extrapolation tasks).[11,31,32] To explore this issue further, we analyzed the similarity between molecules from an anti-HIV screening database.[34] Throughout this article, similarity between molecules is calculated using the Tanimoto coefficient (TC). The properties used for similarity measures were substructures calculated using Extended Connectivity FingerPrint (ECFP), which are topological fingerprints that represent the presence of particular substructures within a compound.[35] It is noteworthy that there are alternatives to ECFP, many of them originating from graph theory, like gSpan[36] and the FSG algorithm[37] that was implemented by Karypis and colleagues.[38] Herein, we prefer to use ECFP as it allows us to easily handle relations between substructures, which is a crucial ability in this study, as discussed below (regarding the shadow approach).

We hypothesized that the task of classifying a molecule is naturally divided into two cases, depending on whether or not this molecule is similar to a known active molecule. Classifying "seen before" molecules (interpolation) and "new" molecules (extrapolation) are tasks that are different in nature, and have a different level of difficulty. This hypothesis suggests that it will be very difficult to develop a method that works well for both interpolation and extrapolation tasks, and we should better develop different methods for these different tasks. Furthermore, it is clear that separate performance evaluations are needed for the two tasks. To differentiate between the classification tasks, we calculated the MTC of each screened molecule to the active molecules within the data set. The results, binned by different MTC values, are shown in Table 1.

Table 1. Most Known Active Molecules Are Similar to Other Active Molecules in the Anti-HIV Data Set

MTC range	#molecules	#active	#active/#molecules
0–0.1	682	0	0
0.1–0.2	19506	0	0
0.2–0.3	17937	4	0.0002
0.3–0.4	2260	9	0.0040
0.4–0.5	700	12	0.0171
0.5–0.6	469	18	0.0384
0.6–0.7	423	64	0.1513
0.7–0.8	363	123	0.3388
0.8–0.9	216	101	0.4676
0.9–1	130	91	0.7

The results confirmed our hypothesis: molecules similar to other known active molecules (defined as having MTC > 0.6) have a probability of 33.5% to be active themselves. In contrast, molecules that share only low similarity with known active molecules (having MTC ≤ 0.6) have a probability of only 0.1% to be active.

Classic sSBVS Methods Perform Better in Interpolation Tasks than in Extrapolation Ones. To confirm the rationale of dividing the test set into "seen before" and "new"

dx.doi.org/10.1021/ci200472s | J. Chem. Inf. Model. 2012, 52, 678–685

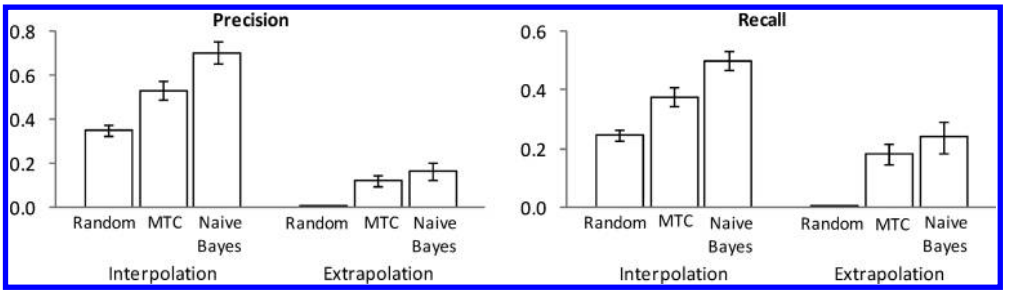


Figure 2. Classic sSBVS methods (MTC and Naive Bayes) perform better in interpolation than in extrapolation tasks. MTC scores the screened molecule as its TC value relative to the most similar training active molecule, whereas NB scores the screened molecule as the sum of the activity contribution scores of all the significant substructures within it. These classifiers' performance was assessed by dividing the data set to training and test sets and further partition the test set to extrapolation and interpolation groups. Then, the classifiers ranked the compounds on each group independently. This entire procedure was repeated 100 times and the averaged precision and recall for the top ranked 100 molecules of the test sets in the anti-HIV data set are presented for the interpolation and extrapolation tasks, separately.

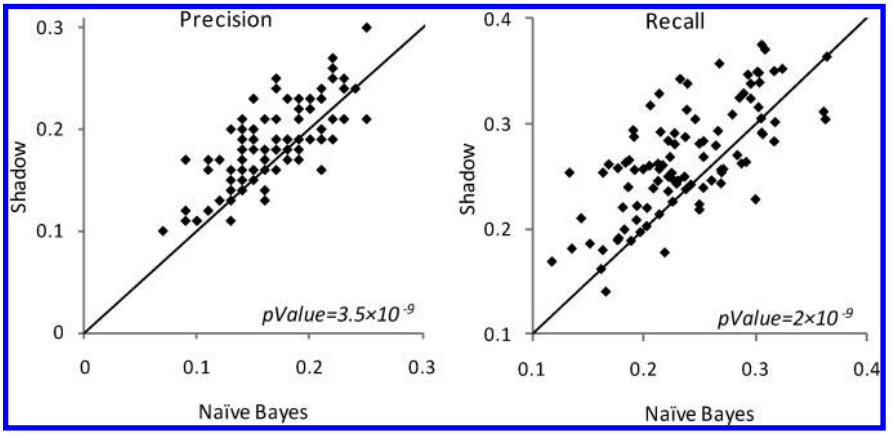


Figure 3. Shadow outperforms Naive Bayes for extrapolation tasks. Each symbol represents the precision or recall achieved for one random split to training and test sets. The precision and recall are calculated for the top-100 ranked test set molecules. Note that most data splits are above the diagonal suggesting that the precision and recall achieved by the Shadow method are higher than those achieved by the classical Naive Bayes method for extrapolation tasks.

molecules, we evaluated the performance of two classic sSBVS methods, MTC and NB. Specifically, we randomly divided the anti-HIV screening data set into training and test sets; the division was performed in a stratified manner, retaining the proportion of active and inactive molecules. The test set was partitioned to molecules having MTC larger than 0.6 (seen before/interpolation) and molecules having MTC $\leq$ 0.6 (new/extrapolation) to the active molecules within the training set. This entire procedure was repeated a hundred times to enable statistical analysis. An average of 141.62 $\pm$ 6.59 active molecules had MTC > 0.6 and these were used for evaluating the performance of the interpolation task, whereas the rest (69.38 $\pm$ 6.59) of the active molecules were used for evaluating the performance of the extrapolation task. For comparison, the theoretical precision and recall achieved by a random classifier were calculated. Figure 2 summarizes the performance for the MTC, NB and random predictors. As expected, in the case of interpolation both classic methods, and even the random predictor, achieve high performance. However, for the extrapolation task, despite an impressive enrichment factor over random (for example, calculated enrichment factor for precision is 37.5 and 49.7 for MTC and NB, respectively), both methods demonstrate much lower performance. The results of the separate analyses encourage the development of a superior method for extrapolation, that is, focusing on virtually screening "new" molecules.

Activity Contribution Score of Shaded Substructures Should Not Be Summed-Up for New Molecules. Both MTC and NB assume independence between substructures. Whereas this assumption is not true, it is often used to achieve low algorithmic complexity, while keeping high performance.[23,39] In this study, we tried to relax this assumption partially for the extrapolation tasks. Toward this end, we decided to consider perhaps the most straightforward dependence, that is, full inclusion of a substructure within another. In such a case, the correlation between the inclusions is the strongest possible, becoming logical implication (if $i \subseteq molecule$ and $j \subset i$, then necessarily $j \subset molecule$). Such an improved classifier should thus take into account only "non-shaded" substructures, that is, significant substructures that are not fully included in any other significant substructure within the same molecule (Figure 1). Considering both non-shaded and shaded substructures is a redundant, and the non-shaded substructures are preferred as they are more specific, and hence more informative. Whereas the NB classifier scores the test set molecules according to eq 1

$$S_{\text{molecule}} = \sum_{i \subseteq \text{molecule}} f_i \qquad (1)$$

dx.doi.org/10.1021/ci200472s | J. Chem. Inf. Model. 2012, 52, 678–685

the shadow approach, scores the test set molecules according to eq 2, which is a modified version of the NB scoring method

$$S_{\text{molecule}} = \sum_{\substack{i \subseteq \text{molecule} \\ \nexists j \; i \subset j \subseteq \text{molecule}}} f_i \qquad (2)$$

where f_i is the activity contribution score of the significant substructure i. The mathematical formulation for a nonshaded significant substructure i is that within the molecule of interest there is no significant substructure j containing substructure i. The performance of this new method was evaluated using the same procedure described in the previous section. The results revealed that the Shadow classifier significantly outperforms both classic methods for the extrapolation tasks using the anti-HIV data set (Table S2, Figure 3). While the difference in performance is not large in absolute terms, it is statistically significant and it suggests that focusing on more refined models of dependence between substructures may be advantageous.

In addition to the precision and recall calculated for the top ranked molecules, an analysis using the entire screened data may also be of interest. Yet, as only the very top of the ranked list of predictions is of interest, standard Receiver Operating Characteristic (ROC) curves are not very useful. Recently, Swamidass, Baldi and colleagues developed a concentrated ROC (CROC) framework that offers an effective way for measuring and visualizing early retrieval performance.[40] Figure 4 shows typical CROC curves for the Naive Bayes and Shadow classifiers, revealing that Shadow outperforms Naive Bayes for all cutoffs.

The entire procedure described above was performed on additional three independent data sets: (i) a data set of molecules inhibiting the dimerization of the HCV core protein (Supporting Information Tables S2–S3, Figure S1); (ii) a data set of molecules tested for their ability to activate the nuclear receptor SF-1 (Supporting Information Tables S4–S5, Figure S2); and (iii) a collection of CYP2C19 inhibitors (Supporting Information Tables S6–S7, Figure S3). For all data sets analyzed, Shadow outperforms the Naive Bayes classifier. This material is available free of charge via the Internet at http://pubs.acs.org. Of course, this does not guarantee that the Shadow classifier always outperforms the classical sSBVS methods. In fact, even the simple measure of Tanimoto coefficient was shown to be sensitive to the data set analyzed.[41] Yet, our findings suggest the generality of the Shadow approach, as it works for four nonrelated molecular data sets, which consist of both inhibitors and activators, and deal with both bioactivity and metabolism.

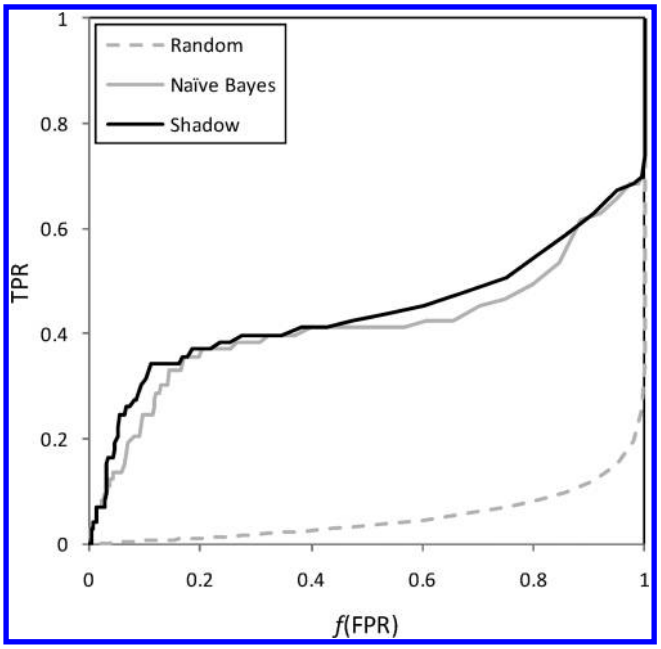


Figure 4. Shadow outperforms Naive Bayes for extrapolation tasks. Typical Concentrated ROC (CROC) graph comparing the Naive Bayes and Shadow classifiers. FPR = false positive rate; TPR = true positive rate. The x-coordinate is transformed by the exponential equation $f(x) = (1 - e^{-20x})/(1 - e^{-20})$ to highlight the early retrieval performance.

DISCUSSION

sSBVS methods offer high performance and computational simplicity and therefore are widely used.[42] Many applications successfully use sSBVS for predicting various molecular properties.[11,14,16,17,19−30,39]

It is noteworthy that in biological and chemical data sets, commonly some of the data are similar to known cases, while others are very different from all the known cases. Of course, prediction of the latter case is more challenging, but also more rewarding. Several groups highlighted the need to evaluate performance, separately, for interpolation and extrapolation tasks.[11,31,32,43] A similar approach was used in other computational tasks, e.g. for proteolytic site prediction.[44] In this study, we quantitatively explored whether there is a need for separately evaluating performance of interpolation and extrapolation tasks. Indeed, this need was confirmed for four different data sets (Table 1 and Supporting Information Tables S1–S7). Furthermore, we showed that whereas the performance of classical sSBVS methods is quite high for interpolation tasks, prediction performance for extrapolation challenges is much lower. Next, we suggested reconsidering the basic assumptions of classical sSBVS methods in an attempt to improve their competence in extrapolation tasks. In this manuscript, we presented the Shadow method, which can be used as a prototype for improving sSBVS methods.

It is noteworthy that the Shadow approach depends on the presence of inclusion relations between significant substructures and the ability to detect them. In other words, when such relations are not present or are sparse the performance of the Shadow approach is expected to resemble the performance of Naive Bayes.

The Shadow approach presented in this paper may be generalized for solving similar problems. In general, it is relevant for cases where considering relations between terms can resolve bias issues when calculating correlation between two vectors. For example, the approach may be profitably applicable to the Tanimoto coefficient (TC) itself, which is widely used for assessing chemical similarity between molecules (eq 2).[45−47] This measure might be overestimated by redundant "on" bits that exaggerate the intersection between the molecules compared. Of course, additional important applications lay outside the field of sSBVS. For example, in correlated mutation analysis,[48−50] a major source of noise is the evolutionary relatedness between sequences in a multiple sequence alignment of the protein of interest and its homologous sequences. Several methods are able to reduce this phylogenetic bias,[51−59] but this basic problem is yet to be solved.

In general, the pharmaceutical industry tightly protects information regarding its chemical substances due to commercial issues. Still, informatics resources and algorithms have the potential to enhance precompetitive collaborations between companies as recently suggested by several industry leaders.[60] We hope that the algorithm presented in this study will contribute to this important initiative.

In summary, bona fide extrapolation is a challenging task. It is not only theoretically interesting, but also has practical importance in medicinal chemistry. Extrapolation tasks may retrieve hits having interesting characteristics, such as structural novelty, patentability, or both.[33] In this study, we only tried to tackle one simple issue: the relations between substructures that are typically ignored for various reasons in sSBVS. In this prototype, we only treat the simplest relation, that is, full inclusion between substructures. Yet, the results are significantly better. Thus, it is expected that considering more complex relations between substructures will further enhance sSBVS methods in the realm of extrapolation.

METHODS

Data Preparation. We use four independent data sets: (i) The NCI HIV antiviral screen.[34] It comprises of 422 active, 1081 moderately active and 41,179 inactive molecules. To sharpen this analysis, all moderately active compounds were removed from the database. (ii) The Scripps Research Institute screening identifying inhibitors of Hepatitis C Virus (HCV) core protein dimerization (PubChem BioAssay ID 1899). It comprises of 998 active and 301 669 inactive molecules. As this data set includes many more inactive molecules than the first data set, we randomly selected 40 000 inactive molecules. (iii) Screening results aiming at identifying activators of the Steroidogenic Factor 1 (SF-1) nuclear receptor (PubChem BioAssay ID 522). It comprises of 1225 active and 63 682 inactive molecules. (iv) The Sanford-Burnham Center for Chemical Genomics screen of inhibitors of CYP2C19 (PubChem BioAssay ID 778). It comprises of 20 295 active and 75 564 inactive molecules. As this data set includes many more inactive molecules than the first data set, we randomly selected 40 000 inactive molecules. All molecules were represented in the canonical SMILES format[61] using the Open Babel software package.[62] Each database was randomly divided 100 times into equal sized training and test sets. The partition of the data was done in a stratified manner, that is, retaining the proportion of active and inactive molecules.

Molecular Fingerprints. Molecular fingerprints are arrays generated for each molecule containing, as elements, binary features representing the presence or absence of particular substructures within that molecule.[63] Various methods for fragmenting molecules were developed.[64−67] In this study, we used the extended connectivity fingerprints (ECFP)[35] since they have been proven consistently superior to others.[43] Fragments were generated for all molecules using the Pipeline Pilot software (SciTegic, Accelrys, Inc.) with a diameter of up to 4 bonds (ECFP_4).

Identification of Significant Substructures. All molecules within the training set were fragmented into substructures using the ECFP_4 method. Next, substructures for which the amount of data was not sufficient to reliably calculate the activity contribution score were discarded. Toward this end, for each substructure i, we define a 2 × 2 contingency table $\{n^i_{b,c}\}$ with $b = \{0,1\}$ and $c = \{0,1\}$, where $b = 1$ signifies activity, $c = 1$ signifies that the molecule contains the specified substructure, and $n^i_{b,c}$ counts the number of molecules of the appropriate type. $n^i_{1,1}$ is the number of active molecules containing the substructure$i$, $n^i_{1,0}$ is the number of active molecules that do not contain the substructure i, $n^i_{0,1}$ is the number of inactive molecules containing the substructure i and $n^i_{0,0}$ is the number of inactive molecules not containing the substructure i. Then, a two-tailed Fisher's exact test of the contingency table was performed, and substructures having $pValue \leq 0.01$ were considered as statistically significant. For each of these significant substructures, an "activity contribution" score was calculated based on the substructure's frequencies within active and inactive molecules in the training set. The activity contribution score was calculated as suggested by[68] (eq 3)

$$f_i = \log\left(\frac{n^i_{1,1}}{n^i_{0,1}} \times \frac{n^i_{0,0} + n^i_{0,1}}{n^i_{1,0} + n^i_{1,1}}\right) \quad (3)$$

This calculated activity contribution score corresponds to the log of the likelihood ratio of a contained substructure to be within active or inactive molecules. Thus, indeed it makes sense to sum these scores as done by NB methods;[17,68] this is equivalent to multiplying the likelihoods, as befits the assumed independence.

Classifiers Construction. *Tanimoto Coefficient-Based Similarity Search.* The Tanimoto coefficient (TC) is a widely used measure for assessing the similarity between two molecules represented by fingerprints (eq 4)

$$\mathrm{TC}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

where A and B are two molecules, $|A \cap B|$ denotes the size of their fingerprints intersection (i.e., the number of 1-bits common to A and B) and $|A \cup B|$ denotes the size of their fingerprints union (i.e., the number of 1-bits in either A or B).[46] TC, which ignores common 0-bits within the molecules compared, ranges from 0 to 1, where 0 signifies no similarity, and 1 signifies fingerprint identity, suggesting high similarity between the two molecules compared. TC-based similarity search[47,69] uses TC values that were calculated between all molecules in the test set and all active molecules within the training set to asses similarity. These TC values were calculated using the Pipeline Pilot software (SciTegic, Accelrys, Inc.). Test set molecules were ranked according to their maximal TC (MTC) relative to the training active molecules.

Naive Bayes. The Naive Bayes classifier[70] was implemented using the Perl programming language. Molecules within the training set were fragmented to their corresponding substructures and significant substructures with their corresponding activity contribution scores were determined. Next, molecules in the test set were scored as the sum of activity contribution scores of the significant substructures they contain (eq 1).

Shadow. The Shadow classifier was implemented using the Perl programming language. Similarly to the Naive Bayes classifier it determines significant substructures and their activity contribution scores based on the molecules within the training set. The inclusions between these significant substructures were determined using Pipeline Pilot software (SciTegic, Accelrys, Inc.). The screened molecules are scored as detailed in the Results section (eq 2). Specifically, only the activity contribution scores of significant substructures not included (non-shaded) by any other significant substructure within the same molecule are considered (Figure 1).

Performance Evaluation of Virtual Screening Methods. In the field of VS, a major goal is to avoid the need for screening a large compound library. Thus, VS methods are evaluated by their ability to rank the compounds in the screening library so that active compounds rise to the top of the list.[71] If this can be achieved, the active compounds should be found more rapidly than if they are screened at random, and experimental validation of only the best scoring compounds in the library should be sufficient for identifying active hits. In this spirit, a practical measure for VS performance evaluation is the distribution of active compounds within the top ranked 100 molecules.[72,73] In this approach, the top 100 ranked molecules are predicted to be active and all the other molecules are predicted to be inactive. A confusion matrix can be calculated where True positives (TP) correspond to the number of real active molecules within the top 100 molecules, false positives (FP)correspond to the number of inactives within the top 100, false negatives (FN) are the number of actives that did not occur within the top 100, and true negatives (TN) are the number of inactives that did not occur within the top 100. The precision is defined as

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

while the recall is defined as

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

The averaged precision and the averaged recall of the top 100 molecules, over the random divisions to training and test sets, are computed for comparison between the methods.

Another important measure is the enrichment factor which is calculated as the ratio of performance (either precision or recall) achieved by a certain method and a random classifier. For such a random classifier, one has

$$\text{precision} = \frac{\text{number of active molecules}}{\text{total number of molecules}}$$

and

$$\text{recall} = \frac{\text{number of top molecules}}{\text{total number of molecules}}$$

Thus, the recall enrichment factor (relative to the random classifier) depends on the limit determined for the number of top ranked molecules (which we picked to be a hundred). Another way to focus on the early retrievals was recently developed by Baldi, Swamidass and colleagues. This analysis, called Concentrated ROC (CROC) offers an effective way for measuring and visualizing early retrieval performance. In this framework, any relevant portion of the ROC curve (in our case, the top ranks) is magnified smoothly by an exponential transformation of the abscissa coordinates.[40]

■ ASSOCIATED CONTENT

Supporting Information

Tables showing statistical comparison between SBVS methods and similarity between known active molecules and figures showing interpolation versus extrapolation in classic SBVS methods. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kliger@compugen.co.il.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are appreciative to Dr. Peter Ertl for analytical support and fruitful discussions. The authors are thankful to Ron Unger, Haim Ashkenazy, Adi Dan, Michael Biniashvili, Rotem Ben-Hamo, Ifat Shub, and Itamar Borukhov for useful comments and helpful discussions.

■ REFERENCES

(1) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discovery* **2011**, *10* (3), 188−95.

(2) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of "drug-likeness". *Drug Discovery Today* **2000**, *5* (2), 49−58.

(3) Chen, J.; Swamidass, S. J.; Dou, Y.; Bruand, J.; Baldi, P. ChemDB: A public database of small molecules and related chemoinformatics resources. *Bioinformatics* **2005**, *21* (22), 4133−9.

(4) Irwin, J. J.; Shoichet, B. K. ZINC—A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177−82.

(5) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37* (Web Serverissue), W623−33.

(6) Swamidass, S. J. Mining small-molecule screens to repurpose drugs. *Briefings in Bioinformatics* **2011**, First published online: June 29, **2011**.

(7) Kolb, P.; Ferreira, R. S.; Irwin, J. J.; Shoichet, B. K. Docking and chemoinformatic screens for new ligands and targets. *Curr. Opin. Biotechnol.* **2009**, *20* (4), 429−36.

(8) Olender, R.; Rosenfeld, R. A fast algorithm for searching for molecules containing a pharmacophore in very large virtual combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 731−8.

(9) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152* (1), 38−52.

(10) Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N. Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput.-Aided Mol. Des.* **2007**, *21* (1−3), 53−62.

(11) Salum, L. B.; Andricopulo, A. D. Fragment-based QSAR: Perspectives in drug design. *Mol. Diversity* **2009**, *13* (3), 277−85.

(12) Cramer, R. D.; Redl, G.; Berkoff, C. E. Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem.* **1974**, *17* (5), 533−5.

(13) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S.; et al. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **1988**, *31* (12), 2235−46.

(14) Andrade, C. H.; Salum Lde, B.; Castilho, M. S.; Pasqualoto, K. F.; Ferreira, E. I.; Andricopulo, A. D. Fragment-based and classical quantitative structure−activity relationships for a series of hydrazides as antituberculosis agents. *Mol. Diversity* **2008**, *12* (1), 47−59.

(15) Batista, J.; Bajorath, J. Mining of randomly generated molecular fragment populations uncovers activity-specific fragment hierarchies. *J. Chem. Inf. Model.* **2007**, *47* (4), 1405−13.

(16) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based

feature selection, and a naive Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 170−8.
(17) Ertl, P. Enhancement of hit rate in high throughput screening by using fragment-based substructure analysis, U.K. QSAR and Chemo-Informatics Group Autumn Meeting, Horsham, U.K., 2001.
(18) Godden, J. W.; Stahura, F. L.; Bajorath, J. POT-DMC: A virtual screening method for the identification of potent hits. *J. Med. Chem.* **2004**, *47* (23), 5608−11.
(19) Hu, Y.; Shamaei-Tousi, A.; Liu, Y.; Coates, A. A new approach for the discovery of antibiotics by targeting non-multiplying bacteria: a novel topical antibiotic for staphylococcal infections. *PLoS One* **2010**, *5* (7), e11818.
(20) Klekota, J.; Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics* **2008**, *24* (21), 2518−25.
(21) Kondratovich, E. P.; Zhokhova, N. I.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Fragmental descriptors in (Q)SAR: Prediction of the assignment of organic compounds to pharmacological groups using the support vector machine approach. *Russ. Chem. Bull.* **2010**, *58* (4), 657−662.
(22) Salum, L. B.; Valadares, N. F. Fragment-guided approach to incorporating structural information into a CoMFA study: BACE-1 as an example. *J. Comput.-Aided Mol. Des.* **2010**, DOI: 10.1007/s10822-010-9375-z.
(23) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **2004**, *47* (18), 4463−70.
(24) Artemenko, N. V.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Artificial neural network and fragmental approach in prediction of physicochemical properties of organic compounds. *Russ. Chem. Bull.* **2003**, *52* (1), 20−29.
(25) Clark, M. Generalized fragment-substructure based property prediction method. *J. Chem. Inf. Model.* **2005**, *45* (1), 30−8.
(26) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43* (20), 3714−7.
(27) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform* **2009**, *1* (1), 8.
(28) Ursu, O.; Oprea, T. I. Model-free drug-likeness from fragments. *J. Chem. Inf. Model.* **2010**, DOI: 10.1021/ci100202p.
(29) Varnek, A.; Kireeva, N.; Tetko, I. V.; Baskin, I. I.; Solov'ev, V. P. Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points? *J. Chem. Inf. Model.* **2007**, *47* (3), 1111−22.
(30) Zhokhova, N. I.; Baskin, I. I.; Palyulin, V. A.; Zefirov, A. N.; Zefirov, N. S. Fragment descriptors in QSPR: Application to magnetic susceptibility calculations. *J. Struct. Chem.* **2004**, *45* (4), 660−669.
(31) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45* (19), 4350−8.
(32) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 1912−28.
(33) Schneider, G. Virtual screening: an endless staircase? *Nat. Rev, Drug Discovery* **2010**, *9* (4), 273−6.
(34) Weislow, O. S.; Kiser, R.; Fine, D. L.; Bader, J.; Shoemaker, R. H.; Boyd, M. R. New soluble-formazan assay for HIV-1 cytopathic effects: application to high-flux screening of synthetic and natural products for AIDS-antiviral activity. *J. Natl. Cancer Inst.* **1989**, *81* (8), 577−86.
(35) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−54.
(36) Yan, X.; Han, J. In gSpan: Graph-Based Substructure Pattern Mining, ICDM'02 (Proc. of 2002 Int. Conf. on Data Mining), 2002; pp 721−724.
(37) Kuramochi, M.; Karypis, G. In Frequent subgraph discovery, ICDM 2001; pp 313−320.
(38) Deshpande, M.; Kuramochi, M.; Wale, N.; Karypis, G. Frequent substructure-based approaches for classifying chemical compounds. *Knowl. Data Eng., IEEE Trans.* **2005**, *17* (8), 1036−1050.
(39) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.* **2008**, *48* (1), 68−74.
(40) Swamidass, S. J.; Azencott, C. A.; Daily, K.; Baldi, P. A CROC stronger than ROC: Measuring, visualizing and optimizing early retrieval. *Bioinformatics* **2010**, *26* (10), 1348−56.
(41) Godden, J. W.; Stahura, F. L.; Bajorath, J. Anatomy of fingerprint search calculations on structurally diverse sets of active compounds. *J. Chem. Inf. Model.* **2005**, *45* (6), 1812−9.
(42) Baskin, I.; Varnek, A. Fragment descriptors in SAR/QSAR/QSPR studies, molecular similarity analysis and in virtual screening. In *Chemoinformatics Approaches to Virtual Screening*; Varnek, A., Tropsha, A., Eds.; Thomas Graham House: Cambridge, U.K., 2008; pp 1−43. DOI: 10.1002/chin.200920272.
(43) Arif, S. M.; Holliday, J. D.; Willett, P. Inverse frequency weighting of fragments for similarity-based virtual screening. *J. Chem. Inf. Model.* **2010**, DOI: 10.1021/ci1001235.
(44) Kliger, Y.; Gofer, E.; Wool, A.; Toporik, A.; Apatoff, A.; Olshansky, M. Predicting proteolytic sites in extracellular proteins: only halfway there. *Bioinformatics* **2008**, *24* (8), 1049−55.
(45) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197−206.
(46) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Kluwer Academic Publishers: The Netherlands, 2007.
(47) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983.
(48) Gobel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* **1994**, *18* (4), 309−17.
(49) Martin, L. C.; Gloor, G. B.; Dunn, S. D.; Wahl, L. M. Using information theory to search for co-evolving residues in proteins. *Bioinformatics* **2005**, *21* (22), 4116−24.
(50) Pazos, F.; Olmea, O.; Valencia, A. A graphical interface for correlated mutations and other protein structure prediction methods. *Comput. Appl. Biosci.* **1997**, *13* (3), 319−21.
(51) Wollenberg, K. R.; Atchley, W. R. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (7), 3288−91.
(52) Vicatos, S.; Reddy, B. V.; Kaznessis, Y. Prediction of distant residue contacts with the use of evolutionary information. *Proteins* **2005**, *58* (4), 935−49.
(53) Noivirt, O.; Eisenstein, M.; Horovitz, A. Detection and reduction of evolutionary noise in correlated mutation analysis. *Protein Eng. Des. Sel.* **2005**, *18* (5), 247−53.
(54) Kundrotas, P. J.; Alexov, E. G. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. *BMC Bioinf.* **2006**, *7*, 503.
(55) Gloor, G. B.; Martin, L. C.; Wahl, L. M.; Dunn, S. D. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* **2005**, *44* (19), 7156−65.
(56) Dutheil, J.; Pupko, T.; Jean-Marie, A.; Galtier, N. A model-based approach for detecting coevolving positions in a molecule. *Mol. Biol. Evol.* **2005**, *22* (9), 1919−28.
(57) Dunn, S. D.; Wahl, L. M.; Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **2008**, *24* (3), 333−40.
(58) Dimmic, M. W.; Hubisz, M. J.; Bustamante, C. D.; Nielsen, R. Detecting coevolving amino acid sites using Bayesian mutational mapping. *Bioinformatics* **2005**, *21* (Suppl 1), i126−35.
(59) Ashkenazy, H.; Kliger, Y. Reducing phylogenetic bias in correlated mutation analysis. *Protein Eng. Des. Sel.* **2010**, *23* (5), 321−6.
(60) Barnes, M. R.; Harland, L.; Foord, S. M.; Hall, M. D.; Dix, I.; Thomas, S.; Williams-Jones, B. I.; Brouwer, C. R. Lowering industry

firewalls: Pre-competitive informatics initiatives in drug discovery. *Nat. Rev. Drug Discovery* **2009**, *8* (9), 701−8.
(61) Weininger, D. SMILES, A chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31−36.
(62) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model.* **2006**, *46* (3), 991−8.
(63) Brennan, R. J.; Nikolskya, T.; Bureeva, S. Network and pathway analysis of compound-protein interactions. *Methods Mol. Biol.* **2009**, *575*, 225−47.
(64) Batista, J.; Godden, J. W.; Bajorath, J. Assessment of molecular similarity from the analysis of randomly generated structural fragment populations. *J. Chem. Inf. Model.* **2006**, *46* (5), 1937−44.
(65) Borgelt, C.; Berthold, M. R. In Mining Molecular Fragments: Finding Relevant Substructures of Molecules, Second IEEE International Conference on Data Mining (ICDM'02), 2002.
(66) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP—Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (3), 511−22.
(67) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38* (19), 2894−2896.
(68) Ormerod, A.; Willett, P.; Bawden, D. Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct.−Act. Relat.* **1989**, *8* (2), 115−129.
(69) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11* (23−24), 1046−53.
(70) Rish, I. In An empirical study of the naive Bayes classifier, IJCAI-01 workshop on Empirical Methods in AI, 2001; pp 41−46.
(71) Bar-Haim, S.; Aharon, A.; Ben-Moshe, T.; Marantz, Y.; Senderowitz, H. SeleX-CS: A new consensus scoring algorithm for hit discovery and lead optimization. *J. Chem. Inf. Model.* **2009**, *49* (3), 623−33.
(72) Vogt, M.; Godden, J. W.; Bajorath, J. Bayesian interpretation of a distance function for navigating high-dimensional descriptor spaces. *J. Chem. Inf. Model.* **2007**, *47* (1), 39−46.
(73) Bologa, C. G.; Revankar, C. M.; Young, S. M.; Edwards, B. S.; Arterburn, J. B.; Kiselyov, A. S.; Parker, M. A.; Tkachenko, S. E.; Savchuck, N. P.; Sklar, L. A.; Oprea, T. I.; Prossnitz, E. R. Virtual and biomolecular screening converge on a selective agonist for GPR30. *Nat. Chem. Biol.* **2006**, *2* (4), 207−12.