Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments

Madhavi Sastry,[‡] Jeffrey F. Lowrie,[†] Steven L. Dixon,[†] and Woody Sherman*[,†]

Schrödinger, 120 West 45th Street, 17th Floor, New York, New York 10036 and Schrödinger, Sanali Infopark, 8-2-120/113, Banjara Hills, Hyderabad 500034, Andhra Pradesh, India

Received February 10, 2010

A systematic virtual screening study on 11 pharmaceutically relevant targets has been conducted to investigate the interrelation between 8 two-dimensional (2D) fingerprinting methods, 13 atom-typing schemes, 13 bit scaling rules, and 12 similarity metrics using the new cheminformatics package Canvas. In total, 157 872 virtual screens were performed to assess the ability of each combination of parameters to identify actives in a database screen. In general, fingerprint methods, such as MOLPRINT2D, Radial, and Dendritic that encode information about local environment beyond simple linear paths outperformed other fingerprint methods. Atom-typing schemes with more specific information, such as Daylight, Mol2, and Carhart were generally superior to more generic atom-typing schemes. Enrichment factors across all targets were improved considerably with the best settings, although no single set of parameters performed optimally on all targets. The size of the addressable bit space for the fingerprints was also explored, and it was found to have a substantial impact on enrichments. Small bit spaces, such as 1024, resulted in many collisions and in a significant degradation in enrichments compared to larger bit spaces that avoid collisions.

INTRODUCTION

Virtual screening[1−3] is a vital element of modern drug discovery, and the debate continues[4−9] on the relative merits of approaches that incorporate three-, two-, and even one-dimensional[8] (3D, 2D, and 1D, respectively) chemical information and representations. While docking[10−14] is frequently applied when suitable structural models of the target are available, purely ligand-based techniques, such as pharmacophore matching,[15−18] shape-based screening,[19−22] and 2D fingerprint similarity[4,5,23,24] provide alternative and complementary approaches, particularly when a target structural model is not available, but active ligand molecules are at hand. Also, due to their relative high throughput, ligand-based methods are attractive when the number of compounds to screen is large and a fast turnaround is needed. Many of these and other virtual screening methods[25−29] may be available to a modeler, resulting in a multitude of choices, and decisions about which strategies to pursue may not be straightforward.

Though 3D approaches are routinely viewed as holding the greatest promise, Occam's razor ultimately governs many of the decisions in a drug discovery campaign, and 2D fingerprints continue to be widely used in industry, sometimes with more success than 3D shape or docking methods.[9] Yet even when the focus is narrowed to 2D fingerprints, there are still numerous possibilities to consider due to the wide variety of available fingerprinting methods, atom-typing schemes, bit scaling rules, and similarity metrics. This is an important consideration for users of the cheminformatics package Canvas,[30] where more than 10 000 combinations of these four types of variables are possible. While previous studies have not explored every conceivable combination of these parameters in large-scale virtual screening experiments, systematic investigation of reasonable subspaces have provided practical guidelines for modelers who are facing a bewildering array of choices.[31−34] Our intention in this work is to expand on the domain of previous studies by considering all of the aforementioned factors simultaneously.

In this work, we have carried out a very large number of screens, incorporating over 1000 active ligands that span 11 pharmaceutically important targets, and more than 24 000 decoys from the MDL Drug Data Report (MDDR).[35] The goal was to identify general trends, such as which fingerprint methods perform well irrespective of the values of other variables as well as specific combinations that are recommended to maximize the success of virtual screening efforts and specific combinations to avoid. All combinations of 8 fingerprint methods, 13 atom-typing schemes, 13 bit scaling rules, and 12 similarity metrics were explored. We present the aggregate results as well as an analysis of each parameter.

METHODS

Fingerprint Types. Table 1 summarizes the eight types of fingerprints studied in this paper, as implemented in the cheminformatics package Canvas.[30] With the exception of MACCS,[36] Canvas fingerprints are encoded by hashing each chemical pattern into an addressable space of user-controlled size and storing only the "on" bits. A 32-bit fingerprint (the default in Canvas) is therefore represented by a list of integers on the interval [1, 2^{32}], where 2^{32} is the size of the addressable space, and each integer represents the position of an "on" bit in this space. This sparse encoding contrasts with some other fingerprint implementations,[37−39] where hashing is

* Corresponding author. Telephone: 646-366-9555. Fax: 646-366-9550. E-mail: Woody.Sherman@schrodinger.com.
‡ Schrödinger, Andhra Pradesh, India.
† Schrödinger, New York, New York.

10.1021/ci100062n
Published on Web 05/07/2010

Table 1. Fingerprint Types

FP type	description
Linear	linear fragments + ring closures
Dendritic	linear and branched fragments
Radial	fragments that grow radially from each atom. Also known as extended connectivity fingerprints (ECFPs)[42]
Pairwise	pairs of atoms,[44] differentiated by type and the distance separating them: $Type_i$−$Type_j$-d_{ij}
Triplet	triplets of atoms, differentiated by type and the three distances separating them: $Type_i$-d_{ij}−$Type_j$-d_{jk}−$Type_k$-d_{ki}
Torsion	four consecutively bonded atoms,[45] differentiated by type: $Type_i$−$Type_j$−$Type_k$−$Type_l$
MOLPRINT2D	a radial-like fingerprint that encodes atom environments using lists of atom types located at different topological distances[46,47]
MACCS	SMARTS-based implementation of the MACCS structural keys[36]

done into a much smaller addressable space, such as 2^{10} (i.e., 10-bit), and explicit on/off values are stored. The advantage of using a large addressable space is that the likelihood of two different chemical features setting the same bit, also known as a *collision*, is exceedingly small. For example, with 32-bit linear fingerprints a collision typically occurs only once for every few thousand structures fingerprinted, whereas collisions are effectively eliminated altogether with 64-bit fingerprints. By comparison, use of a 10-bit fingerprint (i.e., an addressable space of 2^{10} or 1024) usually results in numerous collisions for even a single drug-like molecule. This point will be elaborated upon further in this work, and the implications for the retrieval of active molecules from a database will be discussed.

A more detailed description of the eight fingerprints shown in Table 1 is provided below. For further information on the characteristics and applications of 2D fingerprints, the reader is referred to any of several excellent sources.[4,5,23,24,31,40,41]

Linear. A structure is examined for all linear paths containing up to a user-defined number of bonds, which is seven by default, and a hashing operation is performed on a string-based description of each linear fragment to produce an integer bit address. To improve the description of rings without causing a massive proliferation in the number of fragments, the default maximum path is expanded from 7 to 14 bonds for ring closure only. This effectively encodes information in and around most ring systems, while producing only a fraction of the fragments that one would get if all paths containing 0−14 bonds were enumerated.

Dendritic. To encode both linear and branched features, a structure is decomposed into fragments consisting of linear paths and intersections of linear paths, with a default maximum of five bonds per path. Dendritic fingerprints incorporate no special treatment of rings.

Radial. Also known as extended connectivity fingerprints (ECFPs or FCFPs),[42] radial fingerprints entail growing a set of fragments radially from each heavy atom over a series of iterations. Using a variant of the Morgan algorithm,[43] each chemically unique fragment is mapped to a distinct integer value by hashing a description of the atoms and bonds within the fragment and the bonds that connect it to the surrounding region.

Pairwise. These fingerprints are based on the concept of an atom pair,[44] which is simply two atom types and the distance separating them: $Type_i$−$Type_j$-d_{ij}, where $Type_i \leq Type_j$. This representation is hashed, byte-by-byte, to an integer value. By default, all pairs of atoms in the molecule are considered, and distances are the number of bonds in the shortest path between each pair of atoms.

Triplet. An obvious extension of the Pairwise fingerprint, a triplet consists of a set of three atoms and the topological distances separating them: $Type_i$-d_{ij}−$Type_j$-d_{jk}−$Type_k$-d_{ki}. Since there are six different ways to order the atoms in a triplet, a canonicalization is performed to ensure that there are no bits in the fingerprint that differ only in a permutation of the atoms.

Torsion. Also known as topological torsions,[45] these are a special case of linear fingerprints in which every fragment consists of a linear path of four atoms that are differentiated by type: $Type_i$−$Type_j$−$Type_k$−$Type_l$. Because only fragments containing four atoms are enumerated, the number of on bits in a torsion fingerprint tends to be about an order of magnitude smaller than the number of on bits in a linear fingerprint.

MOLPRINT2D. Each heavy atom in a structure is characterized by an environment that consists of all other heavy atoms within a distance of two bonds.[46,47] A bit in the fingerprint is derived from a tabular data structure that stores the central heavy atom type and lists the atom types found at distances of one and two bonds. Each member of the list is encoded into a string of the form Type-*freq*(Type)-*d*, where *freq*(Type) is the number of times a given atom type is found at a distance *d* from the central atom. These terms are sorted by distance, then by type, and finally hashed to an integer that sets a bit in the fingerprint.

MACCS. These well-known structural keys[36] are encoded using a set of 155 SMARTS[48] patterns, which we note is 11 short of the 166 public keys defined in the MDL documentation. The missing keys correspond to features that are not easily encoded using SMARTS, such as "isotope" or definitions that are ambiguous, for example, "(&..)," which is defined only as, "and other rare features."

Atom-Typing Schemes. Hashed fingerprinting methods are sensitive to the atom-typing scheme used because this determines which atoms are treated as chemically distinct and, therefore, how many unique chemical features a structure possesses. Table 2 contains descriptions and shorthand codes for the 13 atom-typing schemes investigated. They range from purely generic (G), where each structure is essentially transformed to a single-bonded carbon skeleton, to highly specific treatments, such as Daylight invariant atom types, which differentiate by atomic number, formal charge, valence, and the numbers of hydrogen and non-hydrogen connections.[50]

Bit Scaling Rules. Most often, a fingerprint encodes only the presence or absence of a set of chemical features (i.e., binary values), but there are circumstances where it may be important to account for the number of times each feature appears. Furthermore, since systematic fragmentation of a structure can lead to a distribution of fragment frequencies where, e.g., larger fragments dominate, simple counts may have to be normalized in a way that limits the influence of the more prevalent fragments. Table 3 contains a variety of bit scaling rules that are designed to address these scenarios. When scaling is applied, the "on" bit value for a given feature

Table 2. Atom-Typing Schemes

code	description
G	all atoms and bonds are equivalent (generic)
HB	atoms are distinguished by whether they are hydrogen bond (HB) acceptors or donors; all bonds are equivalent
Hybrid	atoms are distinguished by hybridization state; all bonds are equivalent
Fn	atoms are distinguished by functional type: {H}, {C}, {F,Cl}, {Br,I}, {N,O}, {S}, {other}; all bonds are equivalent
Mol2	Sybyl Mol2 atom types;[49] all bonds are equivalent
TXHB	atoms are distinguished by whether they are terminal, halogen, HB acceptor/donor; all bonds are equivalent
Element	atomic number and bond order
ElemR	atomic number and bond order; aromatic distinguished from nonaromatic
ElemRC	atomic number and bond order; aromatic distinguished from nonaromatic; cyclic distinguished from acyclic aliphatic atoms
RTXHB	atoms are distinguished by ring size, aromaticity, whether terminal, whether halogen, HB acceptor/donor; bonds are distinguished by bond order
Carhart	Carhart atom types (atom-pairs approach);[44] all bonds are equivalent
Daylight	Daylight invariant atom types;[50] bonds are distinguished by bond order
Estate	Estate[51] atom types; the EState coordinate is divided into bins of width 0.25, and each atom is typed according to the bin in which its EState value falls; all bonds are equivalent

Table 3. Bit Scaling Rules

code	description
none	no scaling
F	raw feature counts
F1	scale counts by feature size to unity
FF	scale counts by feature size to feature size
FM	scale counts by feature size to molecule size
F^2	squares of raw feature counts
$F^2$1	scale squares of counts by feature size to unity
F^2F	scale squares of counts by feature size to feature size
F^2M	scale squares of counts by features size to molecule size
$F^{1/2}$	square roots of raw feature counts
$F^{1/2}$1	scale square roots of counts by feature size to unity
$F^{1/2}$F	scale square roots of counts by feature size to feature size
$F^{1/2}$M	scale square roots of counts by features size to molecule size

is replaced by a floating point value, thus, producing a nonbinary fingerprint.

Although Table 3 is a useful reference, the brief descriptions therein are not sufficient to fully understand how each scaling rule is applied. Accordingly, we shall examine the F1 scaling rule in greater detail, which will make the meanings of the others clear.

Scale Counts by Feature Size to Unity (F1). Feature size is the number of atoms in a feature, and there are typically many features of a given size in a given structure. With this scaling rule, the sum of the scaled bit values associated with all features containing n atoms will add up to 1.0. This is true for each applicable value of n. So, if a linear fingerprint produces 5 unique fragments of 7 atoms, and the raw counts of those fragments are 2, 1, 3, 1, 3, then this option will yield scaled bit values of 2/10, 1/10, 3/10, 1/10, 3/10. The objective of this bit scaling is for fragments of each size to have the same overall impact on the fingerprint.

Scale Counts by Feature Size to Feature Size (FF). Here, the sum of the scaled bit values for all features of n atoms will add up to n. With the previous example, where the feature size was 7, the scaled values would, therefore, be (7·2)/10, (7·1)/10, (7·3)/10, (7·1)/10, (7·3)/10.

Similarity Metrics. Canvas provides 24 different indices to measure the similarity or distance between a pair of structures. We investigated 12 metrics in detail (Buser, Dice, Dixon, Euclidean, Hamann, Kulczynski, Pearson, Petke, Soergel, Tanimoto, Variance, and Yule). Table 4 contains the formulas for these indices, expressed in terms of the following intermediate quantities, which are functions of the bit values α_i and β_i in fingerprints F_A and F_B, respectively:

$$a = \sum_{i \in F_A} \alpha_i^2 \qquad \text{(number of on bits in } F_A\text{)}$$

$$b = \sum_{i \in F_B} \beta_i^2 \qquad \text{(number of on bits in } F_B\text{)}$$

$$c = \sum_{i \in F_A \cap F_B} \alpha_i \beta_i \qquad \text{(number of on bits shared by } F_A \text{ and } F_B\text{)}$$

$$d = \sum_{i \in F_A \cap F_B} (1 - \alpha_i)(1 - \beta_i) \qquad \text{(number off bits shared by } F_A \text{ and } F_B\text{)}$$

These expressions are generalized for scaled bit values, but they reduce to the simple counts of "on" and "off" bits indicated in parentheses when 0/1 values are used.

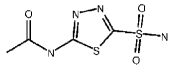
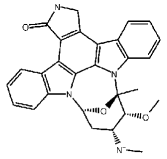
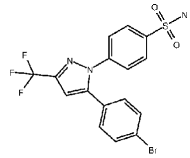
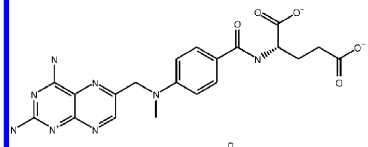
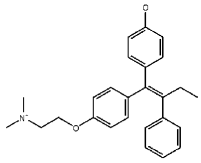
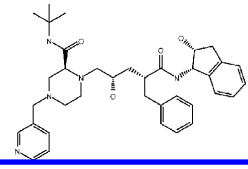
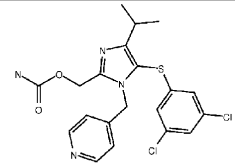
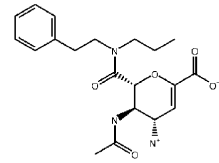
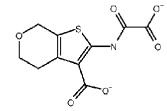
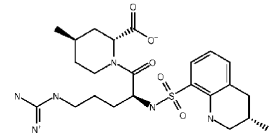
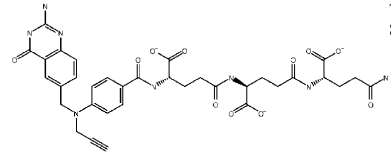
In Table 4, $A = (a - c)$, $B = (b - c)$, and $N = [a + b - c + \min(d,10\,000)]$. The Dixon metric was developed by an author of this work and can be reduced to the product of (1 − Tanimoto) and the squared Euclidean distance.[24] It was devised to overcome weaknesses in its two component indices, which are mathematically inclined to yield higher dissimilarity within collections of either small (Tanimoto) or large (Euclidean) structures. These biases tend to cancel out when a product is utilized.

Data Set. *Target Ligand Set.* The 11 targets from the McGaughey et al. work[9] were used. Ligand queries were extracted from the PDB structures (see Table 5), and the appropriate bond orders and formal charges were assigned. The targets were originally chosen by McGaughey et al. based on having at least one high-resolution crystal structure, a large

Table 4. Similarity/Distance Indices

name	type	formula
Buser	similarity	$((cd)^{1/2} + c)/((cd)^{1/2} + a + b - c)$
Dice	similarity	$2c/(a + b)$
Dixon	distance	$(a + b - 2c)^2/(a + b - c)$
Euclidean	distance	$(a + b - 2c)^{1/2}$
Hamann	similarity	$(c + d - A - B)/N$
Kulczynski	similarity	$0.5(c/a + c/b)$
Pearson	similarity	$(cd - AB)/(ab(A + d)(B + d))^{1/2}$
Petke	similarity	$c/\max(a,b)$
Soergel	distance	$(A + B)/(A + B + c)$
Tanimoto	similarity	$c/(a + b - c)$
Variance	distance	$(A + B)/(4N)$
Yule	similarity	$(c \cdot d - A \cdot B)/(c \cdot d + A \cdot B)$

Table 5. Targets Used in This Study along with Reference PDB Code and Number of Active Compounds[a]

Query	Target	Reference PDB Code	Number of Actives
	Carbonic Anhydrase I (CA)	1azm	80
	Cyclin-dependent Kinase 2 (CDK2)	1aq1	77
	Cyclooxygenase 2 (COX2)	1cx2	257
	Dihydrofolate Reductase (DHFR)	3dfr	26
	Estrogen Receptor Alpha (ER)	3ert	74
	HIV Protease (HIVpr)	1hsh	136
	HIV Reverse Transcriptase (HIVrt)	1ep4	149
	Neuraminidase (NA)	1aq4	12
	Protein Tyrosine Phosphatase 1B (PTP1B)	1c87	8
	Thrombin (Throm)	1dwc	200
	Thymidylate Synthase (TS)	2bbq	31

[a] CDK2, neuraminidase, and PTP1B active compounds were not explicitly labeled in the MDDR, so the original authors either used surrogate ligands or did a similarity search.

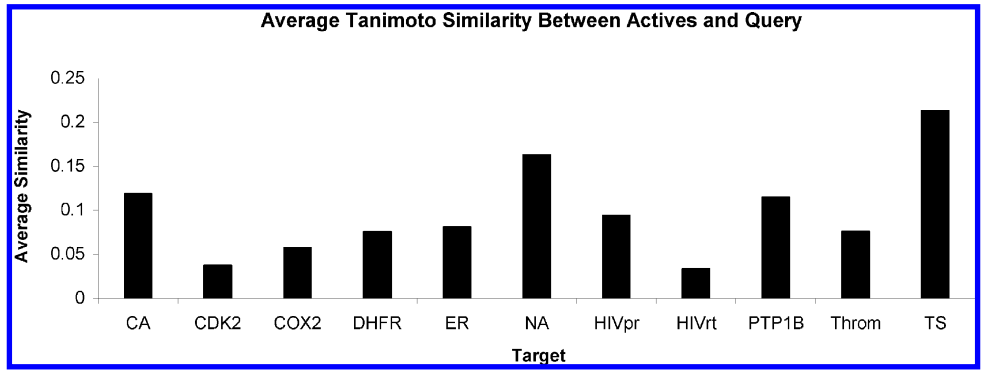


Figure 1. Average similarity of the active compounds to the query using linear fingerprints, Daylight atom types, no bit scaling, and Tanimoto similarity.

number of structurally diverse active compounds in the MDDR, the inclusion of only a single representative target for a given enzyme family, and spanning a diverse set of active sites (i.e., hydrophobic, hydrophilic, small, large, etc.). The active compounds were taken from the MDDR using queries for the target names. Figure 1 shows the average Tanimoto similarity between the actives for each target using linear fingerprints with Daylight atom types and no feature scaling.

Database Compounds. The database compounds were taken from the MDDR, as described in McGaughey et al.[9] The initial database of approximately 129 000 compounds was clustered, and a representative structure from each cluster was chosen. Molecules with greater than 80 nonhydrogen atoms were removed. The 24 580 remaining compounds were used as decoy molecules in this work.

Screening and Enrichment Calculations. Screens were run for each target with the ligand from Table 5 as the query. Fingerprints for each database compound (actives and decoys) were computed with the same set of parameters as the query. All combinations of fingerprint types, atom-typing schemes, bit scalings, and similarity/distance metrics (Tables 1−4, respectively) were explored.

For enrichment calculations we report both the enrichment factor for the top 1% of the database, EF(1%), as well as the Boltzmann-enhanced discrimination of the receiver operating characteristic, BEDROC (Truchon and Bayly, 2007).[52] We used $\alpha = 160.9$ for the BEDROC calculation, which corresponds to 80% of the BEDROC score being accounted for in the top 1% of the database screen. While there are differences between BEDROC and a traditional enrichment factor calculation, as discussed by Truchon et al., for much of this work we use the EF(1%) metric because it is more common and can be compared directly to results from other works.

RESULTS AND DISCUSSION

A total of 1196 different fingerprints were computed for each molecule (7 hashed fingerprint types × 13 atom-typing schemes × 13 bit scaling rules plus an additional 13 feature scalings for MACCS), and 12 similarity/distance metrics were applied for the database screening, resulting in 14 352 combinations of settings for each of the 11 targets. In total, 157 872 screens were performed on a database consisting of 24 580 decoys from the MDDR combined with the known actives for each target. The best average enrichment (EF(1%)) across the 11 targets for a single set of parameters was 35.1 (MOLPRINT2D, ElemRC atom types, no bit scaling, and Buser similarity). This is a substantial improvement over the average EF(1%) of 15.6 observed across all combinations of parameters. It is also significantly better than the value of 19.8 obtained by McGaughey et al. using default Daylight fingerprints on the same data set. The worst EF(1%) across the 11 targets for a single set of parameters was 0.0, which occurred for pairwise fingerprints with G, HB, Fn, or TXHB atomtypes, FF, $F^2$1, F^2F, or $F^{1/2}$ scalings, and Yule or Kulzcynski metrics. Similarly, triplet fingerprints gave zero average enrichments with G or TXHB atomtypes, F, F^2, or $F^{1/2}$ scalings, and Kulczynski metric. This highlights that the choice of parameters can have a substantial effect on enrichment factors. Below, we examine the parameters in more detail and provide recommendations for improving 2D virtual screening enrichments.

Addressable Bit Space. For reference, 10-bit unscaled Linear fingerprints, which have an addressable space of 1024 (the default for Daylight fingerprints), showed an average EF(1%) across the 11 targets of 20.7 using Tanimoto similarity, in rough agreement with the value of 19.8 obtained with Daylight fingerprints by McGaughey et al.[9] on the same data set. Note that the value of 19.8 from McGaughey et al. was obtained using the default minimum/maximum path of 0/7, whereas an improved enrichment value of 24.5 was obtained in that work with a 3/10 configuration. In this work, we use the default path lengths for all fingerprints. Using a 32-bit (i.e., 2^{32} total accessible bits) linear fingerprint, as is default in Canvas, boosts the average enrichment to 32.5. This improvement is because the collision rate in 32-bit fingerprints is essentially 0, whereas 10-bit fingerprints yield a significant number of collisions resulting in a loss of information.

To quantify collisions, we calculated the total number of "on" bits in the fingerprint for each query structure in Table 5 using 2^{10}, 2^{32}, and 2^{64} addressable spaces with a single set of parameters (linear fingerprints with Daylight atom typing, no feature scaling, and Tanimoto similarity). By definition, the number of on bits will increase with the size of the addressable space until there are no collisions. As seen in Table 6, the number of on bits increases for all query molecules going from 2^{10} to 2^{32}, while the 2^{32} and 2^{64} results are identical, suggesting that a 2^{32} addressable bit space is large enough to ensure that in most cases there will be no collisions. The difference in the number of on bits between the 2^{10} and 2^{32} addressable spaces ranges from 3.4% for CA (four additional on bits) to 280% for CDK2, where over 93% of the 2^{10} addresses are occupied. Staurosporine, the CDK2 query ligand, highlights the fact that collisions can result in exceedingly high measured similarities between any two large structures, regardless of whether those structures actually resemble each other. The effect is not as dramatic for smaller molecules, but collisions always result in a loss of information, and they serve no constructive purpose other than to reduce fingerprint storage requirements.

Table 6. The Effect of the Addressable Bit Space on Enrichment[a]

target	heavy atoms	2^{10} space "on" bits	2^{32} space "on" bits	2^{64} space "on" bits	2^{10} space EF(1%)	2^{32} space EF(1%)
CA	13	116	120	120	47.5	52.5
CDK2	35	953	2665	2665	7.8	11.7
COX2	26	264	303	303	10.1	18.7
DHFR	33	371	483	483	15.4	38.4
ER	29	178	193	193	10.8	10.8
HIVpr	45	504	694	694	5.9	28.7
HIVrt	29	337	408	408	2.0	3.4
NA	28	322	371	371	25.0	41.6
PTP1B	18	279	332	332	50.0	50.0
thrombin	35	462	607	607	4.5	30.5
TS	53	439	569	569	48.4	70.9
average	31.3	384.1	613.2	613.2	20.7	32.5

[a] Bit counts are reported for the query structures in Table 5 using linear fingerprints, Daylight atom types, and no bit scaling. EF(1%) data were obtained using Tanimoto similarities. Enrichments for 64-bit fingerprints are omitted since they are identical to those obtained for 32-bit fingerprints.

The consequences of high collision rates in 10-bit fingerprints are manifested as a significant degradation in average enrichment across the 11 targets (from 32.5 to 20.7). In the extreme case of thrombin, the EF(1%) decreases from to 30.5 to 4.5, which translates to a loss of 52 out of 61 actives in top 1%. However, in some cases with small ligands that have low collision rates, such as ER and PTP1B, there is no degradation in enrichment.

The effect of saturating the address space will be more significant with fingerprints that set more bits. Using Daylight atom typing, the fingerprints that set the most bits are dendritic, linear, and triplet, which set an average across the 11 query molecules of 343, 613, and 3,974 bits, respectively. In the extreme case, triplet fingerprints on the TS query produce 16 472 on bits. It may be noted that the TS query is large with polyglutamyl and folate moieties (Table 5). The saturation issue with a small addressable space will be less of an issue for fingerprints that set a small number of bits, such as MOLPRINT2D, MACCS, and torsion, which set 25, 44, and 53 bits, respectively. In fact, in no case does MOLPRINT2D set more than 37 bits for the 11 query molecules. Radial and pairwise fingerprints lie between these extremes, with an average of 94 and 253 on bits for the 11 query molecules.

Table 7. Average and Best Enrichments for Each Fingerprint Across the 11 Targets Varying Atom Types, Bit Scalings, and Metrics

fingerprint	average of all settings EF(1%)	best single setting EF(1%)	average of all settings BEDROC(α = 160.9)	best single setting BEDROC(α = 160.9)
dendritic	16.2	34.7	0.16	0.35
linear	14.5	33.5	0.15	0.33
MACCS	7.3	21.6	0.07	0.22
MOLPRINT2D	22.2	35.1	0.22	0.34
pairwise	13.2	29.5	0.13	0.28
radial	13.3	33.8	0.13	0.33
torsion	15.3	34.0	0.15	0.33
triplet	15.3	34.9	0.15	0.34

Fingerprint Type. Average EF(1%) and BEDROC(α = 160.9) for the different fingerprint types across the 11 targets are shown in Table 7. Results are presented for the average over all settings and for the best single set of parameters (atom type, bit scaling, and metric) for each fingerprint type. MOLPRINT2D shows the highest enrichment for a single choice of settings and performs significantly better than all other fingerprint types on average across all settings. The first point highlights the significant improvements that can be obtained with a good fingerprint type and optimized parameters. The latter finding suggests that MOLPRINT2D is a robust fingerprint method that has the greatest chance of performing well with any settings.

Figure 2 shows the best EF(1%) for each fingerprint type across the 11 targets, and Table 8 summarizes the results. It can be seen that the qualitative trend in the enrichments across the targets is similar for all the fingerprint types. For example, all fingerprint types perform relatively poorly on HIV reverse transcriptase (HIVrt), which is a notoriously difficult target for virtual screening methods because of the diversity of actives (see Figure 1). McGaughey et al.[9] found that across all the methods they studied, which included both 2D and 3D approaches, the results for HIVrt were consistently the worst among the 11 targets.

The HIVrt results in Figure 2, while low compared to other targets, show that considerable improvements can be made with the right choice of parameters. The best enrichment for HIVrt is 10.7, which comes from using MOLPRINT2D fingerprint types with Hybr or Carhart atom types, FM scaling, and Euclidean or variance metrics. This enrichment is higher than the best value that McGaughey et al.[9] obtained from any of the 2D or 3D methods (the SQW method performed best with an EF(1%) of 5.4) and significantly better than the 3.4 enrichment reported using Daylight fingerprints with Daylight atom typing and a 3/10 path configuration. A more detailed analysis of the atom typing for HIVrt and other targets will be presented in the next section.

Other targets also show large spreads in enrichment results for the different fingerprint types. For example, the carbonic anhydrase (CA) enrichment values range from 50.0 with MACCS (scaling or F1) to 80.0 with MOLPRINT2D (Carhart, Estate, or Daylight atom types and most scalings and metrics). A representative set of actives for CA showing

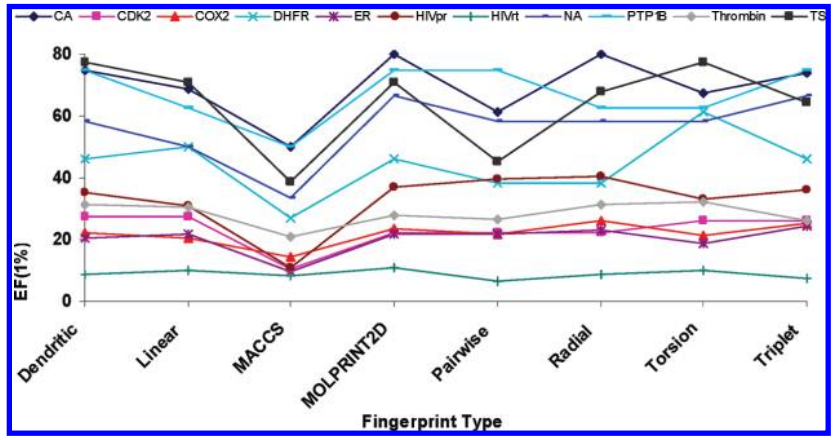


Figure 2. Best enrichment for each fingerprint across all targets.

Table 8. Best EF(1%) for Each Target

	dendritic	linear	MACCS	MOLPRINT2D	pairwise	radial	torsion	triplet
CA	75.0	68.7	50.0	80.0	61.2	80.0	67.5	73.7
CDK2	27.3	27.3	10.4	22.1	22.1	22.1	26.0	26.0
COX2	22.2	20.2	14.4	23.3	21.8	26.1	21.4	25.3
DHFR	46.1	50.0	26.9	46.1	38.4	38.4	61.5	46.1
ER	20.3	21.6	9.5	21.6	21.6	23.0	18.9	24.3
HIVpr	35.3	30.9	11.0	36.7	39.7	40.4	33.1	36.0
HIVrt	8.7	10.1	8.1	10.7	6.7	8.7	10.1	7.4
NA	58.3	50.0	33.3	66.6	58.3	58.3	58.3	66.6
PTP1B	75.0	62.5	50.0	75.0	75.0	62.5	62.5	75.0
thrombin	31.5	30.5	21.0	28.0	26.5	31.5	32.0	26.0
TS	77.4	70.9	38.7	70.9	45.1	67.7	77.4	64.5
mean	43.4	40.2	24.8	43.7	37.9	41.7	42.6	42.8
median	35.3	30.9	21.0	36.7	38.4	38.4	33.1	36.0
SD	24.6	21.1	16.0	25.1	20.7	22.4	23.2	23.6

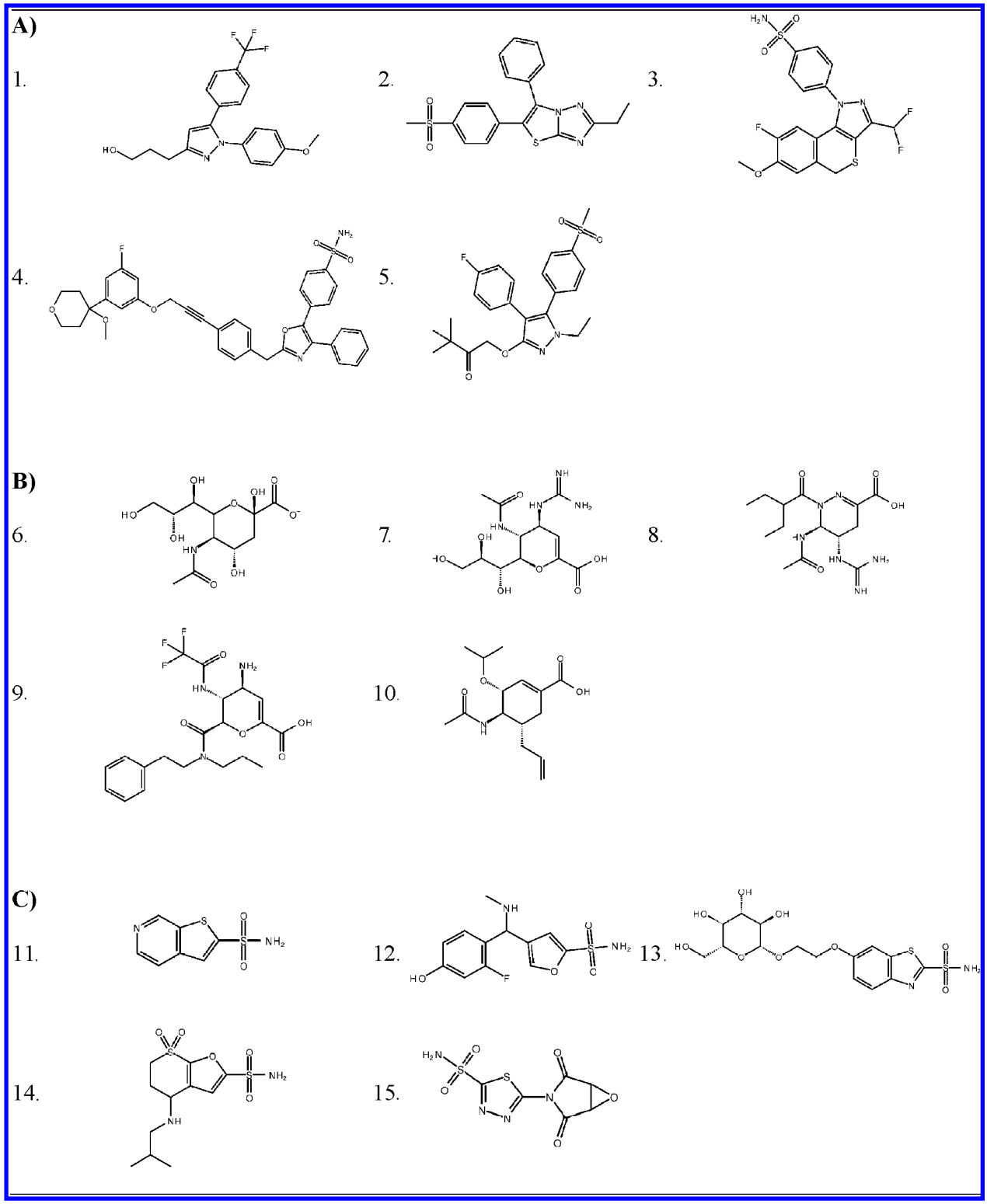


Figure 3. (A) and (B) Representative high-scoring actives for COX2 and neuraminidase, respectively, with RTXHB atom typing. (C) Representative high-scoring actives for carbonic anhydrase (CA) with Daylight atom typing.

the diverse ring systems is shown in Figure 3. Another example with a large variation in enrichments is TS, where unscaled dendritic and torsion fingerprints with Estate atom types perform best with an EF(1%) of 77.4. Interestingly, different combinations of parameters for dendritic and torsion fingerprints can lead to an enrichment factor of zero, again illustrating the complexities of the parameter space and the importance of carefully choosing the right combination of fingerprint methods and associated parameters.

Atom-Typing Schemes. Figure 4 shows the EF(1%) for the different atom-typing schemes across the 11 targets. It is evident that no single atom-typing method performs consistently better than all others across the different fingerprint types and targets. In general, the more discriminating atom-typing schemes, such as ElemRC, ElemR, Carhart, Daylight, and Mol2, perform the best. The average EF(1%) over the targets for the best settings for these atom types is 35.1, 34.3, 34.7, 34.3, and 34.2, respectively. The

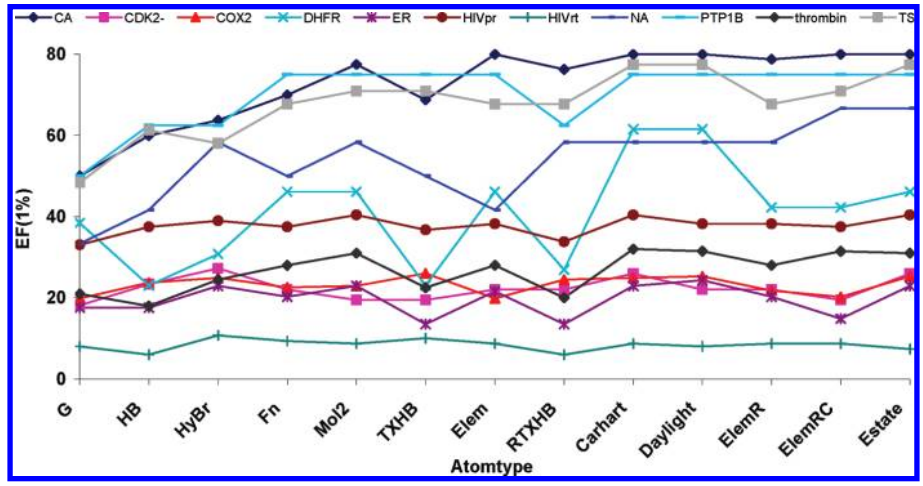


Figure 4. Best enrichment for each atom-typing scheme across all targets.

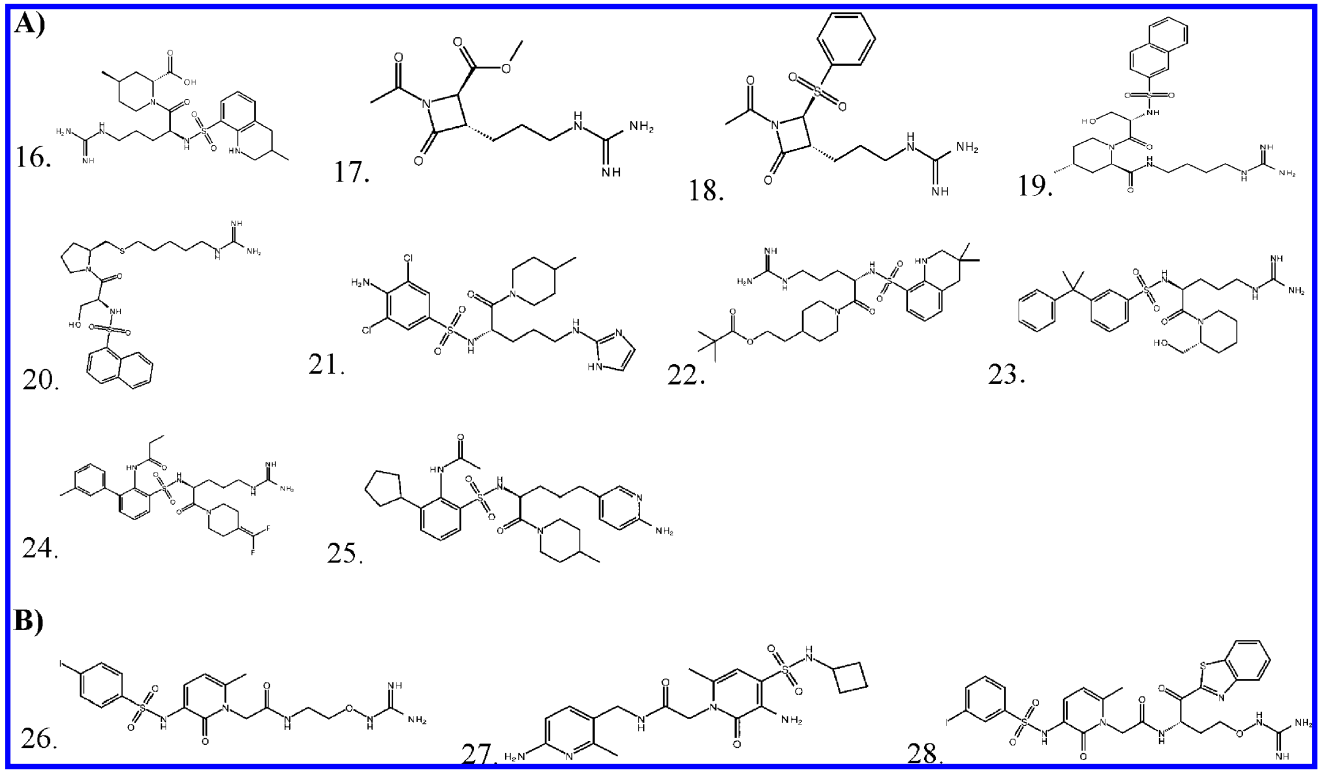


Figure 5. (A) Top 10 thrombin actives with Daylight atom types. Compounds **16**, **19**, **22**, **24**, and **25** are found in the top 1% with both Daylight and G atom typing. (B) Thrombin actives found in the top 1% of the database with G atom types but not with Daylight atom types. The representative molecules are obtained using MOLPRINT2D fingerprint type and no scaling.

more generic atom-typing schemes, such as G and HB, perform worst, with average enrichments of 27.1 and 24.7, respectively. However, due to their generic nature, they are able to pick up on gross characteristics of the molecules, which could be useful for retrieving more diverse virtual screening compounds or for scaffold hopping. Below we examine the kinds of structures that are scored highly using different atom-typing schemes.

As mentioned in the previous section, HIVrt results can be improved considerably with the right choice of parameters. The atom-typing scheme is particularly important with HIVrt due to the diversity of the actives. The best enrichment for HIVrt is 10.7, which comes from using MOLPRINT2D fingerprints with Hybr atom types and FM or $F^{1/2}M$ feature scalings. Hybr atom types also perform the best for linear fingerprints, with an enrichment of 10.1, whereas more specific atom-typing schemes produce poor enrichment results, such as Daylight (EF(1%) = 5.7) and Carhart (EF(1%) = 4.0) using the best feature scaling and similarity metric for those atom types. However, overly generic atom types result in degradation in performance because there is not enough information to capture the key features of the query. This is highlighted by the lack of retrieving any active molecules in the top 1% of the database with linear fingerprints and the most general atom typing (G) with Pearson or Yule metrics.

Thrombin, COX2, and neuraminidase also show improvements when certain generic atom types are employed. Figure 5 contains the top 10 hits for thrombin using MOLPRINT2D fingerprints, no scaling, and either Daylight or generic (G) atom typing. The G atom typing is able to score molecules highly that have direct sulfonamide linkages between two

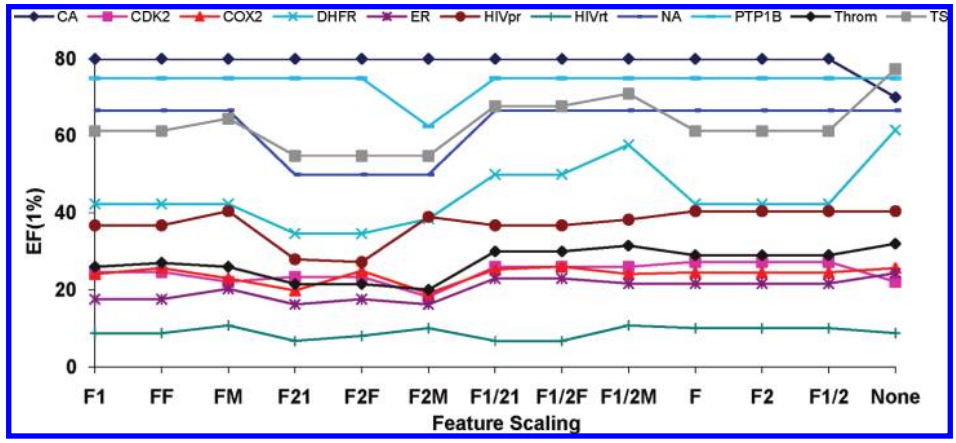


Figure 6. Best enrichment for each feature scaling across all targets.

ring systems (compounds **26**, **27**, and **28**), which is not present in the query molecule. On the other hand, Daylight atom types produce higher enrichments, but there are no top scoring molecules with a direct sulfonamide linkage between two rings, suggesting that less diverse classes of compounds are being retrieved. For COX2 and neuraminidase, schemes that do not distinguish specific atom types but do distinguish specific features, such as ring size, aromaticity, HB acceptor/donor, and ionization potential, perform best. A representative set of active hits for COX2 and neuraminidase are shown in Figure 3. This can be contrasted with CA, for example, where the sulfonamide is present in a similar environment for all actives and, therefore, more specific atom-typing schemes, such as Mol2 and Daylight perform best.

Bit Scaling. Figure 6 shows the EF(1%) for the different bit scalings across the 11 targets. As with the previously explored parameters, there is no single best choice across all targets. However, there are significant variations observed for some targets, such as DHFR, TS, and neuraminidase, whereas HIVrt, ER, CDK2, and COX2 show little sensitivity to the feature scaling option. The lack of a clear trend in the number of features within each set of actives is the most likely reason for the observed insensitivity to feature scaling. Furthermore, the latter set of targets are among the most challenging in the set, and many of the active compounds share little similarity with the query, as can be seen from Figure 7.

Using squares of feature counts scaled to unity ($F^2$1), feature size (F^2F), or molecule size (F^2M), tends to degrade enrichments, as observed in Figure 6. Across all targets, simply excluding feature scaling (None) performs well. However, in some cases feature scaling can result in the retrieval of additional compounds, as seen with PTP1B in Figure 8. Here, scaling the square root of feature size either to unity ($F^{1/2}$1) or feature size ($F^{1/2}$F) results in the greatest enrichment.

In other cases, like CDK2 and HIVrt, there is less consistency between the number of features in the query and the actives, explaining why down-weighting the feature count improves enrichments. Also, the size of the actives varies less for the targets where increased weighting of the counts improves enrichments. For example, neuraminidase and TS have a standard deviation in the molecular weight of 48.8 and 105.8, respectively, and these targets show an improvement by up-weighting the feature count. On the other hand, actives of HIVrt and CDK2 exhibit a larger molecular weight variation of 225.9 and 190.0, respectively, and get better results by down-weighting the feature counts. While the fingerprint screening results are not directly related to the molecular weight, they are related to the number of features of each type, which tends to be correlated with molecular weight. This suggests that up-weighting the feature count is recommended when active compounds of a similar size to the query are desired, whereas to find actives that span a broader range of size the feature counts should be down-weighted. However, there will be exceptions to this since the problem is multidimensional and enrichments will vary depending on which aspects dominate in the data set.

Metrics. Figure 9 shows the best enrichment for 12 metrics from Table 4 (Buser, Dice, Dixon, Euclidean, Hamann, Kulczynski, Pearson, Petke, Soergel, Tanimoto, Variance, and Yule) using the best settings of the other parameters across the 11 targets. There is relatively little variation in the enrichments compared to the variation seen with different fingerprint types, atom typing, and feature scalings presented above. For the majority of the targets (6/11), there is less than a ±5 variation in EF(1%). This means that for any given metric, there is a set of parameters that can be used to produce good enrichment results. The greatest difference in the similarity metrics is observed in DHFR and TS, where the EF(1%) varies by over 15. The distance metric Euclidean performs the worst with average EF(1%) of 28.2 and 28.5 for DHFR and TS, respectively. Variance and Hamann also perform relatively poorly, with an average EF(1%) of 28.5.

While above results for 12 metrics show relatively little variation in enrichment, there are some metrics that perform significantly worse than others, and this effect can be exaggerated when looking at a specific set of parameters rather than the best settings. To understand the effect of the metrics in more detail, we looked at a single set of values for the other parameters. Figure 10 shows the enrichments for linear fingerprints, Daylight atom types, and no feature scaling. Based on the treatment of on and off bits, the metrics are classified into one of three groups. The first class of metrics, called "ON", consider only on bits (Dice, Kulczynski, Petke, Tanimoto, and Soergel). The second class of metrics, called "OFF", include off bits explicitly in addition to on bits (Buser, Hamann, Pearson, Variance, and Yule). The final class of metrics, called "Hybrid" or "H", includes off bits implicitly or partially (Euclidean and Dixon).

As can be seen in Figure 10, there is a large range in enrichments with the results being roughly separable into

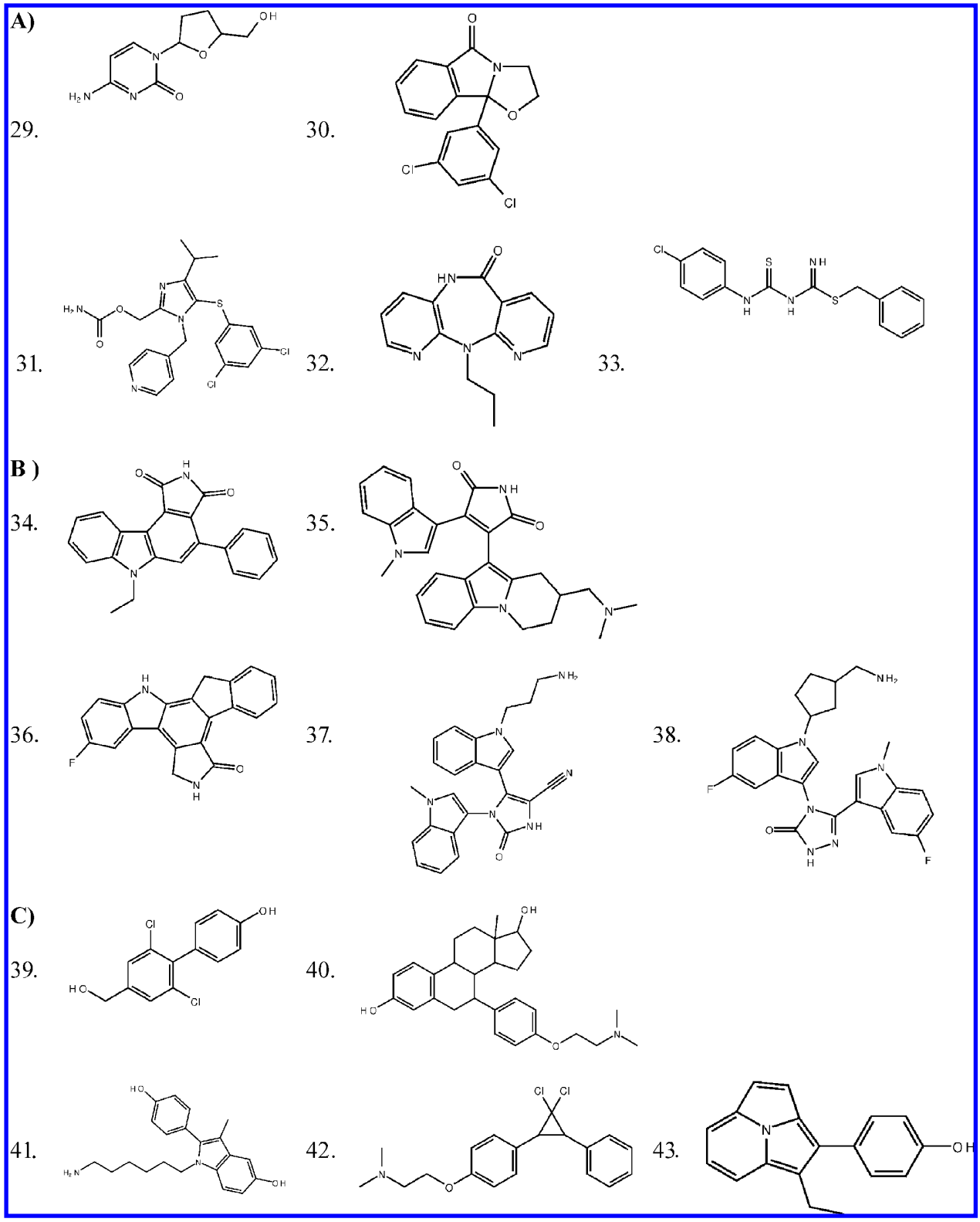


Figure 7. Representative active hits in the top 1% of the database for targets with average enrichments that do not vary significantly with bit scaling. (A) HIVrt, (B) CDK2, and (C) ER. These actives were retrieved with the combination of settings (fingerprint, atom type, bit scaling, and metric), as follows: HIVrt (MOLPRINT2D, HyBr, FM, Euclidean), CDK2 (dendritic, HyBr, F, Dixon), and COX2 (radial, RTXHB, $F^{1/2}F$, Petke).

good and bad performers. Euclidean, Hamann, and Variance perform particularly poorly, while Dixon performs slightly better. The intermediate performance of Dixon is expected, since it is the product of (1 − Tanimoto) and the Euclidean distance squared. The metrics that perform best are generally the similarity metrics belonging to the "ON" class, with Dice, Petke, Soergel, and Tanimoto topping the list and with all having an average EF(1%) greater than 32.

The reason for the poor performance of the OFF and Hybrid metrics can be understood from their formulas in Table 4 and the large addressable space used for Canvas fingerprints. A typical molecule sets a few hundred or a few

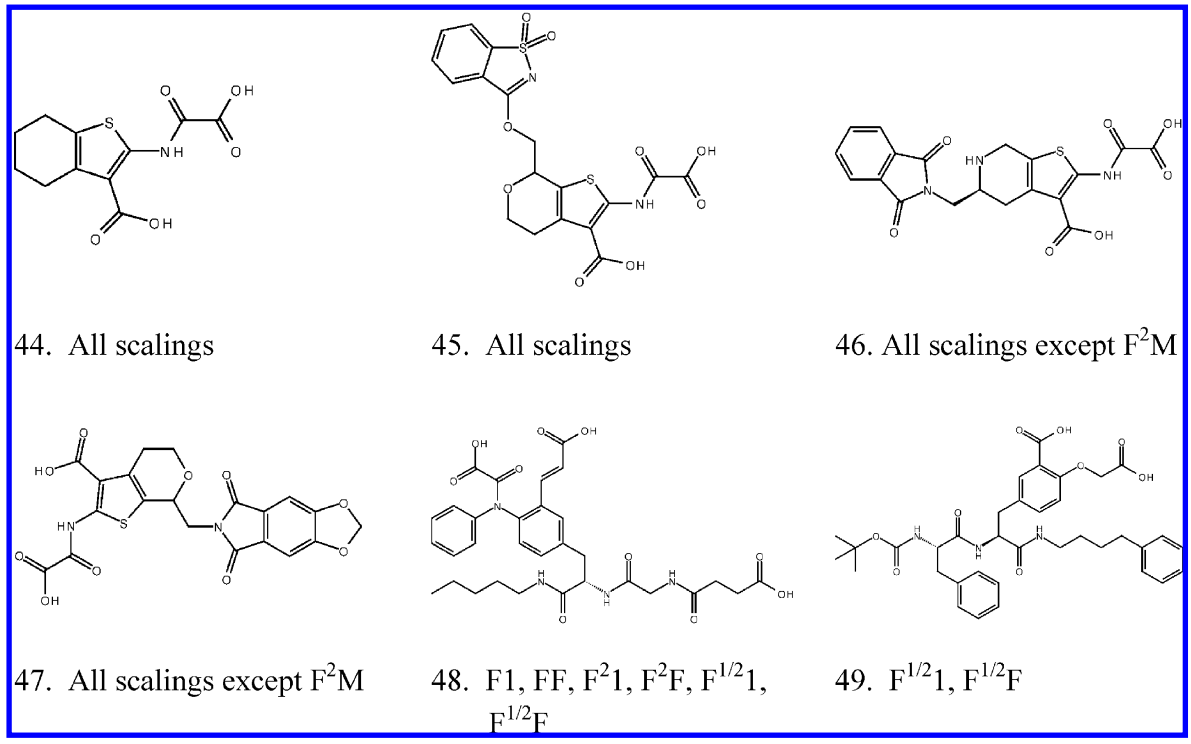


Figure 8. PTP1B actives retrieved in the top 1% of the database identified using different feature scalings (MOLPRINT2D fingerprints, Fn atom types, and Dixon metric). The scalings that retrieved an active are listed beneath the corresponding structure. Compounds **44**−**47** are retrieved by most scalings, including no feature scaling. Scaling the square roots ($F^{1/2}$) of feature counts to unity or feature size is needed to retrieve compounds **48** and **49**. For **49**, only features scalings that include a square root retrieve this compound, presumably because the impact of the additional features relative to the query needs to be attenuated.

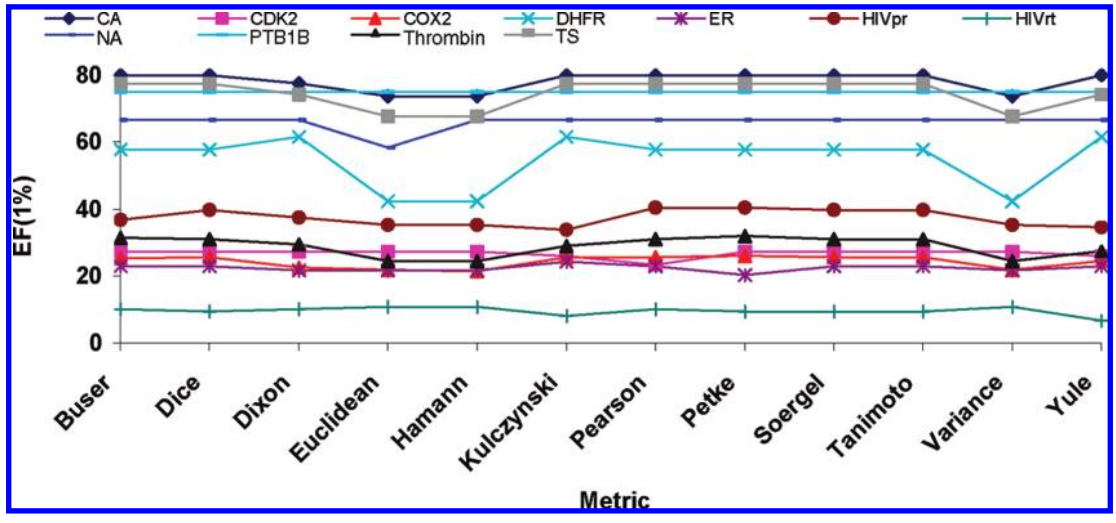


Figure 9. Best enrichment for 12 different metrics across all targets using the best setting of the other parameters for each metric.

thousand bits (see Table 6), which leaves billions of unset bits in the 10^{32} addressable space that will be in common between any given molecule and the query. This is not an issue with smaller bit spaces, such as 10-bit fingerprints, where about half of the bits are set for any given molecule, as seen in Table 6 above. The fact that very small addressable spaces contain a significant number of collisions in the on bits may end up benefiting the metrics that explicitly account for the off bits, as they may result in additional signal. However, as was discussed above, smaller bit spaces result in a loss of information, which is undesirable. Interestingly, not all of the OFF metrics perform poorly, with Buser and Pearson having averages across the 11 targets of 29.9 and 31.6, respectively. These metrics are complex (see Table 4) with many potential contributors to the relatively good performance compared to the other OFF metrics and will be the focus of future work.

Average Enrichments Across All Targets. Figure 11 shows trends in the performance of the fingerprint/atom-type combinations. Pairwise, torsion, linear, and triplet fingerprints perform well with some atom types but consistently underperform MOLPRINT2D, radial, and dendritic. Figure 11 also shows that the more specific atom-typing schemes (Mol2, Carhart, and Daylight) perform best on average, whereas those that treat all atoms and bonds equivalently (G) or distinguish only hydrogen-bond donors and acceptors (HB) tend to be inferior. However, for certain fingerprint types, the more generic atom typing performs well, such as G/HB with radial or HB with MOLPRINT2D. The best overall performance comes from dendritic,

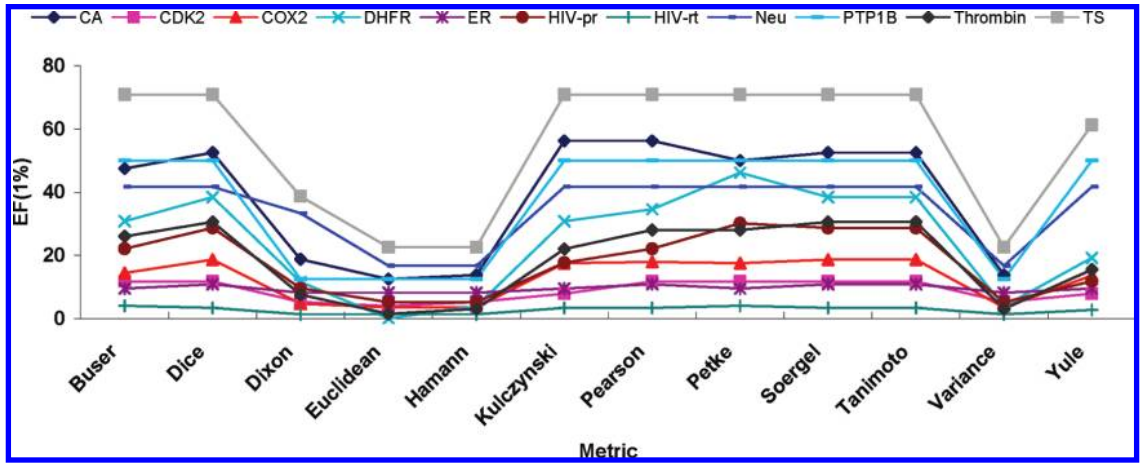


Figure 10. Enrichments for each of the 11 targets for the different metrics with linear fingerprints, Daylight atom types, and no feature scaling.

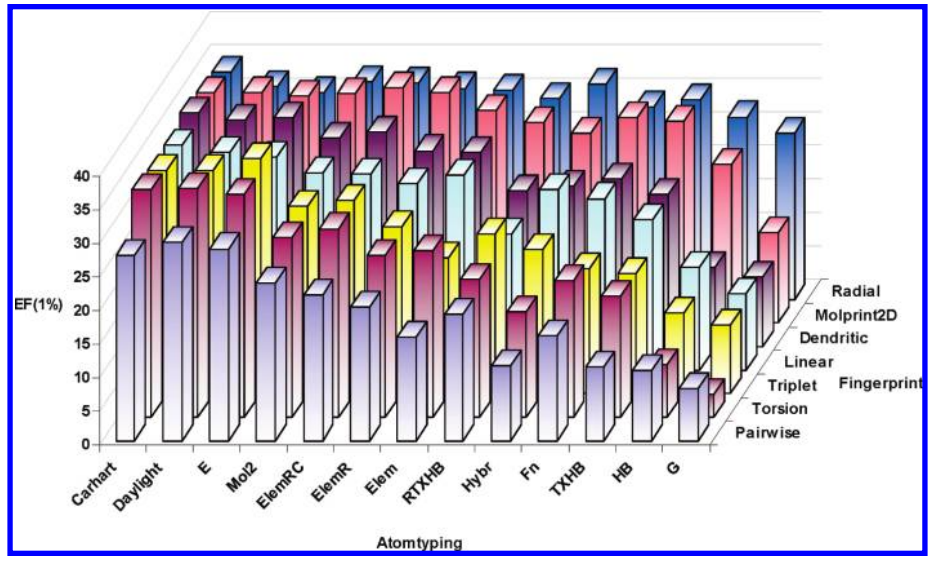


Figure 11. Average EF(1%) across all 11 targets of different atom-typing schemes for each fingerprint. MACCS is not present because there are no atom types. The fingerprint axis is sorted by ascending average results.

MOLPRINT2D, and radial fingerprints using Carhart, Daylight, Estate, or Mol2 atom types.

Figure 12 shows the variation in EF(1%) for different fingerprints, as the feature scaling is varied. While there is not a lot of variation for most fingerprints, radial fingerprints are a notable exception. As described in the Methods Section, radial fingerprints entail growing a set of fragments radially from each heavy atom over a series of iterations. These fingerprints are often very discriminating, since the addition of just one atom has a ripple effect that expands by one bond in each iteration, where the nascent pattern must match exactly to be counted. For a given feature size or a small increment in the size, there is a large number of items to be scaled and, therefore, a greater sensitivity.

CONCLUSION

In this work, we performed a large-scale two-dimensional (2D) virtual screening survey to explore the effects of various parameters on database enrichments. All combinations of fingerprint types, atom typing, bit scaling, and metrics were used to find trends that can help improve results in practical applications of 2D virtual screening. While it is not possible to draw absolute conclusions based on the number of targets studied (11), it is clear that improved parameter settings can be found for a given target compared to the standard settings found in the literature. Some methods, such as MOLPRINT2D, were found to be less sensitive to the precise settings. For example, the average EF(1%) across all parameter combinations was significantly higher for MOLPRINT2D than any other method, and the originally published MOLPRINT2D settings performed well, with an average EF(1%) enrichment of 33.6 across the 11 targets compared to 35.1 with the best MOLPRINT2D setting (ElemRC atom types, no bit scaling, and Buser similarity). We found that for some targets the variation in enrichment with changes in the parameters was substantial, such as carbonic anhydrase, where an enrichment of 55.0 with the original MOLPRINT2D settings was improved to 80.0 with the best settings (see Supporting Information Table S1 for the best settings for each target).

No single combination of fingerprint settings is optimal for all ligand sets, which is expected given the diversity of query molecules and associated active compounds. For example, in the case of HIV reverse transcriptase, which exhibits a great deal of structural diversity within the active set, it was found that the atom-typing scheme that describes the functional atom types and all bonds as equivalent (Fn)

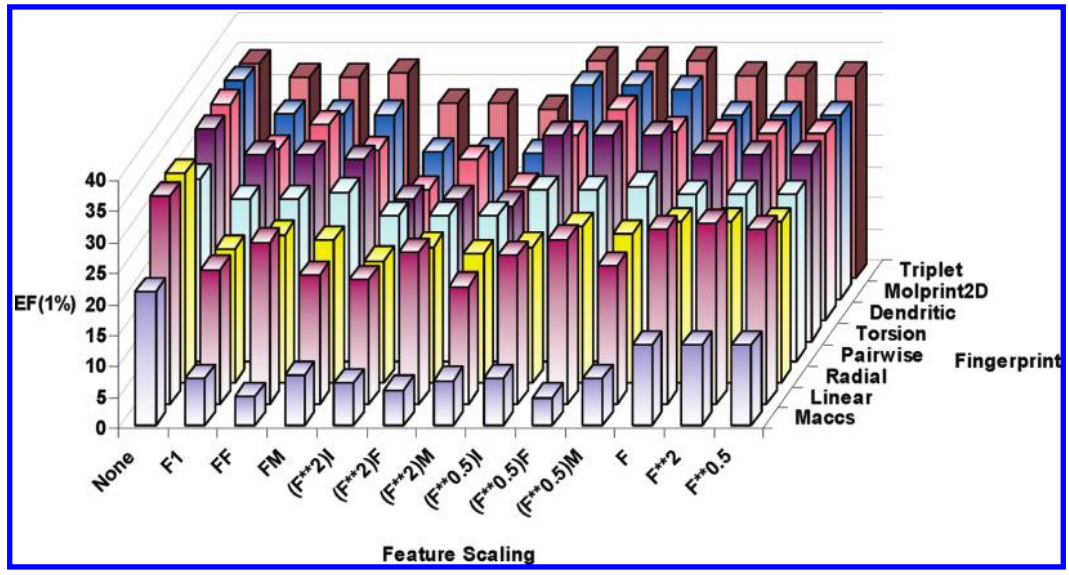


Figure 12. Average EF(1%) across all 11 targets of different feature scalings for each fingerprint type. The fingerprint axis is sorted by average ascending results.

performed very well. On the other hand, with more similar active compounds the most specific atom-typing schemes (Carhart, Daylight, Estate, and Mol2) performed best. However, if the objective of a screen is to identify novel, diverse hits, then a less specific atom-typing scheme may be more appropriate. The top-scoring inactive compounds for each target using three different atom-typing schemes (Daylight, ElemRC, and Fn) with MOLPRINT2D fingerprints, no bit scaling, and Buser similarity can be seen in the Supporting Information Figures S1−S3. These inactive hits give a sense of the ability to find novel compounds that are not known actives and, therefore, scaffold hopping possibilities. However, a detailed study of scaffold hopping is beyond the scope of this paper and will be covered in our future work.

Using a reduced bit-space for the fingerprints was shown to result in considerable degradation in enrichments, as a consequence of the many collisions, and therefore loss of information. The high number of collisions was particularly apparent in the case of staurosporine, which has many more features than the total number of addressable bits in a 10-bit fingerprint. In the case of linear fingerprints with Daylight atom typing, we found that the average EF(1%) over all targets increased from 20.7 to 32.5 when going from an addressable space of 2^{10} to 2^{32}. In the extreme case of thrombin, the EF(1%) increased from to 4.5 to 30.5. Furthermore, the results from the 32- and 64-bit fingerprints were identical, suggesting that 2^{32} addressable bits is large enough to ensure that collisions will be minimal. This is in agreement with the findings of Baldi et al.,[53] suggesting that enlarged bit-spaces should be more successful in practical applications of 2D virtual screening.

In this study we performed an unbiased assessment of fingerprint methods and associated parameters. However, in drug discovery projects there is almost always information that can help an investigator make informed decisions. Custom weighting of specific features is one way to encode a priori knowledge into fingerprints. For carbonic anhydrase, it would be relatively straightforward to weight the sulfonamide moiety more heavily than other features, which would likely result in improved enrichments. More subtle information can also be encoded in fingerprints, such as certain patterns of donors and acceptors at known relative distances that are needed to make hydrogen bonds, as in the kinase hinge interaction. The inclusion of custom fingerprint bit weightings will be the focus of a future paper.

This study resulted in over a terabyte of data, and it is not expected that such an exhaustive set of parameters could be explored routinely or on a broader set of targets. The hope is that the lessons learned here can be used to provide a reference for 2D fingerprint screens or a starting point in further optimizations. Based on the parameter combinations presented here, it should be possible to obtain improved database enrichments and to extend the studies to other variables and data sets.

ACKNOWLEDGMENT

The authors thank Ken Dyall for reviewing the manuscript and providing helpful comments.

Supporting Information Available: Highest ranking inactive compounds for each target using MOLPRINT2D fingerprints, Daylight, ElemRC, and Fn atom typing, no bit scaling, and Buser similarity. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening - An Overview. *Drug Discovery Today* **1998**, *3*, 160–178.
(2) *Virtual Screening for Bioactive Molecules*; Böhm, H.-J., Schneider, G., Eds.; Wiley-VCH: Wienheim, Germany and New York, 2000.
(3) *Virtual Screening in Drug Discovery*, Shoichet, N., Alvarez, J., Eds.; CRC Press: Boca Raton, FL, 2005.
(4) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
(5) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Bond. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
(6) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-

Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219–1229.
(7) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish between "Drug-like" and Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314–3324.
(8) Dixon, S. L.; Merz, K. M., Jr. One-Dimensional Molecular Representations and Similarity Calculations: Methodology and Validation. *J. Med. Chem.* **2001**, *44*, 3795–3809.
(9) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
(10) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
(11) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: A System to Select Quasi-Flexible Ligands Complementary to a Receptor of Known Three-Dimensional Structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 153–174.
(12) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.
(13) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian Docking Functions. *Biopolymers* **2003**, *68*, 76–90.
(14) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shaw, D. E.; Shelley, M.; Perry, J. K.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
(15) Gund, P., Three-Dimensional Pharmacophore Pattern Searching. In, *Progress in Molecular and Subcellular Biology*; Hahn, F. E., Ed.; Springer-Verlag: Berlin, Germany, 1977; Vol. 5, pp 117−143.
(16) Güner, O. F.; Henry, D. R.; Pearlman, R. S. Use of Flexible Queries for Searching Conformationally Flexible Molecules in Databases of Three-Dimensional Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 101–109.
(17) Clark, D. E.; Jones, G.; Willett, P. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Conformational Searching Algorithms for Flexible Searching. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 197–206.
(18) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297–1308.
(19) Meyer, A. Y.; Richards, W. G. Similarity of Molecular Shape. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 427–439.
(20) Putta, S.; Lemmen, C.; Beroza, P.; Greene, J. Novel Shape-Feature Based Approach to Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1230–1240.
(21) *ROCS*, version 3.0; OpenEye Scientific Software: Sante Fe, NM, 2009.
(22) Ballester, P. J.; Richards, W. G. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J. Comput. Chem.* **2007**, *28*, 1711–1723.
(23) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379–386.
(24) Dixon, S. L.; Koehler, R. T. The Hidden Component of Size in Two-Dimensional Fragment Descriptors: Side Effects on Sampling in Bioactive Libraries. *J. Med. Chem.* **1999**, *42*, 2887–2900.
(25) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Measure Based on Tree Matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.
(26) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574.
(27) ato, S. J., Exploring Pharmacophores with Chem-X. In, *Pharmacophore Perception, Development and Use in Drug Design*, Güner, O. F., Ed.; International University Line: La Jolla, CA, 2000; pp 110−125.
(28) Cramer, R. D.; Poss, M. A.; Hersmeier, M. A.; Caulfield, T. J.; Kowala, M. C.; Valentine, M. T. Prospective Identification of Biologically Active Structures by Topomer Shape Similarity Searching. *J. Med. Chem.* **1999**, *42*, 3919–3933.
(29) Grant, J. A.; Haigh, J. A.; Pickup, B. T.; Nicholls, A.; Sayle, R. A. Lingos, Finite State Machines, and Fast Similarity Searching. *J. Chem. Inf. Model.* **2006**, *46*, 1912–1918.
(30) *Canvas*, version 1.2; Schrödinger L.L.C.: New York, NY, 2009.
(31) Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407–1414.
(32) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.
(33) Glen, R. C.; Adams, S. E. Similarity Metrics and Descriptor Spaces - Which Combinations to Choose. *QSAR Comb. Sci.* **2006**, *25*, 1133–1142.
(34) Willet, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
(35) *MDL Drug Data Report*; MDL Information Systems/Symyx: Santa Clara, CA, 2005.
(36) *MACCS-II*; MDL Information Systems/Symyx: Santa Clara, CA, 1984.
(37) *Daylight Fingerprint Toolkit*, version 4.9; Daylight Chemical Systems, Inc.: Aliso Viejo, CA, 2008.
(38) *Unity*, version 4.4; Tripos L.P.: St. Louis, MO, 2003.
(39) *GenerateMD*, version 5.3.2; ChemAxon: Budapest, Hungary, 2010.
(40) *Daylight Theory Manual*; Daylight Chemical Systems, Inc.: Aliso Viejo, CA, 2008.
(41) Lajiness, M. S. Dissimilarity-based Compound Selection Techniques. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 65–84.
(42) Rogers, D.; Brown, R. D.; Hahn, M. Using Extended-Connectivity Fingerprints with Laplacian-Modified Bayesian Analysis in High-Throughput Screening Follow-up. *J. Biomol. Screening* **2005**, *10*, 682–686.
(43) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structure - A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.
(44) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definitions and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 65–73.
(45) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsions: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.
(46) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.
(47) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
(48) *SMARTS - Language for Describing Molecular Patterns*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 2008.
(49) *Sybyl*, version 8.1.1; Tripos L.P.: St. Louis, MO, 2009.
(50) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
(51) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*, Academic Press: San Diego, CA, 1999.
(52) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Comput. Sci.* **2007**, *47*, 448–508.
(53) Baldi, P.; Benz, R. W.; Hirschberg, D. S.; Swamidass, S. J. Lossless Compression of Chemical Fingerprints Using Integer Entropy Codes Improves Storage and Retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 2098–2109.

CI100062N