Molecular Complexity Effects and Fingerprint-Based Similarity Search Strategies

Dissertation zur
Erlangung des Doktorgrades (Dr. rer. nat.) der
Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Yuan Wang
aus Peking

Bonn
2009

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen
Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath

2. Referent: Univ.-Prof. Dr. rer. nat. Andreas Weber

Tag der Promotion: 05 November 2009

Erscheinungsjahr 2009

謹以此致我的祖父

Abstract

Molecular fingerprints are bit string representations of molecular structure and properties. They are among the most popular descriptors and tools in molecular similarity searching because of their conceptual simplicity and computational efficiency. In order to calculate molecular similarity, fingerprints are computed for reference and screening database compounds and their bit settings are quantitatively compared using similarity metrics. One caveat of this approach is the bias caused by complexity effects: complex molecules have higher fingerprint bit density and produce artificially high similarity values.

The asymmetric behavior of Tversky similarity measurement has been reported: comparing A to B is not equal to comparing B to A. This phenomenon can be directly attributed to complexity effects. Hence, preference of parametric settings for Tversky coefficient is determined with regard to the relative difference of molecular complexity. One approach to avoid such effects is using fingerprint representations having constant bit density. Alternatively, emphasizing the absence of bit position features, which is not recorded using conventional fingerprint similarity search methods, provides another approach to address complexity effects. However, in order to optimize search performance, elimination of complexity effects using this approach is not as effective as modulation of complexity effects. In order to evaluate the outcome of virtual screening, search performance is monitored for combinations of different parameters. In general, in similarity searching using highly complex reference compounds it is difficult to recover potential hits that are less complex.

To further investigate complexity effects, the random reduction of fingerprint bit density is also explored. The ensuing loss of chemical information can be compensated for by balancing complexity effects when the fingerprints of reference compounds are modified to reduce their bit density.

When this random process is replaced with iterative bit silencing, the significance of each bit position in similarity searching can be analyzed and different weights can be assigned to each position. Such a weighting scheme emphasizes critical bit positions specific to the reference activity class. Class-specific similarity metrics can be derived by utilizing these weights in similarity calculation. Using these similarity metrics similarity search performance can be improved, especially when conventional methods fail to retrieve potential active compounds.

Information of reference sets can also be directly utilized in the form of

Shannon entropy as a measure of similarity. This simple and efficient similarity search strategy assesses the fingerprint entropy penalty induced by introducing external molecules into the reference set. It has comparable or better performance compared to nearest neighbor approaches but lower computational costs.

Acknowledgments

I would like to thank my supervisor, Prof. Dr. Jürgen Bajorath, for his guidance throughout my study. I also would like to thank Prof. Dr. Andreas Weber for being the co-referent. Thank Dr. Hanna Geppert for her help and advice, and all my colleagues from B-IT for their encouragements and a pleasant working atmosphere. Finally, thanks to my family and my friends for their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the recent decade various computational techniques have become important tools widely used in modern drug discovery.[1–3] *In silico* approaches such as *virtual screening* have become popular in handling increasingly large databases because of their high efficiency and low cost.

Virtual screening (VS) is defined as the computational analog of biological screening, which aims to score, rank, and/or filter a set of compounds using one or more computational procedures.[2] It originates from mainly two areas: protein structure-based compound screening or docking,[4,5] and chemical similarity searching based on small molecules.[1,6] Despite the increasing availability of target protein structures as VS templates, small molecules such as biological screening hit or lead compounds are still the dominant source of information and thus commonly utilized.[1]

Small molecules can be represented using *molecular descriptors*, which are defined as mathematical models of molecular structures and properties.[2] They represent and describe the physicochemical or structural features of molecules, vary in the procedure of computation, complexity of the encoded information, and also the computational complexity. One of the most popular descriptor types for similarity searching and chemical database mining is the simple but effective *molecular fingerprint*.[1,6–8]

1.1 Molecular fingerprints

Molecular fingerprints are bit string representations of molecular structure and properties. Structural and/or physico-chemical property information of a molecule is usually encoded as a binary string where each bit detects the presence or absence of a specific chemical feature or represents a value range of a property descriptor.[1,8] Alternatively, such binary indicators can be replaced with frequency counts of these features and then the molecules are represented as integer strings, also known as molecular holograms.[9–13] In *similarity*

searching, compounds with known biological activity are utilized as reference compounds and their fingerprint representations are calculated. Fingerprints of database molecules are compared with reference fingerprints in a pair-wise manner in order to identify novel active compounds.[6] Hence, this type of similarity searching is carried out in fingerprint space and the overlap between bit string representations is used as a measure of molecular similarity.

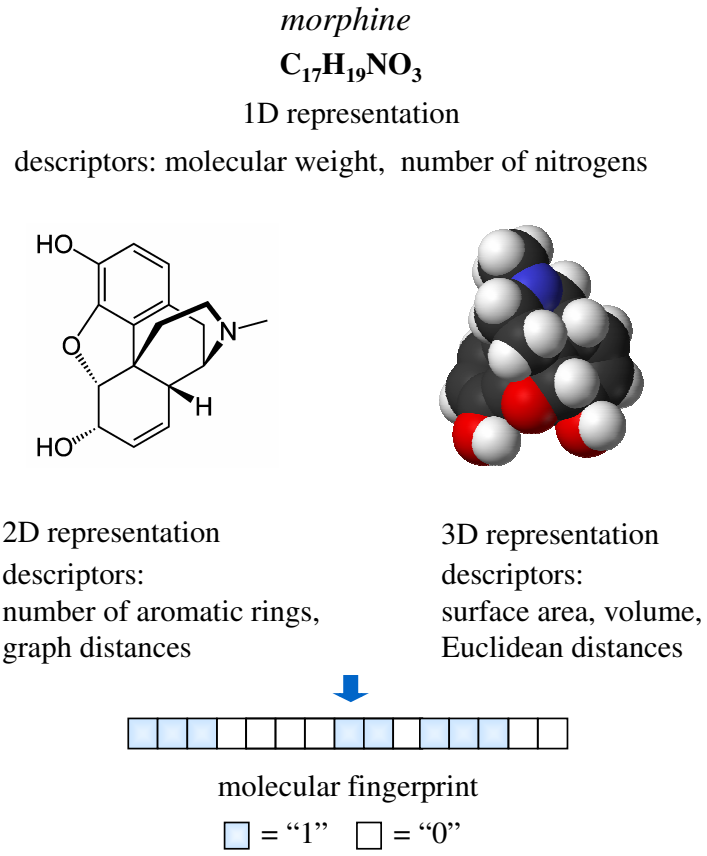


Figure 1.1: Molecular representations and fingerprints. Examples of molecular descriptors and fingerprint are shown for morphine. Molecular representations of different dimensionality (1D, 2D or 3D) produce different descriptors.

Fingerprints are often distinguished based on the dimensionality of the molecular representations from which they are calculated.[9,10] Two-dimensional fingerprints are derived from the chemical graph representation of a molecule and take into consideration information extracted from atom and bond types and graph distances, whereas the calculation of 3D fingerprints requires conformational information, i.e., atomic coordinates.[9] In pioneering investigations, Brown and Martin compared various 2D and 3D descriptors in molecular similarity analysis and concluded that 3D representations were not generally superior to 2D fingerprints,[12,13] although they should in principle contain more

relevant information, simply because molecules are active in three dimensions. The 2D versus 3D descriptor and search method debate is continuing to this date in the literature, but the early views of Brown and Martin have not been fundamentally revised. Two-dimensional molecular representations and search methods are often equally or more successful than 3D methods because they are generally more robust and less error-prone.[1] In particular, 2D fingerprints have been surprisingly successful in many applications, despite their conceptual simplicity.[14,15]

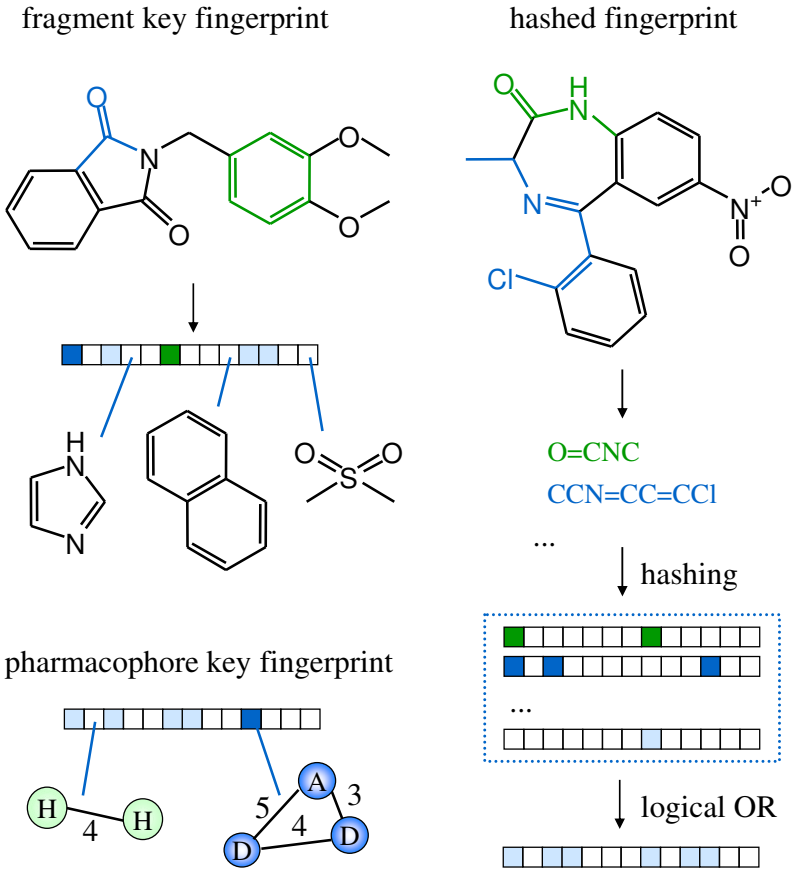


Figure 1.2: Key-type and hashed fingerprints. Fragment key fingerprints, pharmacophore key fingerprints and hashed fingerprints are shown. Fragments, pharmacophore features or paths highlighted in blue or green are projected to the hypothetical fingerprint bit positions filled with the corresponding color. In pharmacophore-based fingerprints, "H", "A" and "D" in colored circles represent hydrophobic group, hydrogen acceptor and hydrogen donor, respectively.

Two-dimensional fingerprints can be classified by considering how their bit strings encode chemical information. In key-type fingerprints such as the MACCS keys,[16] each bit corresponds to a structural feature.[17] The BCI fin-

gerprint is also keyed and for its generation a dictionary of possible fragments is constructed.[18] In pharmacophore-type 2D fingerprints such as TGD (Typed Graph Distance) and TGT (Typed Graph Triangle),[19] atom types and binned 2D graph distances are combined as pharmacophore patterns and each bit represents a possible 2D pharmacophore arrangement. In contrast, hashed fingerprints represent a different design. For example, the pioneering Daylight fingerprint enumerates unique paths up to a specified maximum length in the molecular graph and maps these connectivity pathways onto a bit string of fixed length using a hash function.[20] Following another design strategy, Extended Connectivity Fingerprints (ECFP) generate variable numbers of layered circular atom environments in a molecule-specific manner and hash them into integer representations.[21] In order to compare and group fingerprint representations, Bender et al. have recently conducted a systematic principal component analysis of similarity value distributions of test compounds calculated with various fingerprints, which revealed correlations between different types of fingerprint descriptors.[22] There are in general four broad classes of fingerprints: binary circular fingerprints, circular fingerprints considering counts, path-based and keyed fingerprints, and pharmacophore-based fingerprints. Representative examples of 2D molecular fingerprints and their composition are reported in Table 1.1.

fingerprint	**designation**	**descriptor**	**encoding**	**length**
MACCS[16]	Molecular ACCess System	structural fragments	one-to-one correspondence of bit positions and fragment keys	fixed, 166 bits
TGD / TGT[19]	Typed Graph Distance / Typed Graph Triangle	2D pharmacophore features with atom types and distances	one-to-one correspondence of bit positions and pharmacophore keys	fixed, 420 / 1704 bits
BCI[18]	-	structural fragments	one-to-one correspondence of bit positions and fragment keys from constructed dictionary	dependent on dictionary
Daylight[20]	-	paths or subgraphs	hash function mapping to fixed length	user-defined, e.g. 1024 or 2048 bits
ECFP[21]	Extended Connectivity FingerPrint	extended graph connectivity	hash function mapping to virtual feature space	infinite

Table 1.1: Exemplary 2D fingerprint designs. For each fingerprint the designation of abbreviation, descriptor origin, encoding method and length are reported.

1.2 Similarity metrics

Fingerprint overlap as a measure of molecular similarity is quantitatively determined using various *similarity metrics*. One of the most popular similarity metrics is the Tanimoto coefficient (Tc).[6] The binary form of the Tanimoto coefficient is defined as

$$Tc = \frac{c}{a + b - c}$$

with a being the number of bits set on in the first fingerprint, b the number of bits set on in the second fingerprint, and c the number of bits common to both. Other similarity coefficients have also been applied in the calculation of pair-wise fingerprint similarity, either separately or in combination using data fusion techniques.[6,23–26] Going beyond Tc-like metrics, the Tversky coefficient (Tv)[27] makes it possible to weight the contributions of bit settings of reference and database molecules by introducing the weight parameter α:

$$Tv = \frac{c}{\alpha(a - c) + (1 - \alpha)(b - c) + c}$$

Although many different similarity metrics and coefficients have been reported, systematic comparisons have not revealed a general preference of one method over others.[6,7,24,25] Tanimoto similarity is predominantly calculated to this date because of its simple formulation and stable results over various data sets.[28,29] However, as will be discussed in the following sections, the Tversky formalism offers an opportunity to systematically modify similarity evaluation and study the effects of differential weights on bit settings of reference and database compounds and bits that are set on or off. Table 1.2 reports several similarity metrics that are applied in fingerprint similarity calculation.

It is difficult to establish molecular similarity threshold values that correlate with biological activity. However, this question is particularly relevant for similarity searching because one generally aims at identifying different structures with similar activity, which essentially applies to all virtual screening methods.[1] In a database search, compounds with highest fingerprint similarity are often close analogs of reference compounds and are typically not the molecules one is interested in. Rather, one is mostly interested in structurally increasingly diverse compounds that are typically "further down the list", and this explains why the exploration of activity-relevant similarity threshold values is of high interest.

A traditional way of addressing the question of how calculated similarity is related to activity is provided by cluster analysis.[30,31] For example, molecules can be clustered based on 2D fingerprint similarity and the composition of the computed clusters and the resulting distribution of active and inactive compounds are analyzed. Other studies have been carried out using

coefficient	formula
Jaccard / Tanimoto[6]	$\frac{c}{a+b-c}$
Tversky[27]	$\frac{c}{\alpha(a-c)+(1-\alpha)(b-c)+c}$
Russell / Rao[25]	$\frac{c}{N}$
simple match[25]	$\frac{c+d}{N}$
Forbes[25]	$\frac{Nc}{ab}$
Dice[6]	$\frac{2c}{a+b}$

Table 1.2: Popular similarity metrics. Reported are five similarity coefficients commonly used in fingerprint overlap calculations. a is the number of "1" bits in reference compound, b the number of "1" bits in database molecule, c the number of "1" bits common to both, d the number of "0" bits common to both, and N is the length of the fingerprint. α is the weight on "1" bits in reference compound.

high-throughput screening data sets to analyze the relationship between active and inactive compounds in light of their calculated similarity values.[32,33]

In their seminal publication establishing neighborhood behavior, Patterson et al. showed that for their Unity fingerprints, a Tc value of at least 0.85 corresponded to a high probability that two test compounds shared the same activity.[34] This value has been adopted in many studies to search for bioactive molecules. However, for fingerprints and search conditions other than the originally applied ones, this value was often found to be only a weak indicator of true similarity-activity relationships.[31]

These studies have illustrated that generally applicable similarity threshold values are not available as bioactivity markers. Similarity threshold values can not be generalized because different fingerprints and compound classes require a case-by-case determination of activity-relevant similarity levels.[8]

1.3 Complexity effects

Molecular complexity or size effects are known to bias fingerprint-based similarity evaluation and negatively affect search performance.[10,25,26,35] In a milestone publication, Flower demonstrated that reference compounds of increasing size

generate systematically higher Tc values in databases searching.[10] This is the case because fingerprint *bit density*, defined as the number of "1" bits divided by the length of the fingerprint, typically increases with molecular complexity. High bit density generally favors statistical chance matches in fingerprint comparison and hence might artificially increase similarity values.

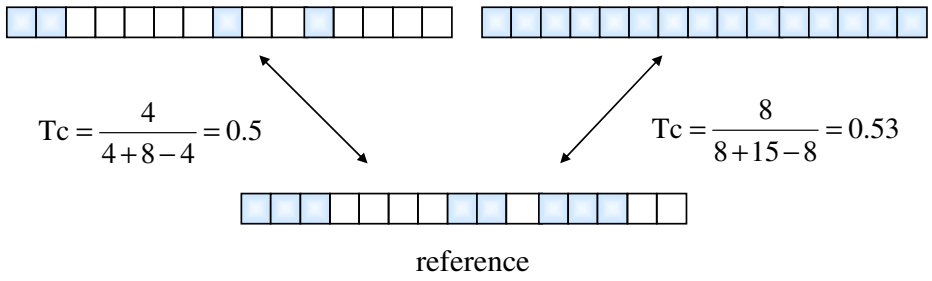


Figure 1.3: Complexity effects in fingerprint similarity calculation. Two candidate fingerprints, one having less "1" bits and the other more, are compared to the same reference fingerprint using Tc similarity metrics. The one having higher "1" bit density (upper-right) yields also higher similarity value, regardless of its actual similarity to the reference. In all fingerprints "1" bits are colored in blue and "0" bits in white.

Molecular size is often, but not always, related to fingerprint bit density. Exceptions include, for example, polymers where fragment-based fingerprints would only account for the presence of a monomer, but not the occurrence of multiple copies. Furthermore, bit density is also influenced by chemical complexity of molecules. When discussing aspects of molecular complexity in the context of similarity evaluation, it should also be considered that alternative molecular representations (for example, 2D versus 3D representations) mirror complexity in different ways. Molecular complexity is determined by multiple components. Depending on the chosen molecular representations, not all factors that contribute to complexity might be taken into account. Table 1.3 provides examples of complexity-relevant factors that can be accounted for at the level of 2D representations and others that require the use of 3D representations. However, regardless of which factors are ultimately considered, when using (2D or 3D) fingerprints, differences in molecular complexity and size typically lead to intrinsically different bit densities.

Figure 1.4 illustrates the principal influence of molecular complexity on fingerprint search calculations on the basis of MACCS Tc distributions. The larger and more complex test compounds are, the higher their bit densities and similarity values in general become. Thus, using reference compounds of moderate to high complexity generally favors the recognition of large and complex database molecules, regardless of whether these molecules are active or not.

2D factors	3D factors
element distribution	conformational entropy
H-bond acceptors/donors	electrostatic potentials
hybridization states	interatomic distance distribution
rigidity	intramolecular interactions
bond topology	stereochemistry

Table 1.3: Factors related to molecular complexity. Examples of factors are listed that contribute to molecular complexity together with the dimensionality of the molecular representation that is required to capture or deduce them.

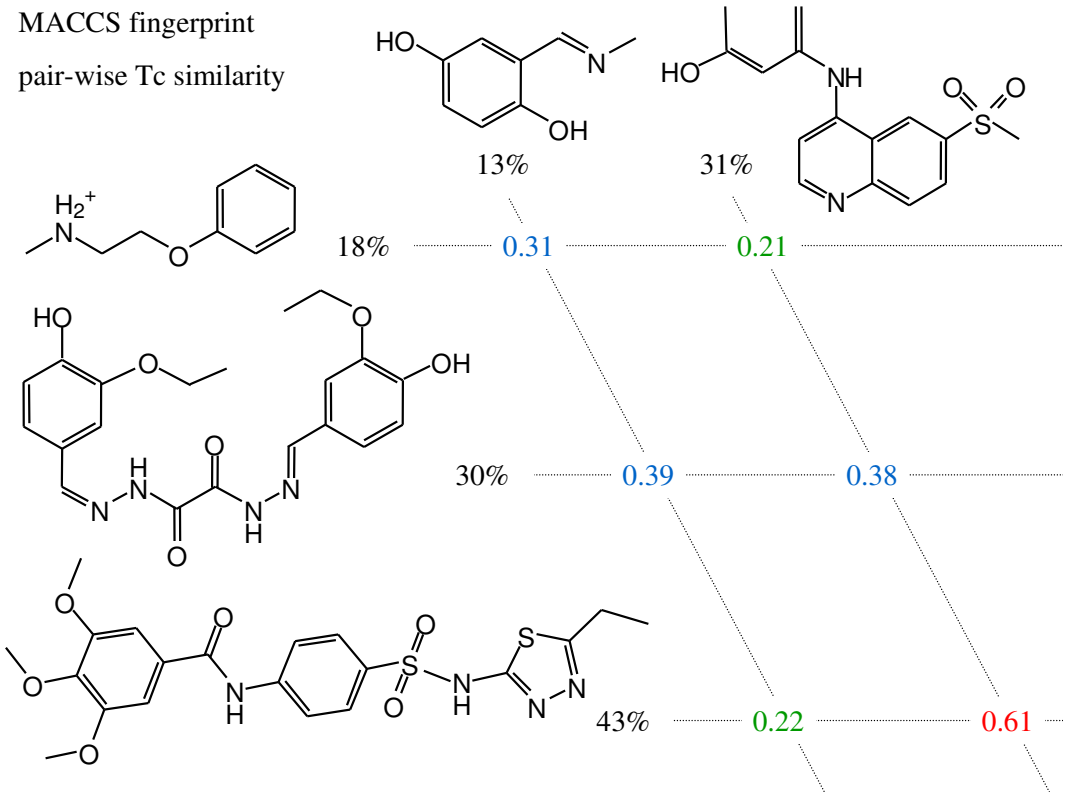


Figure 1.4: Molecular complexity and similarity. Three database molecules (in the left panel) having increasing complexity levels are compared to two reference compounds (depicted in the top) with different complexity using Tanimoto coefficient. Due to the definition of MACCS structural key fingerprint,[16] molecules with higher levels of structural complexity produce MACCS fingerprints with higher "1" bit densities and consequently higher pair-wise Tc similarity values. The bit densities are shown next to the corresponding molecules in percentage and the pair-wise Tc similarities are reported in different colors. Low Tc values are color-coded green, medium values blue, and high values red.

The complexity effects also affect the quality of benchmarking calculations, which are used to evaluate similarity search performance. In a typical benchmarking calculation, a number of known active compounds are added to the background database as targets for the similarity search method under investigation. However, these "hidden" actives, and also the reference compounds utilized to search for them, are usually optimized compounds taken from literature or patent sources that are often more complex than average database molecules. As a result, these complex compounds are easily recognized by similarity searching because of their high similarity values. Thus, the search performance of fingerprints is often artificially high in such benchmark situations and does not accurately reflect a "real life" search scenario. In practical applications, newly identified hits are less complex than optimized lead compounds and hence more difficult to detect.

1.4 Outline of this thesis

This study addresses three major questions:

1. How do complexity effects influence similarity searching?
2. How do they affect virtual screening applications?
3. Can novel computational methods be developed to avoid complexity effects and improve similarity search performance?

In *Chapter 2* fingerprint-based similarity search strategies are introduced together with a general workflow for benchmarking calculations. Concepts and schemes that have been adopted in this thesis are presented. In addition, recent advances in the area of similarity searching using fingerprint-based methods are reported.

In *Chapter 3* the asymmetric behavior of the Tversky coefficient is assessed: given two molecular fingerprints, A and B, comparing A to B might yield different Tversky similarity values than comparing B to A. This phenomenon is shown to be directly related to complexity effects. Also discussed in this chapter is the complexity-independency of a previously developed molecular fingerprint, which can be adopted to avoid biased similarity calculation that is caused by molecular complexity. Then a novel similarity metric, *weighted Tversky coefficient (wTv)*, is introduced as a tool to balance complexity effects. wTv can either eliminate or modulate complexity effects. Calculations reported in this chapter show that modulating complexity effects can improve the search performance more than completely eliminating them.

In *Chapter 4* another novel similarity search method is introduced to address complexity effects from a different angle. This technique, called *random fingerprint bit silencing*, can be applied to highly complex reference compounds

used as templates to search against databases containing less complex structures. Its enhanced performance in systematic test calculations is demonstrated in this chapter.

In *Chapter 5* the *bit position weighted Tanimoto coefficient (bwTc)* is introduced. The bit silencing technique described in *Chapter 4* is employed to derive this novel class-specific similarity metric. Benchmarking test results compared to conventional search methods are presented. The incorporation of class-specific information has been found to significantly improve the results. By combining this metric with the wTv coefficient described in *Chapter 3*, a class-specific similarity metric modulating complexity effects is introduced, the *weighted Tversky coefficient with class-specific bit weighting*, or *wbwTv*. Systematic search calculations revealed better performance of wbwTv compared to its parental methods and other fingerprint-based similarity search strategies.

In *Chapter 6* the Shannon entropy concept is adopted for evaluating bit settings in sets of fingerprints. Its application in similarity searching provides an unconventional yet efficient strategy for molecular similarity calculations.

Chapter 2

Methods in Fingerprint-Based Similarity Searching

Similarity search calculations are conceptually based on the *similarity property principle*: similar molecules are thought to have similar biological activity.[36] That is the case because the interaction of a small molecule and a target protein is dependent on their structures. Small molecules with similar structures are expected to interact similarly with the target. According to this principle, the molecular similarity of screening database molecules to a set of known active reference compounds or an individual reference compound is assessed in similarity searching.[6,37] In order to calculate molecular similarity, fingerprints are computed for reference and screening database compounds and their bit settings are quantitatively compared[15,37] using similarity functions or metrics such as the popular Tanimoto coefficient (Tc).[6]

In this chapter, benchmarking calculations used to evaluate the performance of different computational methods are introduced. This methodology is applied in most of the calculations in this thesis, with minor variations for different approaches. Furthermore, recent discoveries and developments of fingerprint-based search techniques are revisited, including data fusion, frequency-based operations, analysis of complexity effects, and novel fingerprint design strategies.

2.1 Benchmarking of similarity searching

In the benchmarking, compounds that are confirmed to be active are used as templates. A typical source for these compounds is annotated molecular databases containing ligands with confirmed activity. For example, the Molecular Drug Data Report (MDDR)[38] contains structure and activity information of over 150,000 biologically relevant compounds and derivatives[38] and is usually used here as a source of *activity classes* (i.e., sets of compounds that are active

against the same target). In addition to the templates, a number of confirmed active compounds are "hidden" in the background database to be recovered by the search process. They are referred to as the *active database compounds (ADC)* and are extracted from the same activity class of the reference/template compounds.

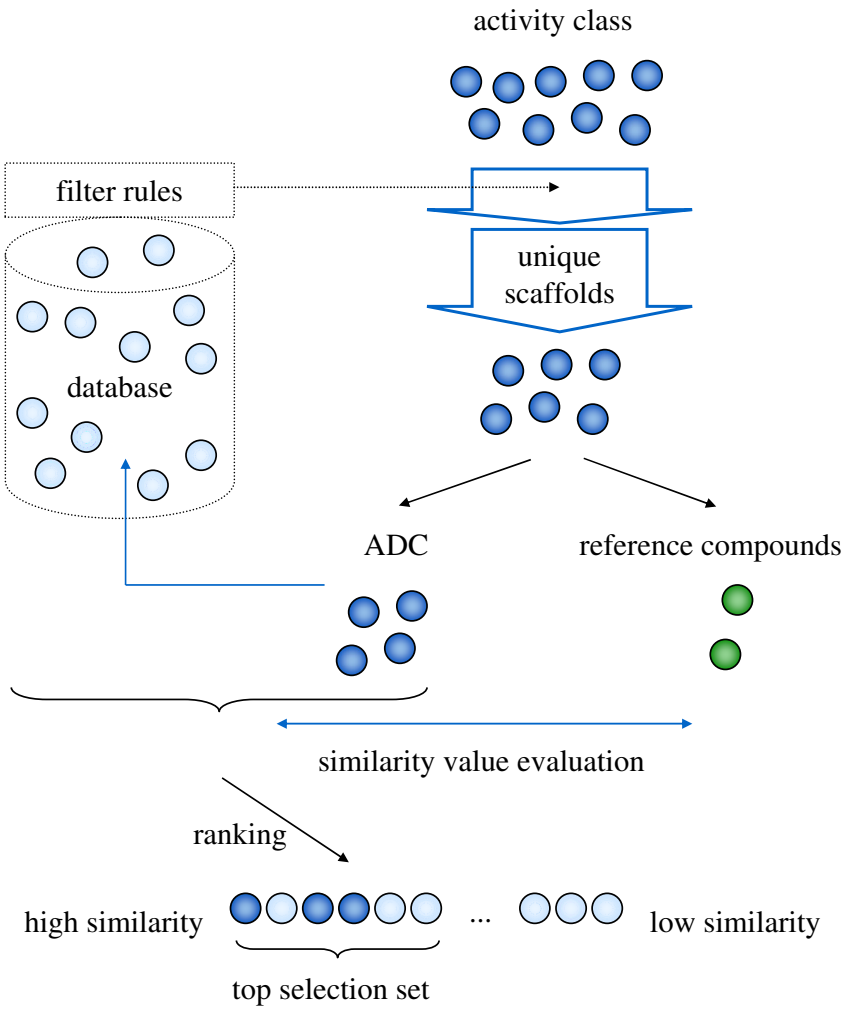


Figure 2.1: General calculation protocol. This flowchart illustrates the setup of the benchmarking system: filtering of activity class, dividing it to reference set and ADC, and carrying out similarity searching and ranking.

To ensure that pre-selected active compounds have molecular properties comparable to background database molecules, they are pre-filtered. For example, the ZINC database that currently contains over eight million small molecules is a public-domain database of compounds that are commercially

available.[39] In a drug-like subset of ZINC, all compounds are required to have a molecular weight of less than 600 Da, a logP value (the logarithm of octanol-water partition coefficient) in the range [-2, 6], between 1 - 10 hydrogen bond donors and 1 - 10 acceptors, and less than 19 rotatable bonds.[39] Similar rules apply to the NCI anti-AIDS database[40], which contains screening results for 42,687 compounds against HIV-related targets.[40] Before similarity searching, active compounds are filtered according to these rules. Furthermore, each pre-selected active compound must have a unique core structure[41] in order to avoid the inclusion of analog series that could potentially bias similarity search results.

Next, the fingerprint of each database molecule is compared to the fingerprints of reference compounds using similarity metrics. As described in section 1.2, determination of an exact activity-relevant similarity threshold is difficult. However, database molecules with the highest similarity values relative to reference compound(s), i.e., the *top-scoring* database molecules, are assumed to have a high probability to be active.

To evaluate the performance of a similarity search strategy, a number of top-scoring compounds are selected, e.g. 100 top-ranking compounds. Such selected compounds are called the *database selection set*, and the number of ADC that occur in this set is assessed. Two quantitative measures are the *hit rate (HR)* and the *recovery rate (RR)*.

Given the total number of ADC (M), the size of the selection set (S), and the number of ADC in the selection set (i.e., the number of "hits", K),

$$HR = \frac{K}{S}$$

and

$$RR = \frac{K}{M}$$

In Figure 2.1, the workflow of the benchmarking protocol is illustrated.

2.2 Merging information of multiple reference compounds

Similarity searching is applicable when only single reference compounds are available, in contrast to other data mining approaches such as cluster analysis or machine learning methods that require multiple active compounds.[1,37] However, fingerprint searching usually becomes more effective when multiple reference compounds (and hence more chemical information) are available.[7,37] For fingerprint searching using multiple reference compounds, different methods have been introduced.[7,14,15,37,42]

For example, fingerprint averaging – also known as the *centroid* method – can be applied to compare a database molecule to a reference set.[11] The

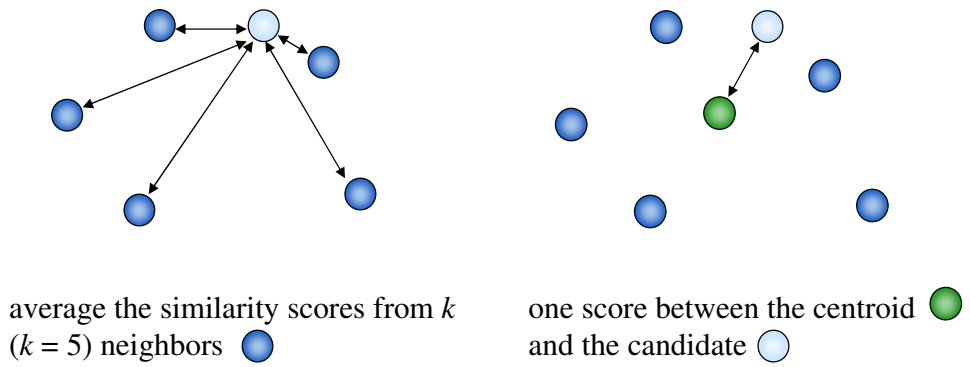


Figure 2.2: Data fusion approaches with multiple reference compounds. Two techniques, k-NN and centroid, are illustrated. k-NN requires k pair-wise similarity calculations (represented as bidirectional arrows to the dark blue circles representing the nearest neighbors) to determine the final average score of the candidate database molecule (blue circle), whereas centroid approach requires only one similairy calculation with the average vector (green circle).

centroid approach calculates an average vector from the fingerprints of the reference compounds. The average fingerprint is thought to represent the property center of the reference set and is compared to fingerprints of individual database molecules – often applying the general form of the Tanimoto coefficient[6]

$$Tc(\boldsymbol{A}, \boldsymbol{B}) = \frac{\sum_{i=1}^{N} a_i b_i}{\sum_{i=1}^{N} (a_i^2 + b_i^2 - a_i b_i)}$$

where $\boldsymbol{A} = (a_1, a_2, \ldots, a_N)$ and $\boldsymbol{B} = (b_1, b_2, \ldots, b_N)$ are two molecular fingerprint vectors of length N. They are not necessarily binary, as a result of the averaging process.

By contrast, data fusion of multiple Tc values relies on pair-wise comparison of a database molecule with all reference compounds and averages the k highest values to produce a final similarity score (*nearest neighbor technique*, or k-NN). For $k = 1$, the average rule becomes the maximum rule and the highest similarity value calculated against individual reference compounds is taken as the final compound score.[11] In comparative studies, 1-NN calculations often produce highest compound recall rates among data fusion techniques and other fingerprint search strategies.[42,43]

2.3 Frequency-based bit-wise techniques

From multiple reference compounds, statistics related to the occurence of bit positions can also be derived to develop methods yielding higher recall. Following the Stigmata approach,[44] fingerprint bit positions that are shared by

a subset of reference compounds of pre-defined size (e.g., at least 50%, 75% or 100% of the reference compounds) are set on as consensus features in a so-called *modal fingerprint* that is then used for database searching. Consensus bit positions have also been explored by fingerprint scaling, which weights different fingerprint bit positions according to their frequency of occurrence in the reference set during similarity searching.[45–47] Conserved bit positions are assigned high *scaling factors*, partly conserved positions are less emphasized, and non-conserved bit positions are not scaled, thus providing a linear compound class-specific weighting scheme.[46]

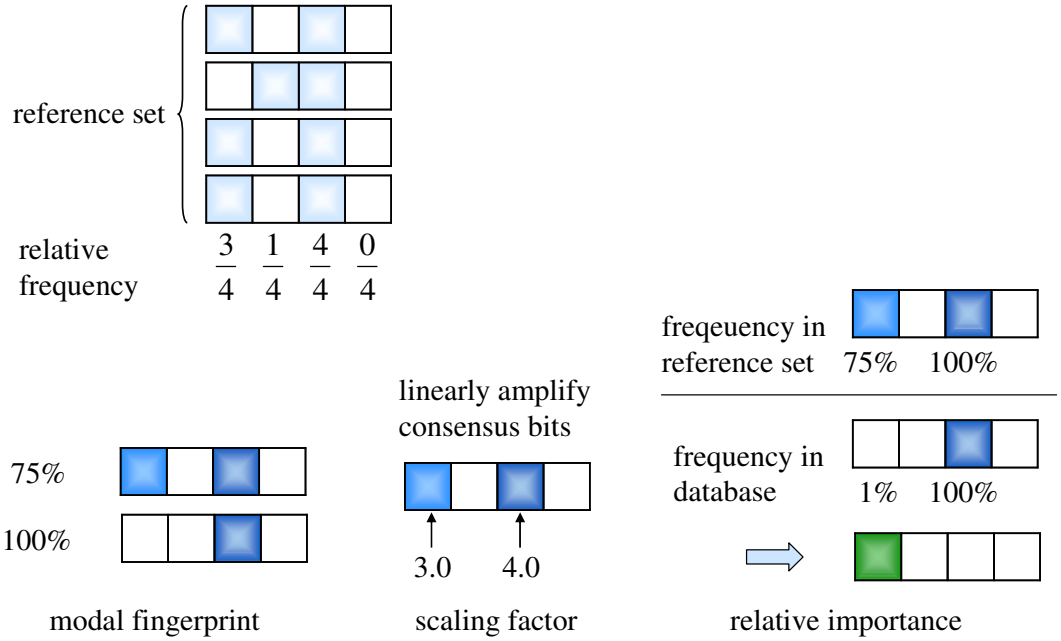


Figure 2.3: Frequency-based approaches. Three bit-wise techniques based on the relative frequency of bit positions are illustrated. Given a hypothetical reference set consisting of four molecular fingerprints, the generation of 75% and 100% modal fingerprints, the application of scaling factor based on the bits' relative frequencies, and the determination of relative bit importance are shown (with high-importance bit highlighted in green).

It should be noted that the derivation of modal or scaled fingerprints exclusively focuses on bit positions that are set on (i.e., set to "1"), but does not consider the absence of features. Nor do they include the occurences of features in the background database. Williams went a step further and introduced the concept of *relative bit importance* by taking not only the frequency of each bit position within the reference set into account but also the relative bit frequency in background database molecules,[9] giving rise to the so-called *reverse fingerprinting* approach that scores bit patterns in reference compounds that

are most discriminatory for active versus database compounds.[9] In Figure 2.3 the three frequency-based similarity search techniques are illustrated.

Feature distributions can also be taken into account in developing a search strategy for extended connectivity fingerprints (ECFPs)[21] that generates sets of layered circular atom environments (i.e., topological features) of varying size in a molecule-specific manner. Thus, these feature ensemble fingerprints depart from the classical fixed-format design of keyed fingerprints. For ECFPs, Hu et al.[48] have introduced the *feature filtering* method that removes features from search calculations that only occur in active, but not in database compounds. Thus, the search is focused on topological features occurring in reference sets. In the context of feature filtering, a simple similarity function that essentially counts reference set features present in database molecules and ranks them accordingly has been shown to be more effective than Tanimoto similarity calculations with increased structural diversity of hits.[48]

2.4 Molecular complexity effects in similarity searching

The influence of fingerprint complexity effects on search calculations has been explored in different ways. For example, in library design, Dixon and Koehler discovered a systematic relationship between molecular size and similarity in Tc calculations: sets of small molecules displayed a general tendency to be more dissimilar than large molecules.[35] Three distance metrics were applied to quantify compound dissimilarity: 1-Tc – the complement of Tanimoto similarity (a measure of distance or dissimilarity), XOR – exclusive OR (accounting for the number of bit positions that differ in fingerprints of two molecules), and the Euclidean distance. Within the same library, 1-Tc calculations preferentially selected subsets of small compounds as being dissimilar, whereas the other two metrics mostly selected subsets of larger compounds.[35] This phenomenon can be explained by the fact that complex compounds generally have more bit positions set to "1" than an average database molecule and thus have an increased probability to match "1" bits in other molecules.[10,26] To study such effects, Flower generated a probability density function for random bit string matching to investigate the theoretical distribution of Tc value ranges.[10] Furthermore, Holliday and colleagues analyzed the relationship between similarity values and relative bit density and found that comparison of low-density fingerprints generally produces lower Tc values than comparison of high-density fingerprints.[26,35]

For reference compounds of increasing complexity, Tc value distributions of database molecules systematically shift towards higher values,[10] as illustrated in Figure 2.4. In this context, molecular complexity essentially refers to topo-

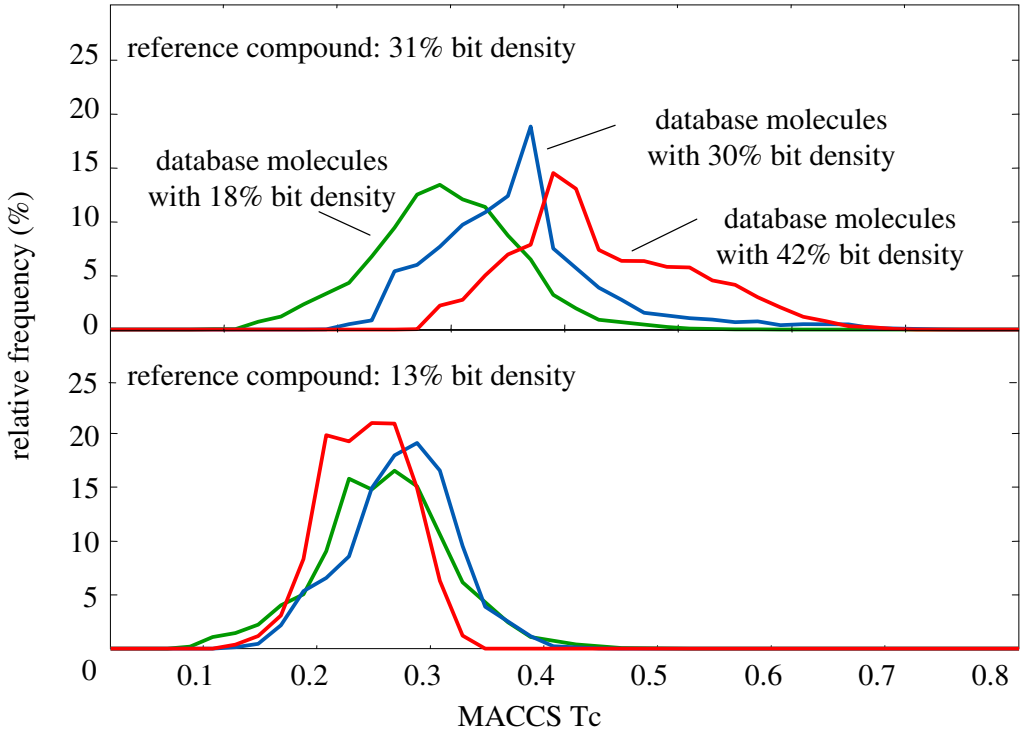


Figure 2.4: Tc similarity value distribution under complexity effects. Shown are the distributions of MACCS Tc similarity values produced by single template similarity searches on three different ZINC[39] subsets containing molecules of increasing bit density (18%, 30%, and 42%). When a reference compound with 31% bit density is used, the higher the bit density of database molecules becomes, the more the distributions are shifted towards higher Tc values. Thus, ZINC molecules with 42% bit density would preferentially be selected, followed by those with 30% bit density. By contrast, when a reference compound with 13% bit density is used, the distributions are shifted towards lower Tc values. However, relative to ZINC molecules with 42% bit density, molecules with 30% and 18% bit density now obtain in part higher Tc values and are more likely to be detected in similarity searching.

logical complexity. The bit density of keyed or hashed fingerprints generally increases with the topological complexity of test compounds. Bit density also tends to increase with molecular weight (size) because larger molecules often have more complex topology than smaller ones, although this is not always the case. Figure 2.4 also shows that simple reference compounds produce narrower Tc distributions in screening databases than more complex queries that typically generate broader value distributions. These effects have different consequences. On one hand, complex reference compounds can be more discriminatory than low-complexity queries because Tc values for fingerprint comparisons are more evenly spread over a wider range.[10] However, on the other hand, the ensuing shift towards higher Tc values also makes it more difficult to distin-

guish active compounds from database decoys. Hence, the outcome of similarity searching using reference compounds of different complexity is hard to predict. As will be discussed in the next chapters, systematic test calculations have revealed substantial complications of fingerprint searching that result from the use of complex queries.

Different similarity coefficients have also been systematically evaluated in fingerprint search calculations utilizing compound reference sets of varying complexity and the best-performing coefficient for each complexity level has been determined.[29] When reference and database compounds had comparable complexity, Tanimoto similarity calculations were found to be preferred over a wide range of experiments. However, when reference compounds were more complex than database molecules, the Forbes or simple match coefficient (see Table 1.2) performed best.[29]

2.5 Property descriptor value range-derived fingerprint

Different from the conventional fingerprint design reported in section 1.1, the so-called property descriptor value range-derived fingerprint, PDR-FP, is a class-directed 2D fingerprint that encodes database value ranges of molecular property descriptors.[49] Following this design strategy, value ranges of 93 property descriptors are determined for a screening database and binned into differently sized intervals so that the amount of screening database molecules falling into each interval is exactly the same (*equifrequent binning*). For a test compound, the matching descriptor intervals are determined and for each descriptor, the corresponding bit is set to "1".[49] The format of this fingerprint is easily adjustable for different screening databases and exactly 93 bits are always set on in this fingerprint, which consists of 500 bit positions in total.

Another unique feature of its design is the training potential for specific compound activity classes. This is achieved by calculating a non-binary bit vector for a compound reference set that emphasizes bit positions of individual value ranges that are conserved in active compounds (Figure 2.5). Applying a dot product similarity metric, this vector is then compared to individual PDR-FP representations of database molecules. This fingerprint has been shown to be particularly effective on compound classes of high structural diversity where other types of fingerprints produce only low compound recall or fail.[42,49]

2.6 Summary

In this chapter, the similarity search benchmarking protocol and workflow are introduced. Benchmarking calculations enable the evaluation of the similarity

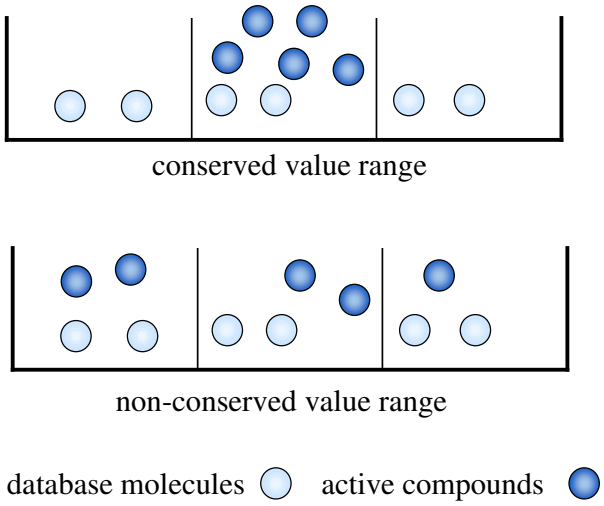


Figure 2.5: Conserved descriptor value ranges. Illustrated is the equifrequent binning of a hypothetical descriptor for hypothetical active and database molecules. The descriptor value range is divided into three bins and molecules are assigned to different bins according to their descriptor values. The number of database molecules assigned to each bin remains constant (two out of six). If all five active compounds have the same value range for this descriptor, then the value range is conserved and likely to be relevant for their activity.

search performance and are therefore applied to assess different methods presented in the following chapters. Similarity searching strategies such as data fusion and frequency-based approaches can be utilized to incorporate information from multiple reference fingerprints, which generally improves the search performance. In addition, molecular complexity effects are discussed for conventional similarity measures and the similarity value distributions are illustrated for comparing molecules with different complexity. Finally a novel fingerprint design, PDR-FP, is introduced, which depends on the value ranges of property descriptors. Conserved descriptors whose value ranges are potentially critical for identifying active molecules can be selected. Similarity searching using PDR-FP has been shown to be more powerful than other fingerprint types, especially in recovering structurally diverse hits.

Chapter 3

Complexity Effects in Tversky Similarity Searching

In similarity searching the evaluation of molecular similarity critically depends on the application of similarity measures for quantitative bit string comparison.[6] In Table 1.2 different similarity metrics are compared. A unique feature of the Tversky coefficient is the ability to put variable weights on the bit settings of molecules that are compared. By contrast, most similarity measures put equal weight on template and database molecules. Thus, these measures are symmetric in nature, which means that the results of pair-wise molecular comparisons are order-independent. Principal and statistical limitations associated with the use of similarity coefficients have been noted previously[10,50] and an elaborate analysis of different similarity measures and their strengths and weaknesses has been presented.[28]

Chen & Brown investigated the behavior of Tversky coefficients in large-scale similarity search calculations using three different 2D fingerprints and found that putting increasingly high weight on the bit string representations of template compounds produced higher hit rates than calculations using a symmetric coefficient with equal weights on template and NCI database molecules.[40,51] Chen & Brown interpreted their findings as "the first evidence of the presence of asymmetry in chemical similarity measures by an empirical study of two large databases".[51] The study by Chen & Brown represents an important advance because it highlights possible complications of molecular similarity assessment that are often not appreciated and enables further analyses of the observed effects, which will be discussed in this chapter. Furthermore, approaches to overcome such limitations of fingerprint comparisons will be discussed. For example, designing fingerprints that have constant bit density regardless of the nature of test molecules could eliminate the relative differences in bit densities and the induced complexity effects. Alternatively, introducing similarity metrics that are independent of bit densities could in principle also

avoid computational bias caused by complexity effects. For example, a modified version of the Tanimoto coefficient has been reported that can be applied to balance discrepancies in bit settings.[52] A bit density-independent variant of the Tversky coefficient, *weighted Tversky coefficient (wTv)*, will be introduced that makes it possible to systematically change the relative contributions of bits that are set on or off in similarity calculations. The behavior of this coefficient in similarity searching will be thoroughly characterized for compounds having different degrees of complexity and the relationship between complexity, similarity values, and search performance will be analyzed.

3.1 Properties of the Tversky coefficient

For two molecules being compared and represented by fingerprint bit strings A and B, Tversky coefficients (Tv) are defined as follows:[27]

$$Tv(A,B,\alpha) = \frac{c}{\alpha(a-c)+(1-\alpha)(b-c)+c} \tag{3.1}$$

with α in [0, 1]. Here, a represents the number of bits set on in A, b the number of bits set on in B, and c the number of bits set to "1" in both bit strings. The α parameter varies between zero and one and determines the relative weight on uniquely set bits. For $\alpha = 0.5$ equal weights are put on both molecules (and the Tversky coefficient becomes the symmetric Dice coefficient,[6] see Table 1.2), whereas for $\alpha > 0.5$ or $\alpha < 0.5$ more weight is put on bits that are exclusively set on in A or B, respectively. If A and B are compared and their bit string representations have exactly the same number of bits set on, Tversky coefficients are symmetric, which means that comparing A with B and B with A produces the same value. If the bit densities of A and B differ, the comparison becomes order-dependent for $\alpha \neq 0.5$ and the corresponding Tversky coefficients are asymmetric.

Tv can be transformed as follows:

$$\begin{aligned} Tv(A,B,\alpha) &= \frac{c}{\alpha(a-c)+(1-\alpha)(b-c)+c} \\ &= \frac{c}{\alpha(a-b)+b} \end{aligned} \tag{3.2}$$

which has the format of a hyperbola function of variable α. Figure 3.1 illustrates this hyperbola function under two situations: $a-b>0$ (left) and $a-b<0$ (right). In both cases only the part with positive $Tv(\alpha)$ values (colored in blue) are considered. It can be seen that when $a-b>0$, $Tv(\alpha)$ increases with α and when $a-b<0$, $Tv(\alpha)$ decreases with α. When $a-b=0$, this function does not depend on the value of α.

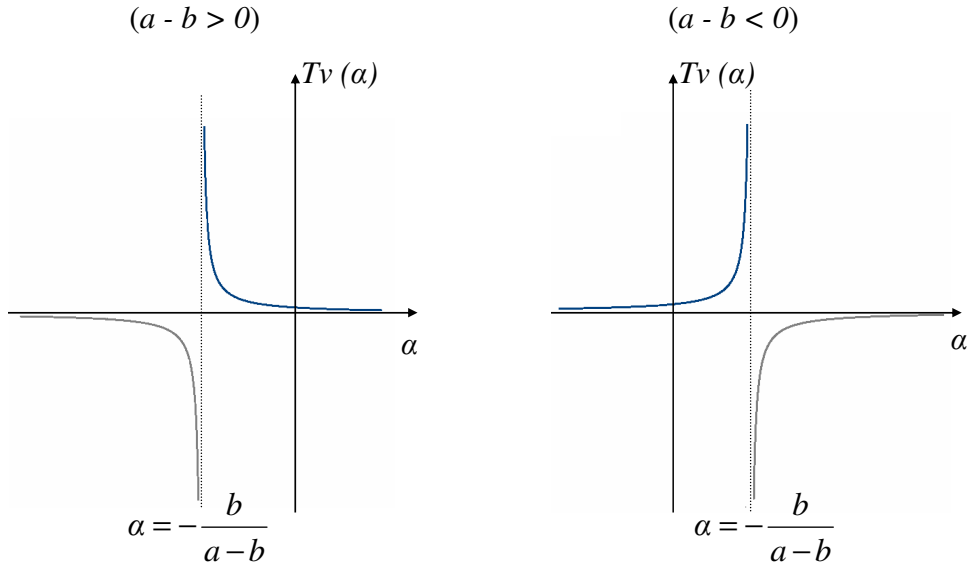


Figure 3.1: Hyperbola function. The hyperbola function $Tv(\alpha)$ is illustrated for two different cases: $a - b > 0$ (left) and $a - b < 0$ (right). The positive part of $Tv(\alpha)$ is colored in blue in both cases. The curve is monotonously increasing when $a - b > 0$ and decreasing when $a - b < 0$.

In the following example, Tversky similarities from relative differences in fingerprint bit settings of hypothetical molecules A, B_1, B_2, and B_3 are determined under systematic variation of α. The corresponding bit numbers are a, b_1, b_2, and b_3, respectively. Characteristic features of Tversky similarity can be best rationalized when studying examples that produce large variations in similarity values. This is the case when comparing a test molecule with a sub- and superstructure and, in addition, another molecule having the same fingerprint bit density.

Figure 3.2 shows the similarity curves for comparisons of A with B_1, B_2, and B_3, respectively. For the A vs. B_1 and A vs. B_3 comparisons, convex curves are obtained whose gradients strongly depend on the differences between a and b_i. Assuming $c \neq 0$, for $a > b_1$ Tv values are monotonously decreasing and for $a < b_3$ they are monotonously increasing. Figure 3.2 also shows the difference in similarity values for comparison of molecules A with B_1 and B_3, respectively, when α is set to 0.5 and Tv becomes a symmetric coefficient. This reflects a general bit density-dependence of the Tversky similarity measure.

In this example, molecule A sets 50 of 100 hypothetical fingerprint bits to one. Molecule B_1 is a substructure of A having 25 fewer bits set on, B_2 is another molecule that – like A – has also 50 bits set on but only 37 in common with it, and B_3 is a superstructure of A having 25 more bits set to one. Comparison of A and B_1 leads to a Tv similarity value of 1.0 for α value

of 0, comparison of A and B_2 to 0.74 for all α values, and A and B_3 to 1.0 for $\alpha = 1$. Thus, for extreme α values Tversky similarity calculations become akin to substructure searching. For α values close to one, test molecules achieve high Tv values if they contain the query compound as a substructure (blue curve in Figure 3.2). By contrast, for α values approaching zero, molecules obtain high Tv values if they themselves are substructures of the query (red curve in Figure 3.2). In Figure 3.3 an example of superstructure searching is shown. Given an arbitrary 4-bit fingerprint design, two molecules, A and B, are compared. In this case A is a superstructure of B ($a > b = c$). As a result, Tv decreases when α increases and its maximal value of 1.0 is achieved when $\alpha = 0$.

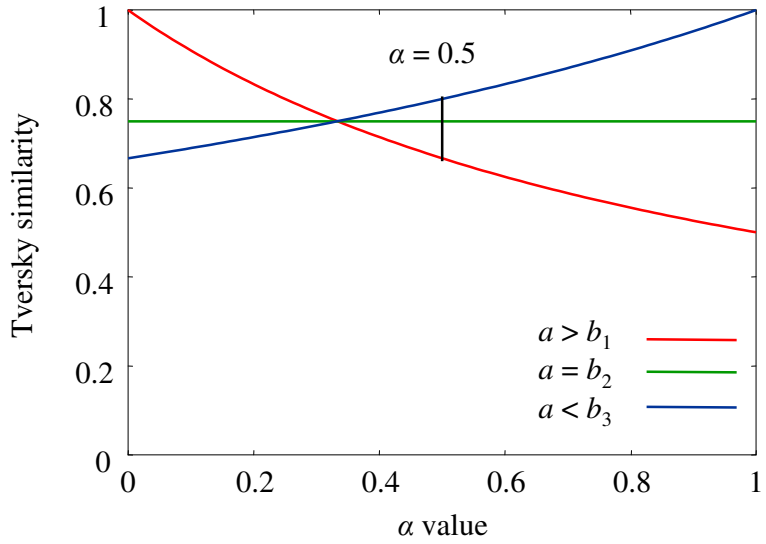


Figure 3.2: Property of the Tversky coefficient. Reported are Tversky similarity values for a template compound A compared to three different database molecules B_i (or hypothetical fingerprints with a and b_i bits set to one, respectively) as a function of the weighting parameter α. Three cases are shown: $a > b_1$ (fewer bits are set on in B_1 than in A), $a = b_2$ (the same number of bits set on in both compounds), and $a < b_3$ (more bits are set on in B_3). The differences, $a - b_1$ and $b_3 - a$, are set to be equal. The black bar marks the difference in the two similarity values of B_1 and B_3 for $\alpha = 0.5$ (symmetric Tversky coefficient).

In addition to differences in specific bit settings, overall differences in bit densities also lead to a separation of molecules depending on α parameter values. For example, if active compounds have comparable bit densities but on average a higher bit density than inactive molecules, the $a > b_1$ case applies for the comparison of active against inactive molecules. As a consequence, if α increases, similarity values decrease for inactive database molecules but are mostly unaffected for active compounds (case $a = b_2$, as shown in Figure 3.2) leading to a preferential de-selection of inactive molecules. By contrast, if bit

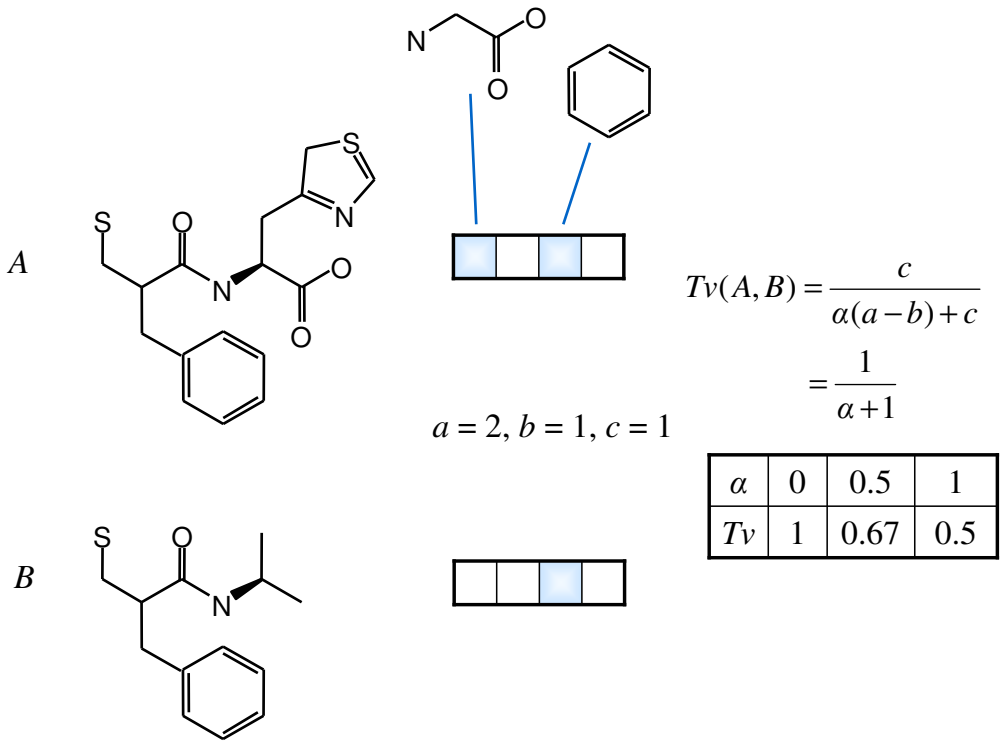


Figure 3.3: Superstructure searching using the Tversky coefficient. Given an arbitrary 4-bit fingerprint design, two molecules, A and B, are compared. A is a superstructure of B and has one more bit set on than B. Tv decreases with increasing α and is maximal when $\alpha = 0$.

strings of active compounds have similar bit densities but systematically lower bit densities than inactive molecules, the $a < b_3$ case applies and, according to Figure 3.2, lowering α will lead to a de-selection of inactive molecules.

Figure 3.2 also reveals another general characteristic of the Tversky coefficient. As discussed above, in its symmetric version ($\alpha = 0.5$), it assigns higher similarity values to molecules that have more bits set on than to molecules with fewer bits, even if their distance to an active reference compound is the same in "bit string space", i.e., molecules B_1 and B_3 both deviate in exactly 25 bit positions from A ($a - b_1 = b_3 - a$). However, comparison of A and B_3 results in a significantly higher similarity value than the comparison of A and B_1. That is because the "1" bits dominate the Tversky similarity comparison: the increase of "1" bits affects the similarity value more than the decrease of "1" bits (i.e., increase of "0" bits). These theoretical considerations apply to any molecular fingerprint design that depends on structural complexity and systematically affct calculations of Tversky similarity.

3.2 Molecular complexity and fingerprint characteristics

One measure of molecular complexity is the number of heavy atoms. In order to investigate the behavior of molecular complexity effects, the number of heavy atoms was assessed for both active compounds and database molecules. In Table 3.1 characteristics of five activity classes extracted from MDDR[38] as well as the background NCI database[40] used by Chen & Brown[51] are shown. For five activity classes and the NCI background database, the average number of non-hydrogen atoms was calculated as a measure of molecular size. Also determined for each compound set was the average number of bits set on in three different fingerprints, MACCS, TGD, and PDR-FP. For the five activity classes, average numbers of non-hydrogen atoms ranged from 14.0 to 32.3 and for the NCI database, the average number was 25.2. Activity class NNI was assembled to consist of on average much smaller molecules than the other classes and showed significantly lower bit density for MACCS and TGD. For PDR-FP, bit densities did not vary because this fingerprint was designed to have a constant number of bits set on, independent of molecular size.[49]

class	**designation**	**number of compounds**	**number of heavy atoms**	**bit density MACCS (%)**	**bit density TGD (%)**	**bit density PDR-FP (%)**
BEN	benzodiazepine agonists	57	25.6	30.8	13.4	18.6
CAT	cathepsin inhibitors	90	32.3	30.2	20.8	18.6
HH2	histamin H2 antagonists	41	27.6	33.5	23.0	18.6
NNI	neuronal injury inhibitors	50	14.0	20.3	6.0	18.6
TNF	TNF-α release inhibitors	65	31.0	31.7	19.7	18.6
NCI	NCI anti-AIDS database	42687	25.2	25.7	13.2	18.6

Table 3.1: Characteristics of compound sets for Tv calculations. Reported are the number of compounds, average number of non-hydrogen (or heavy) atoms, and average bit densities for three different 2D fingerprints, MACCS, TGD and PDR-FP, for each of the five activity classes and the background database.

Compound class complexity and pair-wise Tversky similarity

Pair-wise Tversky similarities were calculated for compounds within each activity class and also between activity classes and NCI compounds under systematic variation of α parameter values. The results are shown in Figure 3.4. For MACCS and TGD, average similarity values within each activity class formed symmetric curves with a minimum at $\alpha = 0.5$. This is the case because for each pair of active molecules A_1 and A_2, both values Tv(a_1, a_2) and Tv(a_2, a_1) contribute to the overall average value.

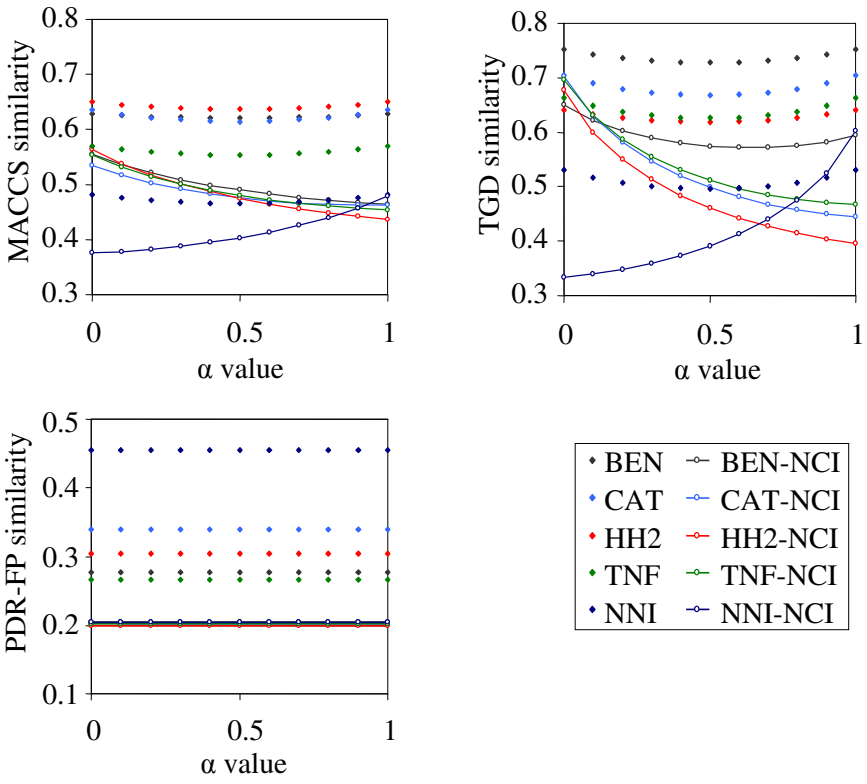


Figure 3.4: Pair-wise Tversky similarity. Shown are the average pair-wise Tv similarity values with varying α (using a step-size of 0.1). Dots represent average similarity within each activity class and the corresponding color-coded lines represent average similarity of NCI database molecules when compared to the classes.

In comparison, average Tv values for activity classes against NCI compounds did not follow symmetric curves but were monotonously decreasing for classes BEN, CAT, HH2, and TNF, and monotonously increasing for NNI. Since average complexity was lower for NCI than BEN, CAT, HH2, and TNF compounds (Table 3.1), similarity values decreased for increasing α values and NCI

molecules were preferentially de-selected, which corresponds to the $a > b_1$ case in Figure 3.2. By contrast, NNI had lower average complexity than NCI, leading to increasing similarity values when α increased and preferential selection of NCI compounds, which corresponds to the $a < b_3$ case in Figure 3.2. As can be seen in Figure 3.4, by far the smallest differences between similarity values for variation of α were observed for BEN relative to the NCI database when using the TGD fingerprint. This was a consequence of the fact that BEN and NCI compounds produced nearly the same average bit density (13.4% vs. 13.2%, Table 3.1). These results were perfectly in accord with theoretical expectations.

For PDR-FP, average similarities formed no monotonously increasing or decreasing curves, but horizontal lines. This was because PDR-FP has consistently 93 bits set on for each molecule and therefore Tv becomes completely independent of the α parameter. This is obvious if the Tversky formula in Eq.(3.2) is transformed accordingly:

$$\begin{aligned} Tv(A,B,\alpha) &= \frac{c}{\alpha(a-b)+b} \\ &\overset{a=b}{=} \frac{c}{b} \\ &= \frac{c}{93} \end{aligned} \tag{3.3}$$

The Tv value now only depends on the number of common "1" bits out of the total number of "1" bits in the fingerprints.

Similarity distribution overlap

In similarity searching, hit rates depend on differences between the distributions of (a) pair-wise intra-class similarity values and (b) similarity values for active vs. database molecules. As can be seen in Figure 3.4, when average similarity values were calculated, maximal differences and lowest similarity values between activity classes and NCI compounds for fingerprints MACCS and TGD were achieved for $\alpha = 1$ (BEN, CAT, HH2, TNF) or $\alpha = 0$ (NNI). Yet it cannot be assumed that performance is optimized at $\alpha = 1$ and $\alpha = 0$, respectively, because until now, only average similarity values have been considered. However, individual molecules can deviate in Tv scores and thus affect hit rates. Therefore, for the comparison of similarity value distributions, one also needs to take standard deviations into account. There are two effects that minimize the overlap of two distributions and hence increase hit rates. First, the larger the difference between average similarity values is, the further the distributions are apart. Second, the smaller the standard deviations are, the narrower the distributions become and the smaller their intersection area is. As an example, distributions for similarity values withing activity class HH2 and between HH2 and NCI are shown in Figure 3.5.

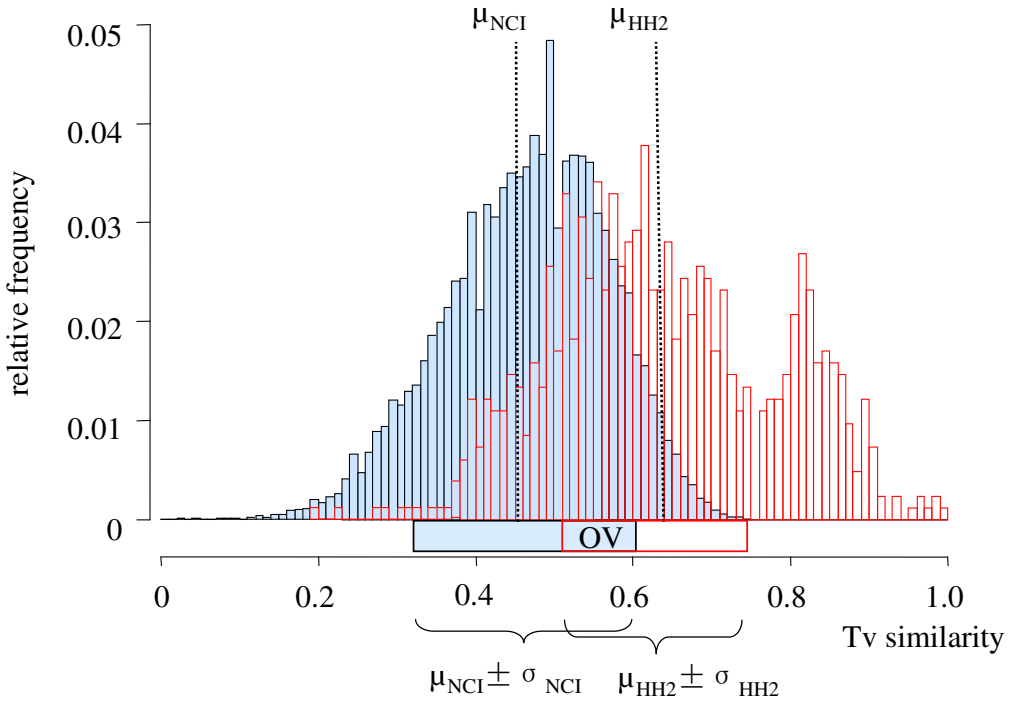


Figure 3.5: Tversky similarity distributions. Value distributions for pair-wise Tversky similarities ($\alpha = 0.5$) within activity class HH2 (red) and between HH2 and the NCI database (blue) are shown. The position of the average value (μ_{HH2} or μ_{NCI}) for each distribution is indicated by a dotted line. The intervals $[\mu_{HH2} \pm \sigma_{HH2}]$ and $[\mu_{NCI} \pm \sigma_{NCI}]$ are represented by a red and blue box, respectively. The area "OV" represents the overlap of the intervals, as discussed in the text.

In light of its relevance, a simple measure that approximates the overlap of two similarity distributions has been defined (see Figure 3.5). Given two distributions of intra-class similarities (AC) and similarities between active and database molecules (DB), the overlap (OV) is defined as:

$$OV \quad = \quad (\mu_{DB} + \sigma_{DB}) - (\mu_{AC} - \sigma_{AC}) \tag{3.4}$$

Here μ_{AC} and μ_{DB} are mean values and σ_{AC} and σ_{DB} standard deviations of the two distributions. For similarity searching it is assumed that $\mu_{AC} > \mu_{DB}$.

By plotting OV as a function of the α parameter (Figure 3.6), α values can be determined that minimize the overlap between the distributions and are thus preferred for similarity searching. These α values (approximated using a step-size of 0.1) are reported in Table 3.2. For MACCS and TGD, optimal α values were greater than 0.5 for activity classes CAT, HH2, and TNF, and smaller than 0.5 for NNI. For BEN, optimal α values were 0.6 for MACCS

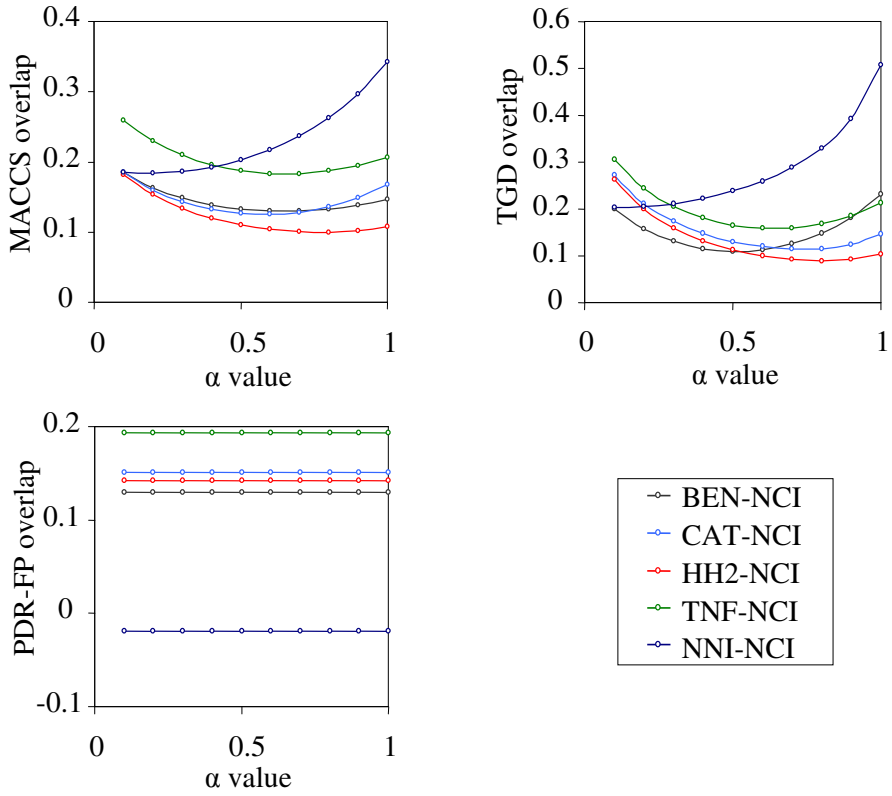


Figure 3.6: Tversky similarity overlap. The overlap OV between intra-class and inter-class Tversky similarity value distributions is shown as a function of the α parameter.

class	MACCS	TGD	PDR-FP
BEN	0.6	0.5	-
CAT	0.6	0.7	-
HH2	0.8	0.6	-
NNI	0.2	0.1	-
TNF	0.6	0.8	-

Table 3.2: Optimal Tv α values. α values producing minimal overlap between intra-class and class-NCI Tversky similarity value distributions are shown as determined by graphical analysis of Figure 3.6. PDR-FP calculations are independent of α values because of its constant bit density. Therefore the overlap is also constant.

and 0.5 for TGD, where average bit densities were nearly identical for BEN and NCI. For PDR-FP, OV was constant because of its constant bit density and the results of search calculations were independent of α values. Taken together, these results confirmed that differences in fingerprint bit densities determine parameter settings for optimal Tversky similarity calculations. With the complexity-independent PDR-FP it is possible to circumvent complexity effects. Yet another possibility is to modify the similarity metric in use so that the fingerprint representation can remain unmodified.

3.3 Development of the weighted Tversky coefficient

When Tc calculations were used to guide the selection of diverse compound subsets from libraries, selected molecules often displayed the tendency to be smaller than average database molecules because larger molecules having higher Tc were determined to be more similar.[53] These observations have prompted Fligner et al.[52] to introduce a modified version of the Tanimoto coefficient (MTc) that takes all bit position into account (i.e. set on or off):

$$MTc(p) = \frac{2-p}{3}Tc_1 + \frac{1+p}{3}Tc_0 \tag{3.5}$$

In this formulation, Tc_1 and Tc_0 are Tanimoto coefficients calculated for bits set on and off, respectively. The parameter p was empirically determined to adjust bit density effects. Using this modified coefficient, Fligner et al. were able to avoid the prevalence of small compounds in diverse subsets taken from the NCI database.[52]

The relationship between "1" bits in two fingerprints A and B also determines the complexity dependence of Tversky similarity calculations. As discussed above, if a reference compound has more bits set on than database molecules, similarity values tend to decrease with increasing α. By contrast, if a reference compound has fewer bits set on, similarity values tend to increase with increasing α. Corresponding relationships between "0" bits in fingerprints also systematically change similarity values when α increases but the directions are reversed compared to "1" bits. Thus, for Tv calculations, it is immediately apparent that taking both "1" and "0" bits into account provides a principal possibility to eliminate the influence of complexity or size effects because complexity effects caused by "1" bits and "0" bits can cancel out each other. A form of the Tversky coefficient accounting for bits that are set off can be written as follows:

$$\begin{aligned} Tv'(A,B,\alpha) &= \frac{c'}{\alpha(a'-c')+(1-\alpha)(b'-c')} \\ &= \frac{c'}{\alpha(a'-b')+b'} \end{aligned} \tag{3.6}$$

where a' and b' denote the number of "0" bits in A and B, respectively, and c' the number of "0" bits common to both. Using a weighted combination of Tv and Tv' (*weighted Tversky coefficient, or* wTv) it is possible to balance different densities of "1" and "0" bits in fingerprints such that neither "1" nor "0" bits dominate similarity evaluation:

$$wTv(A,B,\alpha,\beta) = \beta\frac{c}{\alpha(a-b)+b} + (1-\beta)\frac{c'}{\alpha(a'-b')+b'} \tag{3.7}$$

where β is defined as the weight on "1" bits, i.e., the larger β becomes, the more weight is put on "1"s and the less on "0"s; for $\beta = 1$, $wTv = Tv$ and for $\beta = 0$, $wTv = Tv'$. The above equation can be further transformed:

$$wTv = \beta\left(\frac{c}{\alpha(a-b)+b} - \frac{c'}{\alpha(a'-b')+b'}\right) + \frac{c'}{\alpha(a'-b')+b'} \tag{3.8}$$

In this formulation, the term

$$\left(\frac{c}{\alpha(a-b)+b} - \frac{c'}{\alpha(a'-b')+b'}\right)$$

can be viewed as a coefficient of β. When it is greater than 0, the linear function $wTv(\beta)$ monotonously increases. By contrast, when the coefficient is negative, the function monotonously decreases. The characteristics of this coefficient are determined by the value of α and the intrinsic bit settings of the fingerprints that are compared. The bivariate function $wTv(\alpha,\beta)$ is expected to have a nontrivial value distribution surface for different (α,β) combinations and systematic variation of the α and β parameters best describes this similarity metric. However, some general characteristics can be deduced by comparing cases where search templates and active database compounds (potential hits) have significant differences in bit density and where bit densities are similar.

When all other parameters in Eq.(3.7) remain constant and the reference compounds have fewer bits set on than potential hits, i.e. $a < b$, then the term

$$\frac{c}{\alpha(a-b)+b}$$

increases due to the decrease of the denominator. If $a < b$, it also follows that $a' > b'$ (because a' and b' are complementary to a and b). This reduces the term

$$\frac{c'}{\alpha(a' - b') + b'}$$

and, as a result, the term

$$\beta(\frac{c}{\alpha(a - b) + b} - \frac{c'}{\alpha(a' - b') + b'})$$

increases relative to the situation where bit densities are similar. Increasing α and β values will further amplify this trend, which also favors the detection of hits.

By contrast, when reference compounds have more bits set on than potential hits, i.e., $a > b$, the term

$$\frac{c}{\alpha(a - b) + b}$$

decreases and the term

$$\frac{c'}{\alpha(a' - b') + b'}$$

increases, thereby reducing

$$\beta(\frac{c}{\alpha(a - b) + b} - \frac{c'}{\alpha(a' - b') + b'})$$

and the resulting wTv values. The larger the difference between a and b is, the more difficult it becomes to achieve high wTv values for comparisons between reference compounds and active database compounds. In fact, the term

$$\beta(\frac{c}{\alpha(a - b) + b} - \frac{c'}{\alpha(a' - b') + b'})$$

could potentially become negative, which would significantly reduce wTv values for potential hits and make it very difficult to distinguish them from other database molecules. Thus, differences in complexity between reference and active database compounds might significantly complicate similarity evaluation and present difficult fingerprint search situations. Modulating α and β parameters accordingly can reverse the trend, as further analyzed and discussed below.

Balancing complexity effects

To study the effects of fixed β values under systematic variation of α, calculations were carried out on five compound classes assembled from the MDDR database.[38] These classes included benzodiazepines (abbreviated BEN; 57 compounds), cathepsin inhibitors (CAT; 90), vasopressin antagonists (VAS; 109), neuronal injury inhibitors (NNI; 50), and tumer necrosis factor α release inhibitors (TNF; 65). With the exception of VAS, these activity classes were previously used in calculations in section 3.2 (Table 3.1). They were designed to produce fingerprints with different average bit densities. VAS was newly assembled from the MDDR and had by far the highest average bit density among the classes studied. The NCI database[40] was adopted as background database (see Table 3.1).

class	**number of compounds**	**bit density (%)**
BEN	20	26.0
CAT	20	30.8
CAT	40	31.0
CAT	60	30.8
CAT	80	31.0
TNF	20	40.8
VAS	20	46.0
NNI	20	15.2

Table 3.3: Reference sets for pair-wise wTv similarity calculation. Reported are the number of compounds and average MACCS bit densities for eight reference sets extracted from five activity classes. The background database, NCI, contains 42,687 compounds and their average MACCS bit density is 25.7%.

For similarity calculations, subsets of 20 compounds were selected from each activity class (except for CAT, where subsets of 20 to 80 compounds were generated to assess the parametric dependence on reference set size). The MACCS fingerprint[16] bit densities of these activity classes significantly differed. Table 3.3 summarized the reference sets and their bit densities. For these compound classes, MACCS "1" bit densities range from 15% - 46%. Thus, "1" bits are sparsely set and "0" bits dominate the fingerprint bit settings.

Each active compound was used as an individual template and searched against the background database. For each reference set, average pair-wise wTv similarity values were determined for α values ranging from 0 to 1 and constant β values of 0, 0.5, and 1, respectively. The similarity profiles in Figure 3.7 and Figure 3.8 report the average database similarity for given β and systematically changing α values. For $\beta = 0$, all weight is put on the "0" bits and for $\beta = 1$ all weight on the "1" bits. For $\beta = 0.5$, "0" and "1" bits are equally weighted. Thus, β settings of 0 or 1 emphasize complexity effects, whereas 0.5 eliminates

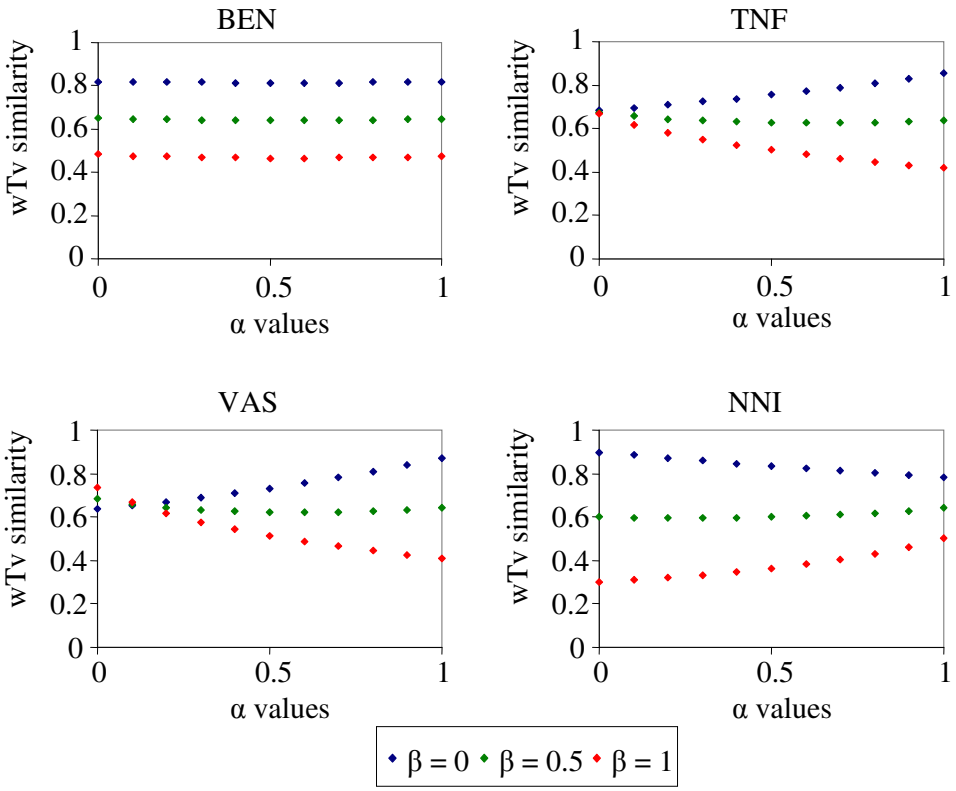


Figure 3.7: Pair-wise wTv using reference sets with different complexity levels. For four activity classes, average weighted Tversky similarity of background database molecules was calculated using the MACCS fingerprint. For each class, three curves were recorded for systematic variation of α and β values of 1 (i.e. complexity-dependent calculations over-weighting "1" bits), 0.5 (complexity-independent), and 0 (over-weighting "0" bits).

them from similarity evaluation. For α values ranging from 0 to 1, increasing weight is put on the bit settings of reference compounds; $\alpha = 0.5$ equally weights reference and database molecules. Thus, wTv values calculated with $\alpha = 0.5$ and $\beta = 1$ are proportional to conventional Tanimoto similarity.

As can be seen from Figure 3.7, asymmetric similarity curves were obtained for activity classes whose bit densities differed from the database average. When bit densities of active molecules were higher than the database, the curves were monotonously increasing for $\beta = 0$ and decreasing for $\beta = 1$. When bit densities of active molecules were lower, these trends were reversed. Only BEN produced similarity values that were essentially constant over the entire α range because its bit density was very similar to the background database. When β was set to 0.5 complexity effects were balanced and the similarity values were largely constant over the α range. Although BEN matched the bit density of

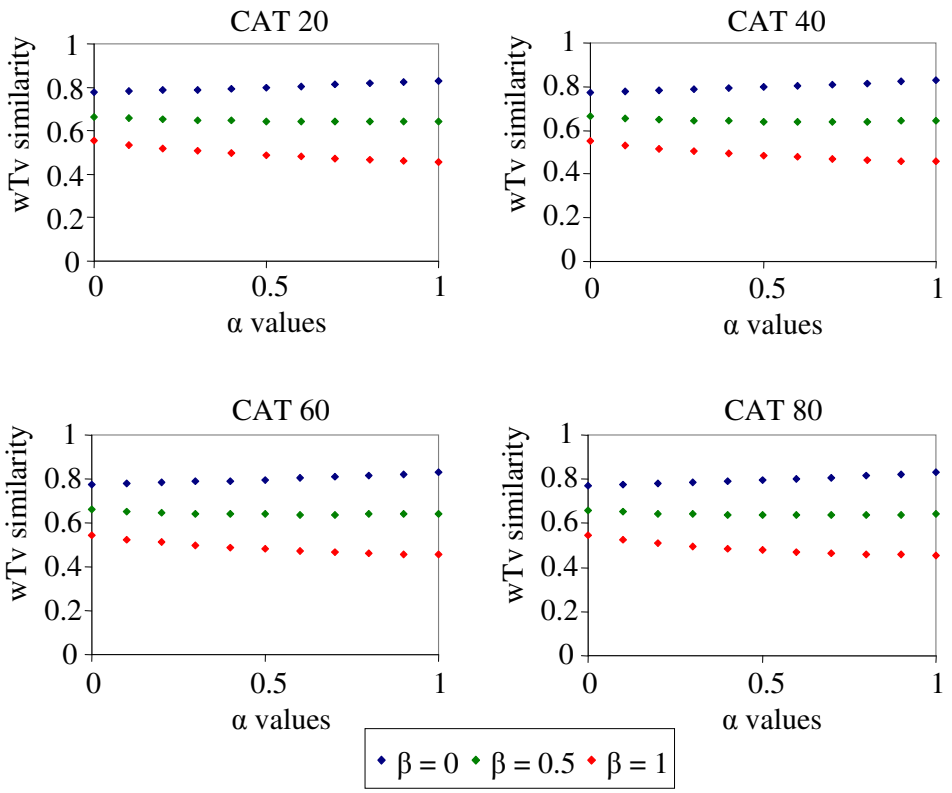


Figure 3.8: Pair-wise wTv using reference sets with different set sizes. For reference sets of four different sizes of class CAT, average weighted Tversky similarity of background database molecules was calculated using the MACCS fingerprint. For each class, three curves were recorded for systematic variation of α and β values of 1 (i.e. complexity-dependent calculations over-weighting "1" bits), 0.5 (complexity-independent), and 0 (over-weighting "0" bits).

the database, curves for β settings of 1 and 0 illustrate the consequences of sparsely set "1" bits in the MACCS fingerprint (bit density of 26%). At the ($\alpha = 0.5$, $\beta = 1$) reference point, the average similarity of 0.47 was artificially low; when complexity effects were balanced, i.e. ($\alpha = 0.5$, $\beta = 0.5$), the average similarity was 0.64. Fingerprints of all activity classes and database molecules contained more "0" than "1" bits and thus similarity values for $\beta = 0$ were generally higher than $\beta = 1$. Balanced average similarity relative to the database was ~0.65 for four activity classes and 0.6 for NNI. Thus, as one should expect, the average similarity calculated for a large number of database molecules was comparable for different activity classes when complexity no longer influenced the calculations. The CAT profiles (Figure 3.8) show that the similarity curves did not depend on the size of the reference set.

Taken together, these data illustrate the influence of complexity effects

on similarity calculations and show that wTv calculations with $\beta = 0.5$ produce essentially constant similarity values that are independent of relative weights on reference and database molecules. Thus, in this case, database search calculations on active molecules are no longer biased by artificially increasing or decreasing similarity values.

Active compounds of different complexity

Retrieval of active compounds and determination of hit rates present challenges that go beyond the similarity evaluation presented in Figure 3.7 because the detection of molecules having similar activity requires successfully distinguishing potential hits from average database molecules. Specific bit patterns must be detected that are only shared by active molecules.

To investigate the role of varying bit densities in similarity search calculations under systematic variation of α and β, a set of 1,214 tyrosine kinase inhibitors (TKI) was assembled from the MDDR[38] and divided into four subsets with increasing average MACCS fingerprint "1" bit density (from TKI01 to TKI04), as reported in Table 3.4. The lowest- (TKI01) and highest-complexity (TKI04) subsets were used as reference sets in separate calculations where the remaining three subsets were added to the background NCI database as potential hits. For each reference compound, search calculations were carried out under systematic variation of α and β, the top scoring 100 or 500 database molecules were selected, and hit rates calculated and averaged for each subset, thus producing set-specific $HR(\alpha, \beta)$ values. For example, $HR(0.3, 0.6)$ reports the hit rate calculated for wTv ($\alpha = 0.3$, $\beta = 0.6$) used as the similarity coefficient. $HR(\alpha, \beta)$ can be plotted as a 2D landscape map illustrating the relationship between the two parameters and the search results.

subset	number of compounds	bit density (%)
TKI01	300	18.8
TKI02	300	25.2
TKI03	300	31.0
TKI04	314	39.5

Table 3.4: Subsets of TKI for wTv similarity calculation. Reported are the number of compounds, and average MACCS bit densities for four TKI subsets used in calculations of Section 3.3. The background database, NCI, contains 42,687 compounds and their average MACCS bit density is 25.7%.

For low-complexity reference set TKI01 (Figure 3.9), top hit rates between 25% and 45% were obtained with MACCS for selection sets of 100 database molecules. For high-complexity reference set TKI04(Figure 3.10), hit rates were generally lower (10% to 20%). In both cases, it can be observed

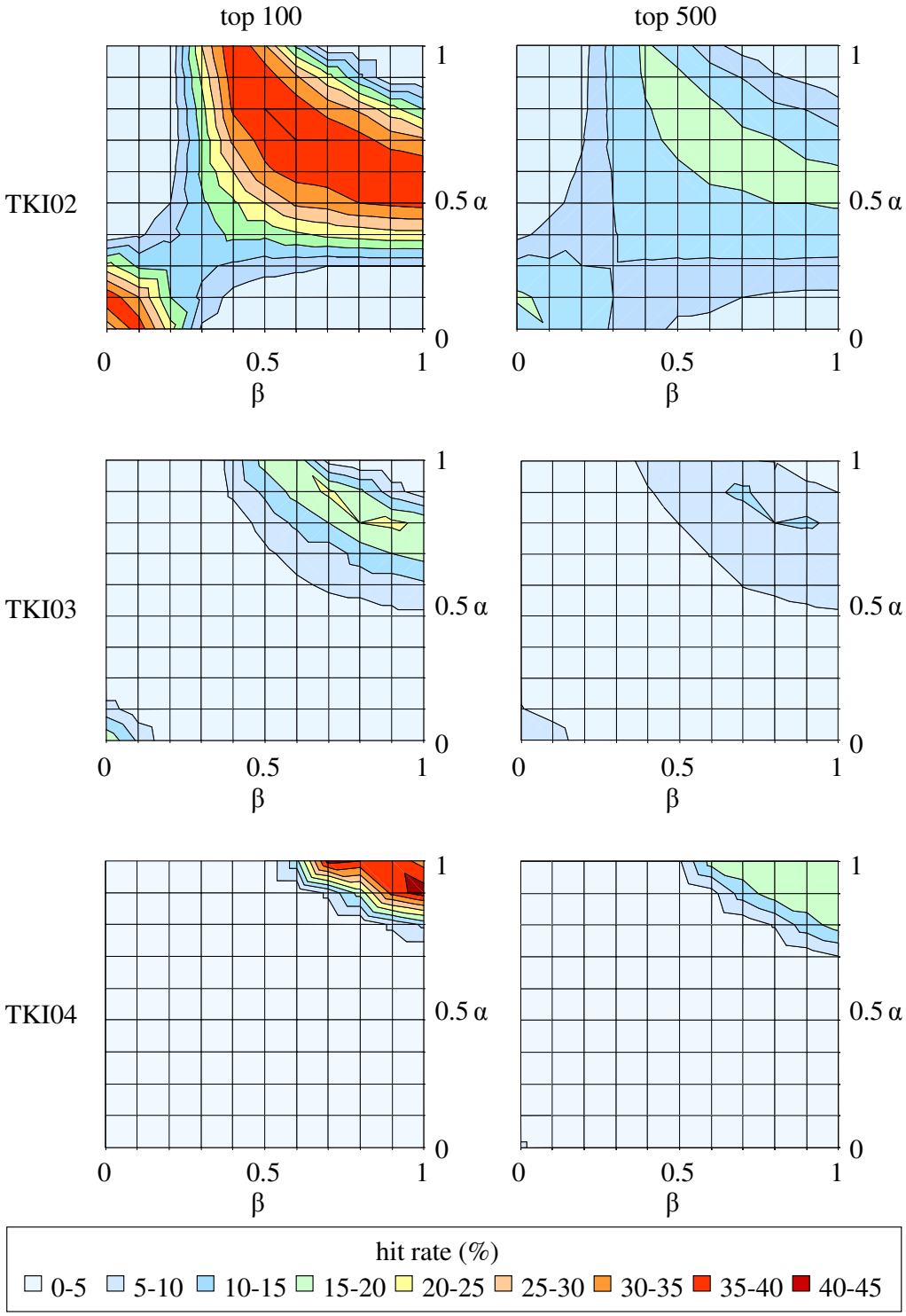


Figure 3.9: Hit rate landscapes using simple references. Reported are similarity search results for reference set TKI01 and ADC sets TKI02 (top), TKI03 (middle) and TKI04 (bottom). Hit rates from top 100 (left) and top 500 (right) molecules are reported under systematic variation of the α and β parameters in increments of 0.1.

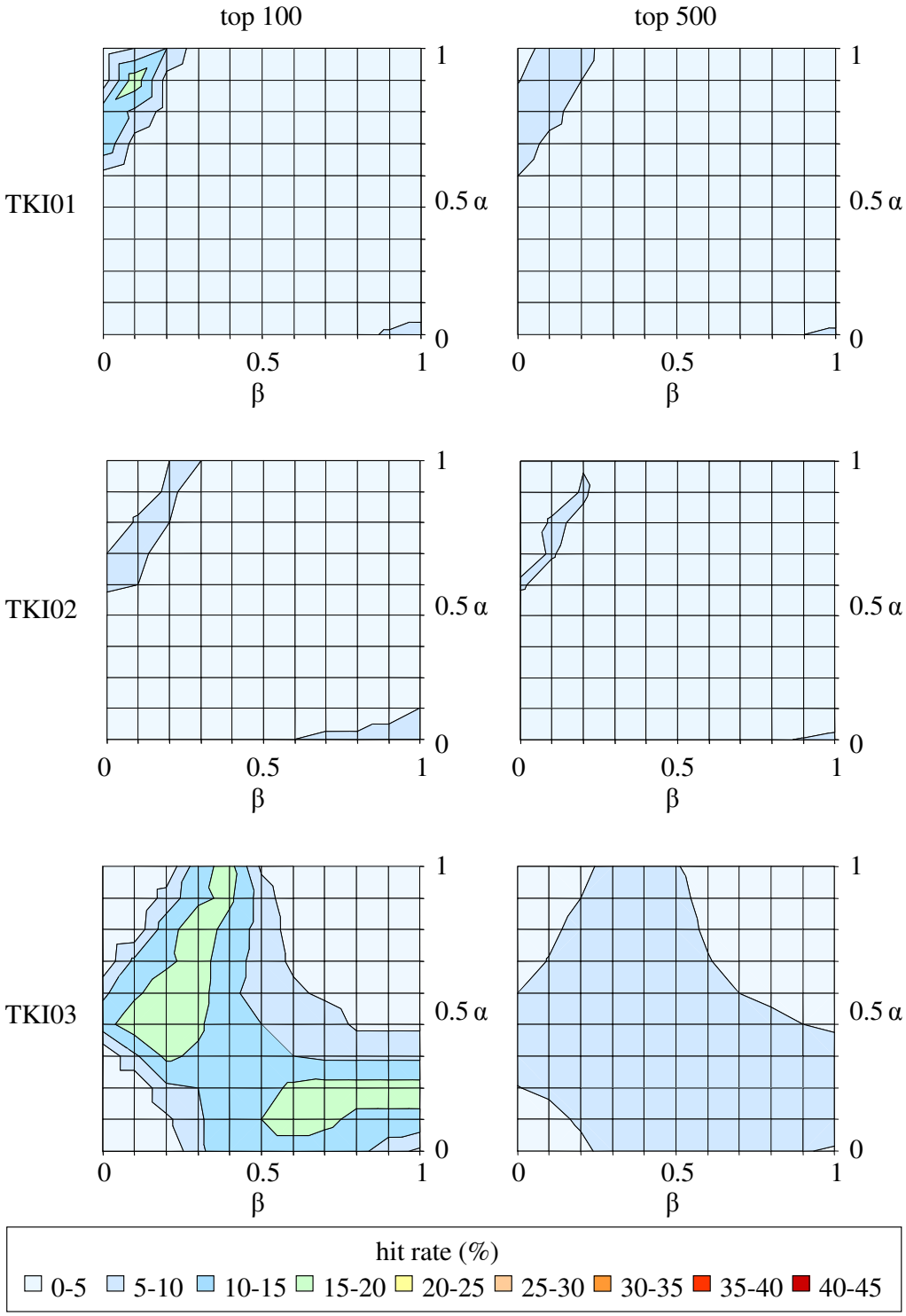


Figure 3.10: Hit rate landscapes using complex references. Reported are similarity search results for reference set TKI04 and ADC sets TKI01 (top), TKI02 (middle) and TKI03 (bottom). Hit rates from top 100 (left) and top 500 (right) molecules are reported under systematic variation of the α and β parameters in increments of 0.1.

that multiple (α, β) combinations produced preferred hit rates. However, top hit rates were generally not observed at the ($\alpha = 0.5$, $\beta = 1$) reference point for conventional similarity assessment. In fact, when bit densities of reference compounds and hits were different, similarity calculations using these parameter settings generally failed. However, top hit rates were typically also not produced by the ($\alpha = 0.5$, $\beta = 0.5$) parameter settings, i.e. when complexity effects were balanced ($\beta = 0.5$) and equal weight was put on the bit settings of reference and database molecules ($\alpha = 0.5$). In calculations with reference compounds and potential hits having similar bit density (top panel in Figure 3.9, bottom in Figure 3.10), different (α, β) combinations produced top hit rates. When bit densities of reference compounds, potential hits, and database molecules were comparable, complexity effects only played a minor role. However, as discussed above, "0" bits dominated all fingerprint settings and therefore, increasing weight on shared "1" bits (i.e. increasing β) often improved hit rates in these cases. The top panel in Figure 3.9 and bottom panel in Figure 3.10 also show an apparent approximate symmetry of hit rates along the ($\alpha = \beta$) diagonal because complementary combinations of (α, β) values produce equivalent (high or low) hit rates. Importantly, when the complexity of reference compounds and potential hits differed, clear preferences for (α, β) combinations were observed. If the bit density of reference compounds was lower than that of potential hits (reference set TKI01, Figure 3.9) combinations of high α and high β values produced best hit rates. By contrast, if the bit density of reference compounds was higher than that of potential hits (reference set TKI04, Figure 3.10) combinations of high α and low β values were preferred. In both cases, these parameter combinations increased wTv values for potential hits, which can be deduced from the wTv formula. Thus, these results are generally expected for reference compounds and hits having different fingerprint bit density. In these cases, modulating complexity effects, rather than eliminating them, and putting high weights on the bit settings of reference compounds optimized retrieval of active compounds.

Virtual screening scenario

In the previous section, the complexity of potential hits was systematically changed and the search results illustrated in Figure 3.9 and Figure 3.10 reveal systematic trends of parametric preference. In this section, search calculations are analyzed for potential hits that closely matched the bit density of the background database and reference compounds of different complexity. The two instances where reference compounds have bit densities higher than or comparable to the database typically apply to practical virtual screening situations. This is the case because reference compounds for virtual screening are often optimized leads or drug candidates (having high complexity) or, alternatively,

hits taken from experimental screening campaigns (with complexity comparable to the database).

The average MACCS "1" bit density of the background database (25.7%) was taken as a reference point to search for molecules that closely matched this density (i.e. hits with complexity comparable to an average database molecule). For two activity classes (TKI and TNF), sets of compounds were assembled from the MDDR having bit densities very similar to the background database (TKI: 250 compounds, average bit density 25.2%; TNF: 250, 25.8%). These sets were added to the background database as potential hits. Then other sets of 50 compounds having average bit densities smaller than, comparable to, or larger than the background database were used as search templates as reported in Table 3.5. For all reference compounds, wTv similarity calculations were carried out under systematic variation of α and β, as described above, and set-specific $HR(\alpha, \beta)$ values were calculated for the top scoring 100 database molecules, as shown in Figure 3.11. The calculations were repeated applying Tanimoto similarity and the comparison of results is shown in Table 3.6.

reference set	**TKI**	**TNF**
low complexity	18.7	19.1
medium complexity	25.2	25.5
high complexity	39.2	34.4

Table 3.5: Bit densities of TKI and TNF subsets. Reported are the average MACCS bit densities (in %) of reference sets of class TKI and TNF used in calculations of Section 3.3.

		hit rate (%)		**bit density (%)**	
reference set		**Tc**	**wTv**	**Tc hits**	**wTv hits**
TKI	low	23	36	24.5	24.7
	medium	28	30	25.1	24.8
	high	0	1	-	25.3
TNF	low	20	19	25.6	25.7
	medium	8	21	27.4	25.4
	high	0	3	-	25.7

Table 3.6: Hit rates of wTv and Tc. Best hit rates for selection of the top 100 database molecules are reported for TKI and TNF search calculations when potential hits closely match the MACCS bit density of the background database (25.7%). For each activity class, three sets of reference compounds with increasing bit density are used, as reported in Table 3.5. "bit density Tc/wTv hits" stands for average bit density of hits identified on the basis of Tanimoto or weighted Tversky similarity.

As shown in Table 3.6, wTv calculations produced overall better hit rates than control calculations using standard Tanimoto similarity. Figure 3.11 reveals trends similar to those seen in Figure 3.9 and Figure 3.10. Best hit rates were apparent after modulating complexity effects through variation of α

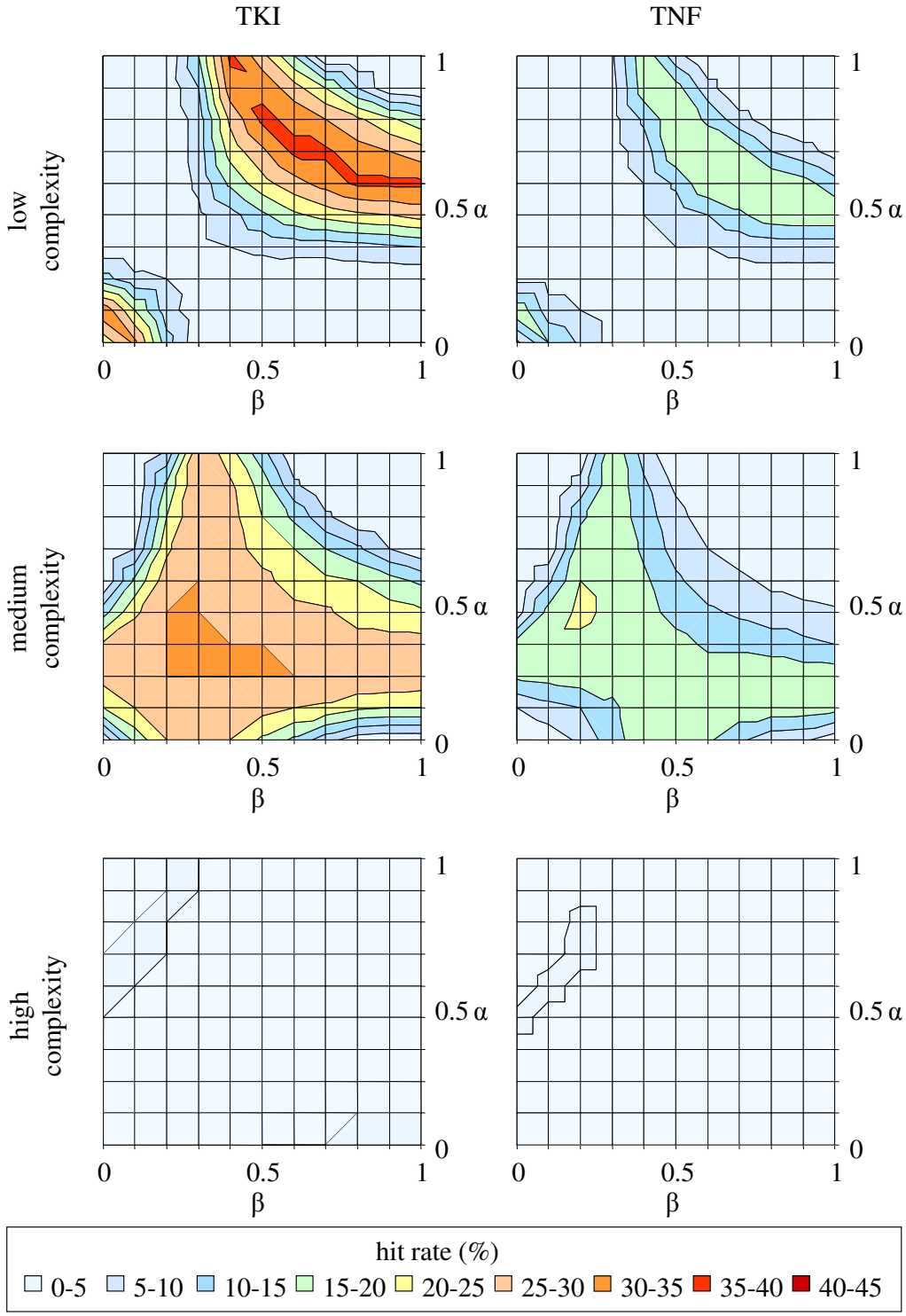


Figure 3.11: Virtual screening using different reference sets. Reported are similarity search results for low complexity (top), medium complexity (middle) and high complexity (bottom) reference subsets of class TKI (left) and TNF (right). The potential hits had comparable average bit density to the database. Hit rates from top 100 molecules are illustrated under systematic variation of the α and β parameters in increments of 0.1.

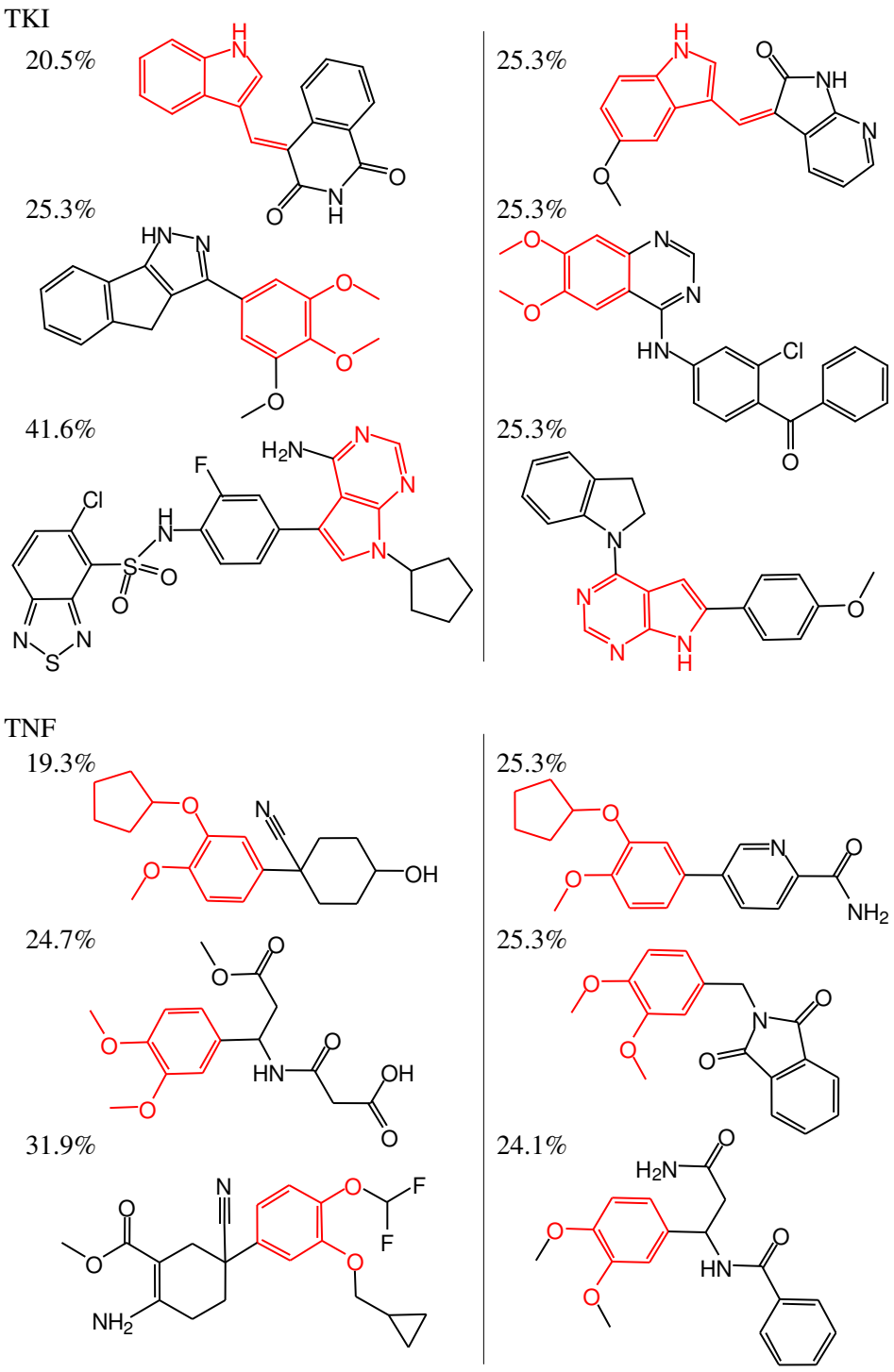


Figure 3.12: Structures of templates and hits. Examples of TKI (top) and TNF (bottom) reference compounds of varying complexity (left) are shown together with hits identified using these compounds (right) in the calculations summarized in Figure 3.11. Their bit densities are reported and substructures shared by corresponding reference compounds and hits are colored red.

and β parameters. Furthermore, these results make it possible to distinguish between three search situations. Calculations with reference compounds having lower complexity than the database are less relevant for virtual screening than the other two cases. Here combinations of high α and high β values were preferred, as discussed above. By contrast, when the complexity of reference compounds, database molecules, and hits were comparable many (α, β) combinations produced top hit rates. However, for reference compounds of higher complexity than potential hits or the background database, which is highly relevant for virtual screening, hit rates were much lower. Despite these very low hit rates that made the evaluation of parameter combinations difficult, there was also a preference for high α and low β values, at least in the case of TNF. Clearly, the case where reference compounds had higher complexity than potential hits presented the most challenging search scenario (where evaluation of standard Tanimoto similarity failed). These findings are well in accord with principal expectations derived from the formula of wTv. Thus, the trends observed here should generally apply to wTv calculations and related similarity metrics. Figure 3.12 shows examples of reference compounds of varying bit density and corresponding hits. These figures also illustrates that the density of "1" fingerprint bits provides a meaningful measure of molecular complexity.

3.4 Summary

Fingerprint search performance is determined by intrinsic features of fingerprint descriptors, chosen search strategies, and the way fingerprint similarity is quantified. For conventional 2D fingerprints such as MACCS, bit density is usually much influenced by molecular size. This chapter has uncovered a direct relationship between fingerprint bit densities and asymmetry of Tversky similarity calculations and demonstrated that differences in bit densities determine preferred Tv parameter settings for similarity searching.

Application of the Tversky similarity measure makes it possible to calculate molecular fingerprint similarity in a symmetric and asymmetric fashion. For fingerprints having different complexity, mathematical analysis has been conducted to describe the characteristics of Tversky coefficient with regard to the weight put on reference compounds. Furthermore, similarity search results have confirmed such characteristics and explained the asymmetric behavior of Tv similarity calculations. Evaluation of Tv distributions has enabled the determination of optimal α values in similarity searching, which is dependent on different fingerprint bit densities of the reference classes.

In addition to the demonstration of complexity effects and their direct influence on Tv similarity searching, two possible approaches have been suggested to avoid complexity effects. First, for a fingerprint design with constant bit density such as PDR-FP, Tv calculations are always symmetric and independent of

α parameter settings. Therefore, development of complexity-independent fingerprints can circumvent search difficulties that occur when complex optimized lead structures are used to search for relatively simple non-optimized hits.

Second, the weighted Tversky coefficient (wTv) has been introduced, which is a versatile similarity metric taking the weight on "0" into consideration. With the wTv it is possible to study and balance complexity effects and differently weight contributions of reference and database molecules. The interplay between these parameters produces complex similarity value distributions that have been analyzed to study the influence of molecular complexity on fingerprint searching in detail. Balancing complexity effects leads to constant similarity values for reference and background database molecules, independent of how compound contributions are weighted. Under these conditions, no systematic errors occur in calculating the similarity of database molecules.

Moreover, taking differences in molecular complexity into account also provides opportunities to optimize the retrieval of active compounds. Accordingly, in fingerprint searching for active compounds having different complexity, modulating complexity effects, rather than eliminating them, and putting high weight on reference compounds led to best hit rates in the analysis. Hit rate landscape maps have revealed preferred parameter combinations for similarity searching and helped to better understand preferred characteristics of reference compounds, which has implications for virtual screening. In wTv calculations, highly complex molecules are, for principal reasons, much less suitable as references than active compounds having complexity comparable to the screening database. The findings reported herein provide the basis for further analyses of similarity metrics and aid in the design of sound fingerprint search protocols. For example, in *Chapter 5* an activity class-specific similarity metric will be discussed, which has been developed based on wTv to account for complexity effects.

Chapter 4

Random Reduction of Fingerprint Bit Density

In the previous chapter, apparent asymmetry in search calculations on large databases using the Tversky coefficient[51] was shown to be a direct consequence of differences in molecular complexity. Similarity search calculations using conventional fingerprints such as, for example, MACCS structural keys[16,54] and similarity metrics like the Tanimoto coefficient (Tc)[6] are sensitive to differences in complexity between reference compounds and database molecules, which correlate to differences in fingerprint bit density.

There are two typical scenarios for practical fingerprint search applications. First, one uses hits from screening data sets as reference compounds for additional virtual screening. These hits usually have complexity and size comparable to average database molecules (from which they were selected). Second, one selects known active compounds from the scientific or patent literature as references to search databases for novel hits, which is probably the most common search situation. Typically, these templates are chemically optimized and potent compounds that are larger and more complex than average database molecules and hits from which they originate. In the previous chapter, it has been discussed that that the more complex reference compounds are, the lower the search performance becomes.

In this chapter, complexity effects are further investigated with regard to bit density of fingerprints, rather than similarity metrics. It is shown that when the number of bits set on in the fingerprints of complex reference compounds is randomly reduced, search performance notably increases, although random bit density reduction – also termed *random bit silencing* – reduces the chemical information content of fingerprints and biases similarity evaluation. This at first glance unexpected finding is analyzed and a generally applicable strategy is suggested to improve the performance of search calculations using conventional fingerprints.

4.1 Bit silencing experiment

For a binary fingerprint, *bit silencing* of a "1" bit is to set this bit from "1" to "0". It differs from modification of the fingerprint through removal of individual bit positions because in bit silencing, the length of the fingerprint is kept constant. As a result, the bit density of the fingerprints is reduced. Yet the presence of the correnspondingfeature in the molecule is no longer encoded and the loss of information is expected to affect pair-wise similarity comparison. As can be observed from previous studies (see for example, Figure 2.4 in Section 2.4), the search difficulties induced by complexity effects are directly related to high bit densities of reference compounds. Thus, experiments that reduce the bit densities of reference sets through random bit silencing are designed to systematically evaluate the interplay of complexity effects and fingerprint information on similarity search performance.

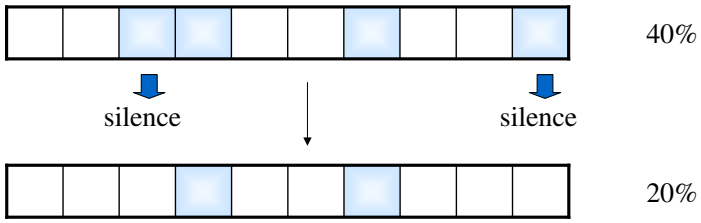


Figure 4.1: Bit silencing. Example of silencing the third and last bit position out of a hypothetical 10-bit fingerprint. As a result, the bit density reduces from 40% to 20%.

In order to generate sets of active compounds with systematically varying fingerprint bit density, five activity classes were initially assembled from the MDDR[38]: cyclooxygenase inhibitors (COX), leukotriene antagonists (LKT), phospholipase A2 inhibitors (PA2), reverse transcriptase inhibitors (RTI), and protein tyrosine kinase inhibitors (TKI). It was critically important to obtain subsets of each activity class with fingerprint bit densities similar to or larger than background database molecules, which limited the initial choice of MDDR activity classes. As background database (termed BGDB) for similarity searching, 5,000 molecules were randomly selected from ZINC.[39]

First the average bit density of the MACCS fingerprint[16] for BGDB molecules was calculated to be 22.3% (of 166 MACCS bits). Then, from each activity class, 100 compounds with comparable average bit density (22.3 - 22.7%, depending on the class) were extracted as active database compounds (ADC), to be added to BGDB as potential hits. In addition, as reference sets for similarity searching, for each activity class four subsets were assembled (termed reference sets RS1 - RS4) with 20 compounds each of systematically increasing average MACCS bit density per set of approximately 22%, 29%, 33%, and 39%.

class	ADC	RS1	RS2	RS3	RS4
COX	22.5	22.3	28.5	30.3	39.3
LKT	22.6	22.4	28.7	33.5	38.6
PA2	22.3	21.9	-	34.1	39.2
RTI	22.7	23.2	28.4	32.7	38.7
TKI	22.3	21.7	28.8	32.4	41.3

Table 4.1: Bit densities of active database compounds and reference sets. Reported are average MACCS bit densities (in %) calculated for active database compounds ("ADC") and four different reference sets ("RS1" - "RS4"). ADC and RS1 were selected to have bit densities comparable to BGDB (22.3%). Reference sets RS2, RS3, and RS4 were designed to contain molecules of increasing bit densities. For activity class PA2, no reference set with an average bit density of 29% could be identified and, therefore, RS2 was not available in this case.

For RS1, these 20 compounds had to be extracted from the ADC sets, because not sufficient additional active molecules were available at this bit density level. Therefore, for similarity searching using RS1, only 80 instead of 100 ADC were available. Table 4.1 summarizes the different ADC and reference sets and their bit densities. The design of these compound sets has enabled the evaluation of the influence of increasingly complex search templates on fingerprint similarity searching and also provided a basis for set-directed modification of fingerprint bit settings.

Three different types of fingerprint search calculations were carried out using MACCS. First, for each activity class, reference sets RS1-RS4 were separately used to search for ADC and hit rates were calculated for the 100 top-scoring database molecules.

Second, fingerprints of reduced bit density were generated for reference compounds, while fingerprints of ADC and BGDB compounds remained unmodified. To decrease the average bit densities of a reference set by 5%, 10%, 15%, etc., fingerprint bit positions were randomly selected and set to "0" in all compounds of this reference set until the desired bit density level were achieved. For some compounds in the set these positions were set to "0" before silencing and they remained to be "0" in the process. For RS1, three reduction levels were generated, for RS2 four, for RS3 five, and for RS4 six. At each reduction level, similarity search calculations on all compound sets were performed with ten different versions of randomly silenced fingerprints. In each case, hit rates were determined for the top-scoring 100 database molecules and the results were averaged.

Third, MACCS fingerprints with randomly reduced bit densities were created for all compounds, i.e. reference, ADC, and BGDB molecules. Bit positions were randomly chosen and set to "0" in all compounds until average bit densities were reduced by 5% or 10%. Larger reductions (e.g., 15%) were not meaningful because the BGDB average bit density was only 22.3%. Then

search calculations were carried out for ten different random fingerprint versions at each reduction level and hit rates were calculated and averaged as described above.

All search trials using unmodified and bit density-reduced fingerprint versions were conducted using a 20-nearest neighbor approach (20-NN)[11] and Tanimoto similarity was calculated. That is, the pair-wise Tc similarity of a database molecule was determined against each of the 20 reference compounds and the average of these individual Tc values was used as final similarity score. The 20-NN strategy was chosen here in order to equally weight contributions of the fingerprints of all reference compounds. For comparison, a number of test calculations were also carried out using a 1-NN search technique, i.e. using only the highest similarity value. Control calculations at different bit density reduction levels were carried out with TGD and TGT that are 2D two- and three-point pharmacophore-type fingerprints, respectively (see Table 1.1).[19] Like MACCS, these fingerprints are keyed, i.e. each bit is associated with a defined feature, but they monitor atom pair (TGD) or three-point pharmacophore patterns (TGT) and are larger than MACCS (with 420 and 1704 bit positions, respectively).

4.2 Random bit silencing of reference sets

The results of standard MACCS calculations applying the 20-NN ranking scheme are reported in Table 4.2. Given the set-up of the test calculations, the probability of identifying an active compound by random selection was ¡2%. As can be seen, hit rates were strongly dependent on the bit density of reference compounds, irrespective of the activity class. When searching with reference set RS1 (having about 22% bit density), hit rates of 32–45% were achieved for activity classes LKT, PA2, RTI and TKI. Only for COX, a hit rate of ¡20% was obtained. For reference set RS2 (29% bit density), search performance notably decreased and top hit rates were only 26% (for classes LKT and TKI). For reference sets RS3 and RS4 (with 33% and 39% bit density), hit rates were further reduced to between 0% and 12%. In the case of RS4, the most complex reference compounds with bit densities ¿38%, similarity search calculations failed for all classes but PA2 (producing a low hit rate of 6%). No single active molecule was recovered among the top 100 database molecules for classes COX, LKT, RTI and TKI. These results clearly illustrate the consequences of using complex reference compounds in fingerprint searching and the correlation between bit densities and search performance. The more complex the reference compounds are, the lower the compound recall becomes. Moreover, search calculations that produce reasonable hit rates for reference compounds with bit density comparable to database molecules (RS1) essentially fail when reference compounds with high bit density are used (RS4). On the basis of these obser-

vations, reducing the bit density in fingerprints of reference compounds can be expected to balance complexity effects and increase search performance. However, setting "1" bits to "0" also reduces the chemical information content of fingerprint representations, making the net effect of such modifications difficult to predict. Thus, it is necessary to systematically study the consequences of bit density reduction in fingerprints of reference compounds.

class	RS1	RS2	RS3	RS4
COX	17	13	1	0
LKT	45	26	4	0
PA2	39	-	12	6
RTI	32	6	0	0
TKI	42	26	3	0
average	35	18	4	1

Table 4.2: Search performance using unmodified MACCS fingerprints. For reference sets of increasing bit densities ("RS1" to "RS4"), hit rates (in %) are reported for selections sets of 100 compounds. In similarity searches using RS1 as templates, 80 potential database hits were available and for RS2, RS3 and RS4, 100 potential hits.

Table 4.3 summarized the results of randomly silencing the reference set only. For RS1-RS4, the bit density of their MACCS fingerprints was randomly reduced in a step-wise manner down to a level of 7–8% and at each reduction level, fingerprint modification was performed ten times to avoid chance effects. Then systematic search calculations against unsilenced fingerprints of database molecules were carried out.

When searching with reference set RS1, step-wise bit density reduction led to consistently lower hit rates over the three reduction levels; starting from, on average, 35% original hit rate to 30%, 19%, and 9%. This gradual decrease in hit rates can be attributed to the loss in fingerprint information content considering that RS1 and database molecules have comparable bit density. Thus, complexity effects are negligible in this case and silenced fingerprint representations lead to lower search performance, as one would expect. By contrast, for reference sets RS2–RS4 having higher bit densities than ADC and BGDB molecules, bit density reduction systematically improved search performance. For RS2, optimal hit rates were reached at the 12–13% bit density level for activity class COX (16%) and at the 17–18% bit density level for classes LKT and RTI (30% and 11%). Here class TKI was an exception because bit density reduction did not increase hit rates. For RS3, bit density reduction led to an in part significant improvement in hit rates taking into account that the original hit rates were overall low for these complex reference compounds. At the 12–13% bit density level, hit rates of 9% instead of 1% were observed for activity class COX, 6% instead of 0% for class RTI, and 13% instead of 3% for TKI. Thus, in contrast to RS2, in this case, bit density reduction for the more com-

reference set		bit density level 7-8%	12-13%	17-18%	22-23%	27-29%	30-34%	39-41%
RS1	COX	6	13	**17**	**17**			
	LKT	12	20	40	**45**			
	PA2	8	19	33	**39**			
	RTI	5	17	27	**32**			
	TKI	12	25	32	**42**			
	average	9	19	30	**35**			
RS2	COX	13	**16**	15	**16**	13		
	LKT	14	19	**30**	29	26		
	PA2	-						
	RTI	5	8	**11**	**11**	6		
	TKI	5	14	23	**26**	**26**		
	average	9	14	**20**	**20**	18		
RS3	COX	2	**9**	8	5	2	1	
	LKT	**15**	7	10	14	8	4	
	PA2	4	9	9	8	8	**12**	
	RTI	2	**6**	**6**	4	1	0	
	TKI	4	**13**	8	4	3	3	
	average	5	**9**	8	7	4	4	
RS4	COX	4	4	**5**	3	4	2	0
	LKT	7	5	**8**	5	7	1	0
	PA2	4	6	7	7	**9**	6	6
	RTI	**2**	**2**	1	1	0	0	0
	TKI	**9**	**9**	4	2	1	0	0
	average	**5**	**5**	**5**	4	4	2	1

Table 4.3: Search performance using randomly silenced reference sets. Hit rates (in %) are listed for reference sets of increasing bit densities and selection sets of 100 compounds. In each block (RS1, RS2, RS3 or RS4), hit rates in the rightmost column indicate that original instead of silenced fingerprints of reference compounds are used as search templates; bold hit rates indicate the best performance within each row. Numbers in column titles show the actual bit density of template fingerprints. In all calculations, bit strings of database compounds (and ADC hidden among them) remained unmodified.

plex TKI molecules also led to an increase in hit rates. Furthermore, for class LKT, the hit rate increased from 4% to 14% at the 22–23% bit density level. Finally, when searching with reference set RS4, random bit density reduction led to the correct detection of several hits for each activity class, whereas the original search calculations with unmodified MACCS fingerprints completely failed in four of five cases (except PA2). For these classes, top hit rates under silencing conditions ranged from 2% (RTI) to 9% (PA2 and TKI).

Comparison of the preferred bit density levels showed that highest hit rates were obtained at different reduction levels, dependent on the class. However, a general trend was observed when average hit rates were monitored over all activity classes, as shown in Figure 4.2. The preferred bit density reduction level shifted towards lower bit densities with increasing original reference set

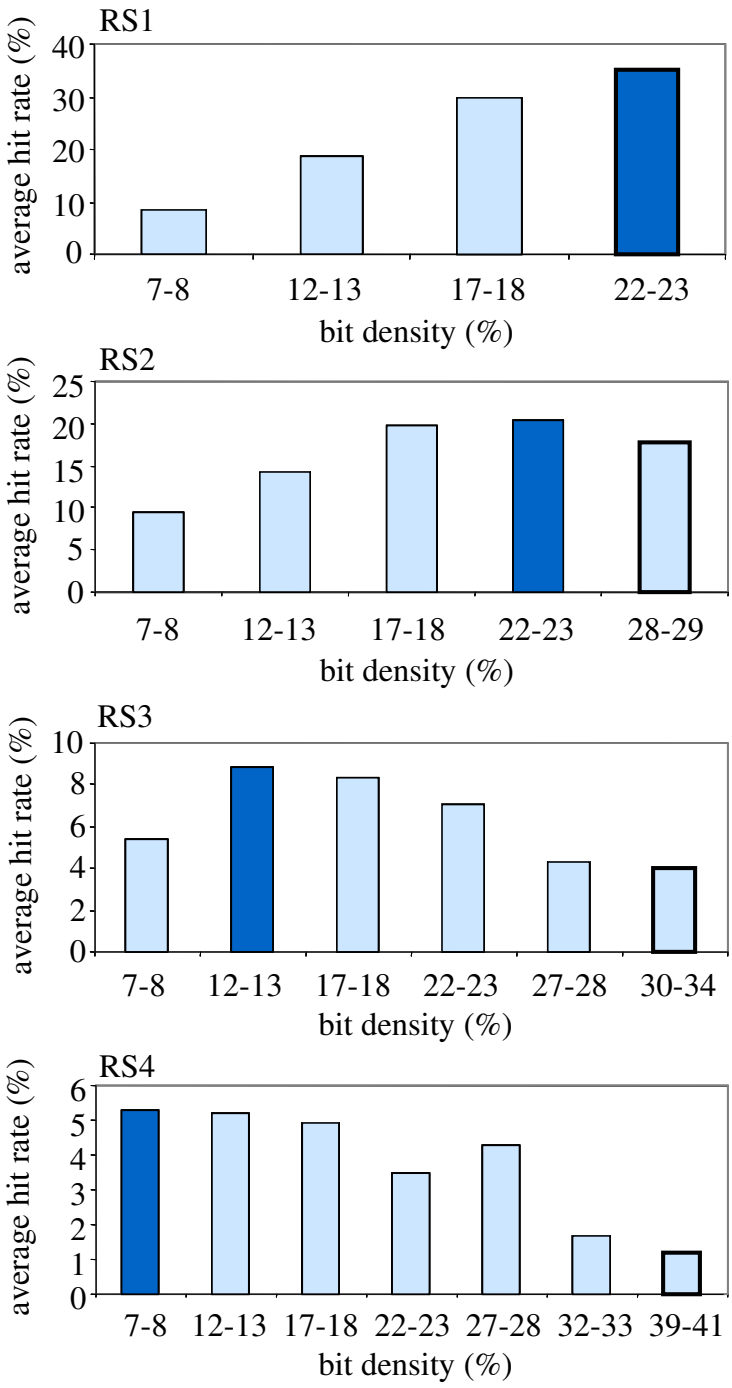


Figure 4.2: Hit rates after bit silencing of reference sets. Hit rates averaged over the ten independent trials of all five activity classes are reported using reference set RS1, RS2, RS3 and RS4. For each reference set, MACCS bit density was randomly reduced to different levels. Bars with bold borders are the hit rates for unmodified fingerprints used in similarity searching, while bars colored in dark blue are the optimal hit rates.

bit density, from RS1 to RS4. For reference set RS1 highest hit rates (on average 35%) were obtained for original bit densities because complexity effects were negligible here, as discussed above. For reference set RS2, the bit density levels 17–18% and 22–23% led to highest average hit rates (with an average of about 20%); for RS3, preferred levels were 12–13% and 17–18% (with average hit rates of 8–9%), and for RS4, best hit rates were obtained at 7–8% and 12–13% bit density levels. Thus, the higher the original bit density of a reference set was, the more its bit density had to be reduced to optimize compound recall. Furthermore, preferred bit density levels were often lower than the average BGDB fingerprint bit density. Because of complexity effects, a given reference compound does not preferentially recover database molecules of comparable bit density, but rather molecules with higher bit density. By contrast, when the reference compound has a lower bit density than the database molecules, bit density differences between database molecules no longer play a significant role. However, the average BGDB bit density level of approximately 22–23% still provided an attractive search level, as shown in Figure 4.2.

reference set		**bit density level**					
		7-8%	**12-13%**	**17-18%**	**22-23%**	**27-29%**	**30-34%**
COX	20-NN	2	**9**	8	5	2	1
	1-NN	12	**25**	23	24	**25**	20
LKT	20-NN	**15**	7	10	14	8	4
	1-NN	23	20	30	**40**	30	19
PA2	20-NN	4	9	9	8	8	**12**
	1-NN	18	**26**	17	15	13	**11**
RTI	20-NN	2	**6**	**6**	4	1	0
	1-NN	6	19	16	**26**	18	10
TKI	20-NN	4	**13**	8	4	3	3
	1-NN	5	**28**	16	12	20	13

Table 4.4: Comparison of 20-NN and 1-NN as rules of data fusion using randomly silenced reference sets. Similarity calculations as reported in Table 4.3 were carried out with 1-NN rules of data fusion. For each database molecule, the highest Tc value from pair-wise comparison with the compounds in reference set RS3 were retained for ranking. Then the highest hit rates over the multiple trials were recorded (labeled "1-NN"). They were compared with the corresponding RS3 data in Table 4.3 (labeled "20-NN"), which was calculated according to the 20-NN or averaging rule of data fusion.

In addition, using the 1-NN search strategy, which usually improves similarity search performance[11,42,43] instead of 20-NN, random silencing of reference set yielded improved performance as well. As shown in Table 4.4, when reference set RS3, which was more complex than BGDB, was used as template, bit density reduction produced in general higher hit rates with 1-NN similarity calculations.

Further calculations were carried out on two activity classes using the TGD and TGT fingerprints[19] (see Table 1.1) instead of MACCS. Detailed data

are shown in Table B.1 and B.2. TGD and TGT displayed trends similar to MACCS when bit densities were reduced. Thus, the effects discussed above were not MACCS-dependent, but generally applies to key-type fingerprints.

4.3 Random bit silencing of all fingerprints

In this section, the bit density in both reference and database molecules was randomly reduced such that relative differences in bit densities remained approximately the same. These modifications generally reduce fingerprint information content but maintain complexity relationships. The results of systematic similarity search calculations using these reduced fingerprint representations are summarized in Table 4.5.

reference set		**10% bit density reduction**	**5% bit density reduction**	**original**
RS1	COX	15	**18**	17
	LKT	40	**46**	45
	PA2	33	37	**39**
	RTI	23	28	**32**
	TKI	32	41	**42**
	average	29	34	**35**
RS2	COX	8	**13**	**13**
	LKT	**29**	27	26
	PA2	-		
	RTI	21	**23**	**26**
	TKI	8	13	26
	average	17	**18**	**18**
RS3	COX	**2**	1	1
	LKT	3	**4**	**4**
	PA2	9	11	**12**
	RTI	0	0	0
	TKI	**3**	2	**3**
	average	3	**4**	**4**
RS4	COX	0	0	0
	LKT	0	0	0
	PA2	**6**	**6**	**6**
	RTI	0	0	0
	TKI	0	0	0
	average	**1**	**1**	**1**

Table 4.5: Search performance after random bit silencing of all fingerprints. Hit rates (in %) are listed for reference sets of increasing bit densities and selection sets of 100 compounds. At each reduction level (5% or 10%) bit densities of reference and database compounds were simultaneously reduced. "original" refers to unsilenced fingerprints.

For reference set RS1, where average bit densities of reference and

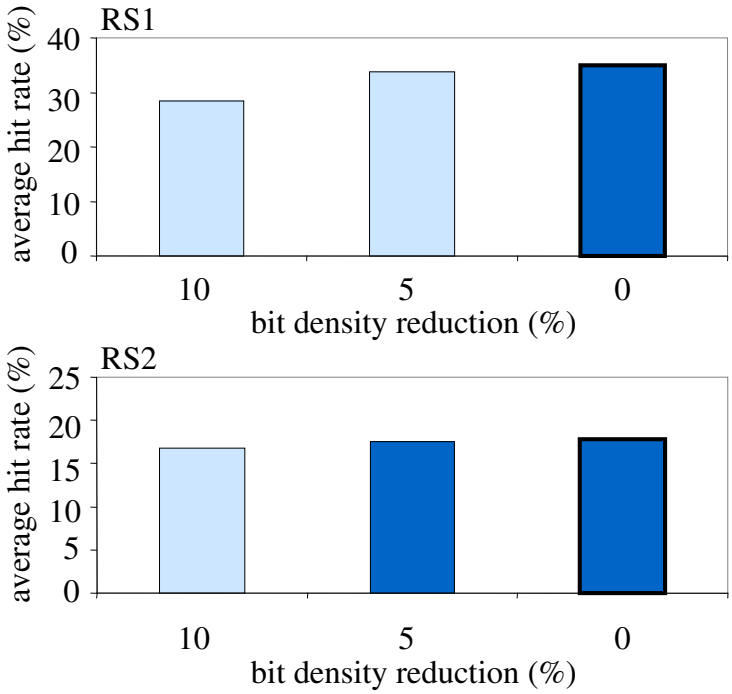


Figure 4.3: Hit rates after random bit silencing of all sets. Hit rates averaged over the ten independent trials of all five activity classes are reported using reference set RS1 and RS2. For each reference set, MACCS bit density of reference and database molecules was randomly reduced at the same time to different levels. Bars with bold borders are the hit rates for unmodified fingerprints used in similarity searching, while bars colored in dark blue are the optimal hit rates. For RS3 and RS4 similar trend was observed (Figure B.1).

database molecules were comparable, bit density reduction led to a consistent decrease in hit rates over the different reduction levels, on average from 35% to 25%. These observations were consistent with the notion that fingerprints with reduced information content lose predictive capacity. For reference set RS2, a decrease in hit rates was only observed for two classes (COX and TKI), whereas hit rates increased for RTI and remained essentially constant for LKT. Thus, RS2 calculations showed that a loss in fingerprint information content led to unpredictable results in the presence of complexity effects. For RS3 and RS4, original hit rates were in part very low and no significant changes were observed. Taken together, these results show that universal bit density reduction decreases fingerprint search performance. By contrast, bit silencing only on reference compounds balances complexity effects and improves compound recall, as discussed in the previous section.

4.4 Summary

In this chapter, an alternative approach to balance complexity effects through random bit silencing has been introduced and tested. Systematic similarity searching using compound reference sets of variable but controlled fingerprint bit density show that the more complex reference compounds are, the lower the recall of active compounds with average complexity becomes. Through random reduction of fingerprint bit density of reference compounds complexity effects can be balanced for standard fingerprints.

The fingerprint bit silencing causes two opposing effects: a general loss of chemical information leading to a decrease in search performance and compensation of complexity effects leading to higher hit rates. Similarity search results show that balancing molecular complexity effects outweighs the information loss associated with bit density reductions and leads to in part significant increases in the recall of active compounds, especially when the reference compounds are much more complex than the database molecules. Importantly, bit positions can be randomly selected and silenced in order to achieve a net increase in hit rates. Without computational analysis, it could not have been predicted that random bit silencing leads to an increase in search performance when reference compounds of above average complexity are used.

These findings suggest that random bit silencing can be applied as a search strategy. Because it is straightforward to calculate and compare average bit densities, one can easily detect whether available reference compounds have higher bit density than database molecules. If so, it is possible to carry out search calculations after random reduction of reference fingerprint bit density to the level of database molecules or below, where complexity effects become negligible. Under these conditions, search calculations using standard fingerprints should have an increased probability of identifying novel hits.

Chapter 5

Bit Position-Weighted Similarity Metrics

In the previous chapter it has been shown that random bit silencing of fingerprints of complex reference compounds enhances search performance. However, this unsupervised process does not depend on whether the silenced/remaining bit positions are critical for the identification of active compounds or not. There is no preference with regard to which bit position to silence. The contribution of individual bit positions to similarity search performance has not yet been systematically analyzed. One possible strategy to address this question is to perform bit silencing in a controlled manner.

In this chapter, bit silencing is utilized as an approach to systematically determine the contribution of each bit position to similarity search performance. For a given fingerprint and compound activity class, bit silencing makes it possible to derive a bit position-dependent weighting scheme that can then be used to modify similarity metrics in a compound class-specific manner. As a result, a bit position-dependent weighted variant of the Tanimoto coefficient, bwTc, is designed, which is found to increase hit rates of conventional search calculations.

Complexity differences between reference compounds and database molecules often systematically affect the result of similarity searching. For Tversky similarity calculations, such biasing effects could be corrected by introducing the weighted Tversky coefficient (wTv, as discussed in *Chapter 3*), which made it possible to set relative weights on "1" and "0" bits and thereby balance complexity differences between reference and database molecules. However, fingerprint searching with chemically optimized reference compounds that were more complex than average database molecules generally made it most difficult to identify novel hits.

Therefore, in this chapter another similarity metric will also be introduced that simultaneously balances complexity effects and emphasizes com-

pound class-specific bit settings during fingerprint searching. This class-directed similarity coefficient is generated by combining the wTv and bwTc functions. The resulting "weighted Tversky coefficient with class-specific bit weighting", or wbwTv, represents a parametric approach of modulating similarity and complexity. In systematic search calculations utilizing compound reference sets of increasing complexity, wbwTv outperformed its parental methods and other similarity metrics.

5.1 Systematic bit silencing and generation of a bit weight vector

The derivation of bit position-weighted similarity metrics consists of two stages: the training stage and the test stage. In the training stage, each individual bit position in a keyed fingerprint is systematically set to "0" for all reference compounds prior to similarity searching, as described in *Chapter 4*. For a fingerprint with N bits, a total of N search calculations (training searches) are carried out with variable settings on $(N-1)$ bits, except for the silenced bit that is constantly set to "0" and does not contribute to the search.

In this study MACCS keys[16] with 166 bits have been subjected to the bit silencing procedure. Hit rates were calculated for 166 silencing calculations and recorded in a bit position-dependent hit rate profile. From the hit rate profile, a *bit position-dependent weight vector* is calculated on the basis of weights that are assigned to each bit position according to the effects of silencing. If silencing of a bit position leads to a reduction in search performance, the bit makes a positive contribution and is emphasized. By contrast, if silencing of a bit increases search performance, it negatively contributes and is de-emphasized. If silencing has no effect, the bit makes no contribution and is not weighted. Accordingly, the weight vector can be derived as follows: if hr_O is the hit rate obtained with the unmodified fingerprint and $(hr_1, hr_2, \ldots, hr_N)$ are N hit rate values that correnspond to the similarity search with each of the N bits in the fingerprint silenced individually, the weight on the i-th bit, w_i, is defined as

$$w_i = (1 + (hr_O - hr_i) \cdot sf) \cdot 100\% \tag{5.1}$$

where sf is a pre-defined scale factor reflecting the magnitude of change observed in the hit rate profile. The higher sf is, the more sensitive the weight vector becomes to fluctuation in hit rates as a consequence of silencing. For example, if sf is set to 100 and silencing of the i-th bit reduces the hit rate by 3%, then $w_i = (1 + (3\%) \cdot 100) \cdot 100\% = 400\%$, which means that the corresponding bit is scaled four-fold relative to the original 100% weight because of its positive contribution. With $sf = 200$ and a 3% reduction in hit rate, the value of w_i becomes 700%. By contrast, if silencing of a bit leads to a 2%

increase in hit rate and $sf = 200$, then the weight on this bit position becomes -300%, which corresponds to three-fold negative scaling.

The bit position-dependent weight vector $\boldsymbol{W}$ consists of the weights of all N bit positions ($\boldsymbol{W} = (w_1, w_2, \ldots, w_N)$) and mirrors the significance of each individual bit. The calculation of $\boldsymbol{W}$ is fingerprint- and compound class-dependent and influenced by the composition of the reference set. For example, for class COX (cyclooxygenase inhibitor) assembled from MDDR[38] and a background database consisting of 5,000 molecules randomly extracted from ZINC[39], the hit rate profile and the derived weight vector are shown in Figure 5.1. A subset of COX consisting of 102 compounds was taken as training set and from this set, a reference subset of 20 compounds was randomly selected and the remaining compounds were added to the background molecules for deriving the bit silencing hit rate profile. 166 bit silencing calculations were carried out in combination with 20-NN ranking (to equally take contributions of all reference molecules into account) and hit rates were calculated for the top-ranked 100 database molecules. In this example, MACCS Tc calculations produced a hit rate of 23%. Individual silencing of 17 of 166 bits reduced this hit rate by 1% to 4%, whereas silencing each of 55 other bits resulted in higher hit rates between 24% and 35%. Thus, silencing of individual bits led to increases in hit rate of up to 12%, which represents a significant improvement of search performance. In this case, silencing of the remaining 94 bit positions did not change the hit rate. Many of these were "0" bits. These findings illustrate that individual "1" bits can significantly compromise the ability to detect active compounds, and that only subsets of fingerprint bits determine search performance. For COX, nearly one third of MACCS bit positions did not detectably contribute to search performance.

To extensively test bit silencing and systematic similarity search calculations, 20 more activity classes were assembled from the MDDR (Table 5.1). The same ZINC subset was used as background database. For each activity class, a training set was assembled as reported in Table 5.1. The number of training compounds ranged from 84-605 for different classes. From each training set, a reference subset of 20 compounds was randomly selected and the remaining compounds were added to the background molecules, as in the COX case described above. Training of weight vector was repeated ten times with ten different reference subsets to avoid random bias and the activity class-dependent weight vector was derived by averaging these ten vectors. Weight vectors of all activity classes are compared in Figure 5.3. In this heat map it is shown that these weight vectors significantly differ in bit position weights and are thus class-specific. It is therefore not possible to select MACCS bit positions that are generally associated with different biological activities. However, bit silencing allows to derive bit weight vectors specific to the corresponding class with information relevant to the identification of active compounds.

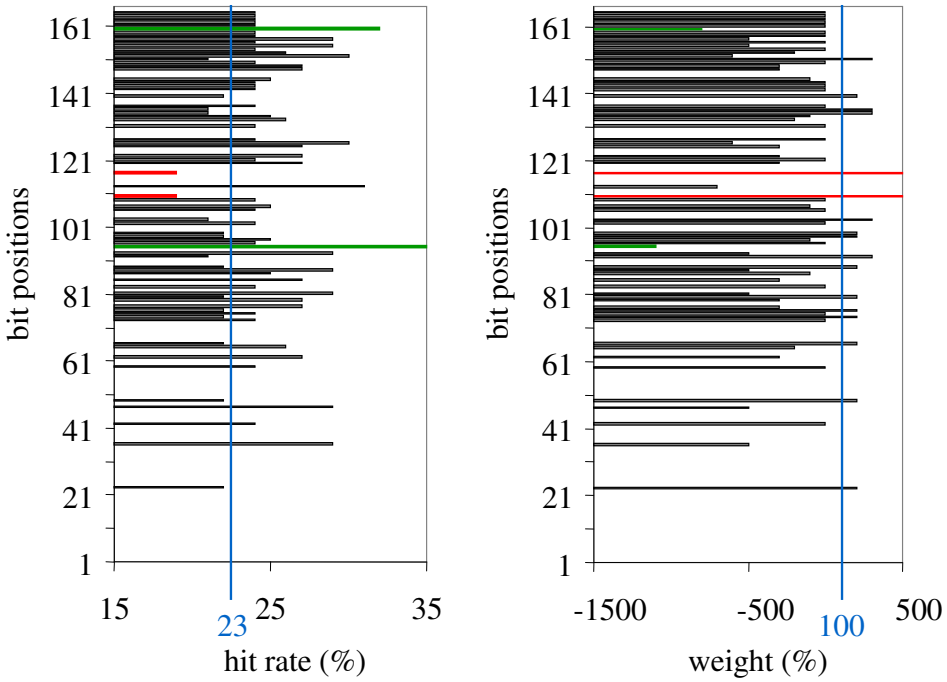


Figure 5.1: Bit silencing-derived hit rate profile. Shown is the hit rate profile of activity class COX derived from bit silencing (left) and bit position-dependent weight distribution generated using a scale factor of 100 (right). Weights of bit positions that increase or decrease the hit rate during silencing are displayed and bits whose silencing does not affect the hit rate of 23% (and thus obtained weights of 100%, shown as blue lines) omitted for clarity. Bit positions with maximum weight (positive scaling due to decrease in hit rate) and minimum weight (negative scaling due to increase in hit rate) are shown in red and green, respectively.

5.2 Bit position-weighted Tanimoto similarity

The weight vector discussed in the previous section makes it possible to generate a bit position-dependent weighted Tanimoto coefficient. Given two molecular bit vectors of length N, $\boldsymbol{A} = (a_1, a_2, \ldots, a_N)$ and $\boldsymbol{B} = (b_1, b_2, \ldots, b_N)$, the general form of Tc[6] is

$$Tc(\boldsymbol{A}, \boldsymbol{B}) = \frac{\sum_{i=1}^{N} a_i b_i}{\sum_{i=1}^{N} (a_i^2 + b_i^2 - a_i b_i)} \tag{5.2}$$

In this formulation, a_i and b_i are binary variables representing the i-th bit in fingerprint $\boldsymbol{A}$ and $\boldsymbol{B}$, respectivly, and $a_i b_i$ their product. Variable weights to each individual bit position can be added corresponding to the results of silencing by calculating the product of the Tc and weight vector $\boldsymbol{W}$. Thus, given a vector of N elements, $\boldsymbol{W} = (w_1, w_2, \ldots, w_N)$, representing the weights

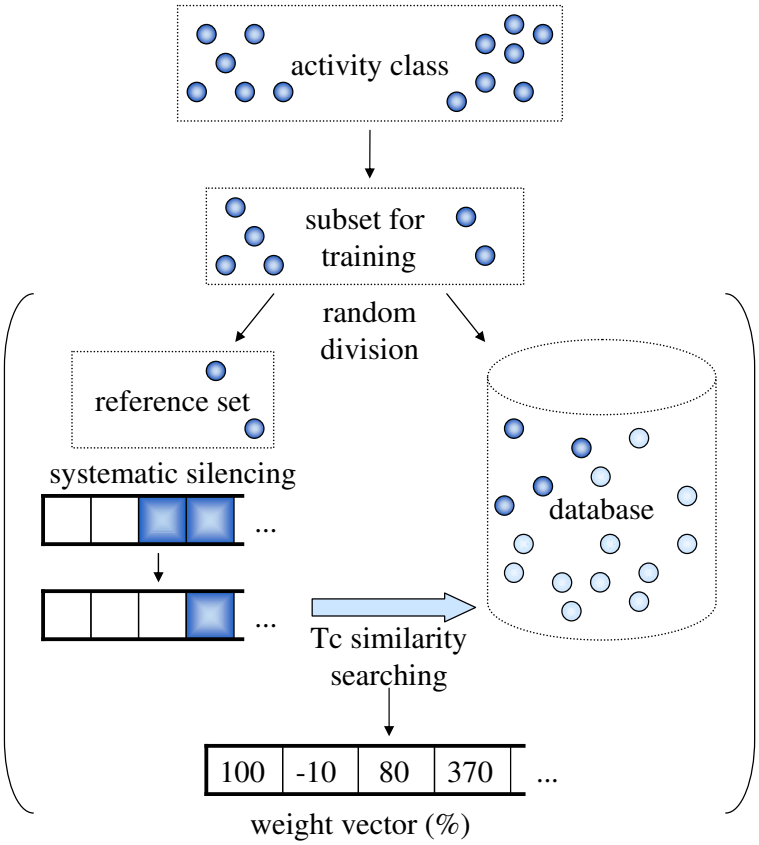


Figure 5.2: Training of bit weight vector. For a given activity class (such as COX), a training subset is assembled. From this subset a reference set is randomly extracted and the remaining compounds are hidden in the background database. In this calculation the reference set consists of 20 compounds. Systematic silencing is carried out on the reference set and similarity searching with Tc is applied to derive the bit weight vector. The training procedures described in Section 5.1 are summarized within the brackets.

on the N bits of the fingerprint, the bit position-dependent Tc, bwTc, is defined as

$$bwTc(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{W}) = \frac{\sum_{i=1}^{N} a_i b_i w_i}{\sum_{i=1}^{N} (a_i^2 + b_i^2 - a_i b_i) w_i} \tag{5.3}$$

The calculation of bwTc is illustrated in Figure 5.4. Two hypothetical fingerprints with ten bits are compared using the conventional Tc and bwTc. For the latter a hypothetical weight vector represented in percentage format is used. Because negative values are permitted for the weight vector's elements, as discussed above, bwTc similarity values can also become negative. Thus, compared to Tc-based ranking, larger value ranges and differences between similarity values are possible in bwTc calculations.

class	designation	number of training compounds	number of potential hits
ACE	angiotensin-converting enzyme inhibitor	215	30
ADR	aldose reductase inhibitor	250	70
CAM	cell adhesion molecule antagonist	133	10
CLG	collagenase inhibitor	146	20
COX2	cyclooxygenase-2 inhibitor	122	40
COX	cyclooxygenase inhibitor	102	140
ELA	elastase inhibitor	112	10
FXA	factor Xa inhibitor	605	40
HIV	HIV-1 protease inhibitor	148	50
LKT	leukotriene antagonist	181	120
LPO	lipid peroxidation inhibitor	138	70
MM1	muscarinic M1 agonist	178	20
NEP	neutral endopeptidase inhibitor	196	60
PA2	phospholipase A2 inhibitor	84	100
PAF	platelet-activating factor antagonist	198	50
PDV	phosphodiesterase V inhibitor	327	10
PKC	protein kinase C inhibitor	129	70
RTI	reverse transcriptase inhibitor	177	100
SST	squalene synthetase inhibitor	99	40
TKI	tyrosine-specific protein kinase inhibitor	253	250
TNF	tumor necrosis factor inhibitor	185	50

Table 5.1: Activity classes for bwTc similarity calculation. For 21 activity classes, "training compounds" were used in bit silencing calculations and the derivation of the class-specific bit position-dependent weight vectors and "potential hits" for similarity searching using MACCS Tc and bwTc calculations. Training and potential hit sets were distinct in each case.

Because the different effects of bit silencing described above were consistently observed for all 21 activity classes, the derivation of class-directed bit position-dependent similarity metrics is expected to be a promising approach of general relevance. Therefore, the derived class-specific weight vectors have been used to systematically compare bwTc calculations with standard MACCS Tc similarity searching and MACCS bit scaling calculations. A separate test set of active database compounds (ADC) was extracted from MDDR for each of the 21 activity classes. The number of these potential hits ranged from 10-250. ADC sets for each activity class were added to the ZINC background database and search calculations were carried out as described above (Section 5.1) for bit silencing. The reference compounds for these search calculations were taken from the training sets, as shown in Figure 5.5. In each case, hit and compound recovery rates were determined for the top-ranked 100 database compounds. Figure 5.6 shows a graphical comparison of hit rates for Tc and bwTc calculations using a scale factor of 100. In Figure 5.7, bwTc control calculations using different scale factors (50, 100, 200) are reported. In comparison, fingerprint

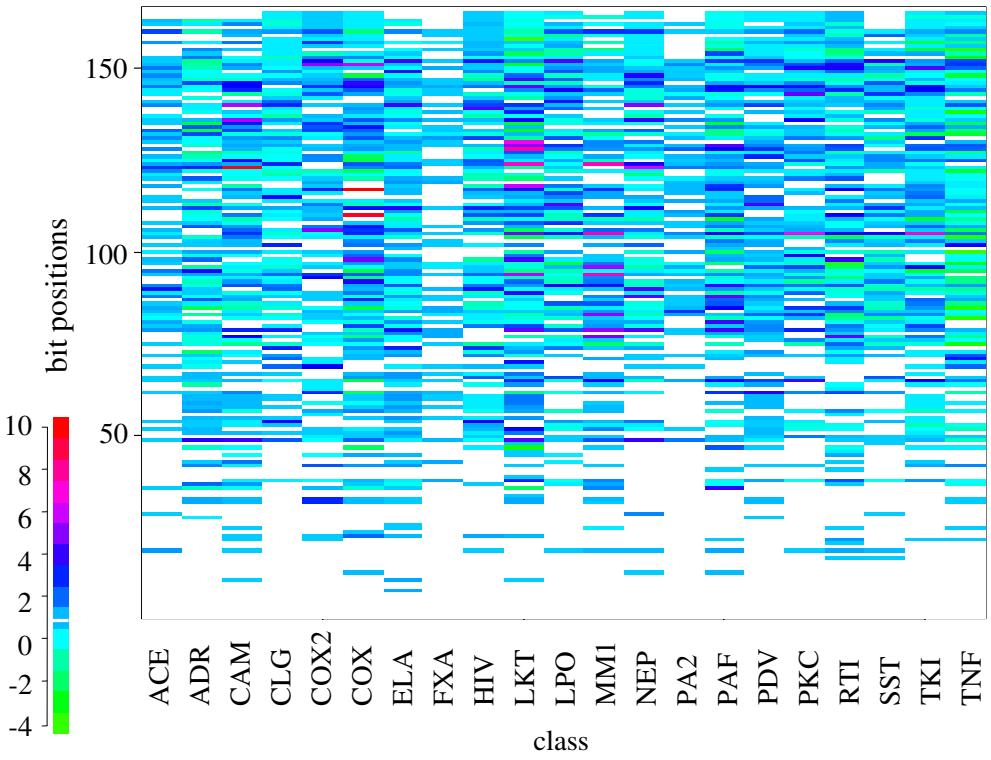
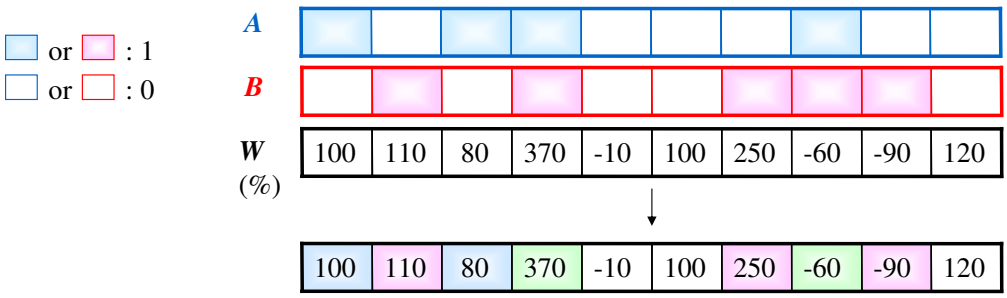


Figure 5.3: Heat map of bit weight vectors. Average bit weight vectors of the 21 activity classes are represented as a heat map. Bit positions with unmodified weight (100%, or 1) are omitted for clarity. The different color distributions show that the weights on bit positions are largely class-specific.

scaling[45] with a scaling factor of 3.0 to consensus bits was carried out as control calculation. Table 5.2 reports the hit and recovery rates for all test calculations.

The results in Table 5.2 and Figure 5.6 show that the application of bwTc generally increased hit and recovery rates of conventional MACCS Tc calculations. COX2 was the only of 21 classes for which Tc calculations produced higher rates. The average hit rate over all activity classes increased from 5% for Tc to 12% for bwTc calculations and the average recovery rate from 8% to 20%. For most classes, applying increasingly large scale factors for the generation of weight vectors did not substantially affect bwTc search results, as illustrated in Figure 5.7, i.e. a scale factor of 50 essentially produced results comparable to those obtained with scale factors of 100 or 200. Test calculations with scale factors of 400 and 800 were also carried out and generally reduced hit and recovery rates. The average hit rates of the 21 activity classes for $sf = 400$ and 800 were 9% and 7%, respectively, whereas for $sf = 100$ or 200 the average hit rates were 12%.

Depending on the activity class, the magnitude of hit rate improvements



$$Tc(\boldsymbol{A},\boldsymbol{B}) = \frac{\sum_{i=1}^{N} a_i b_i}{\sum_{i=1}^{N} (a_i^2 + b_i^2 - a_i b_i)} = \frac{2}{7} = 0.29$$

$$bwTc(\boldsymbol{A},\boldsymbol{B},\boldsymbol{W}) = \frac{\sum_{i=1}^{N} a_i b_i w_i}{\sum_{i=1}^{N} (a_i^2 + b_i^2 - a_i b_i) w_i} = \frac{3.7 - 0.6}{1 + 1.1 + 0.8 + 3.7 + 2.5 - 0.6 - 0.9} = 0.41$$

Figure 5.4: Calculation of the bit position-dependent weighted Tc. Two hypothetical fingerprints consisting of ten bits each are compared using Tc and bwTc. The latter value is calculated on the basis of a hypothetical weight vector. In this calculation, the numerator contains the sum of the weights over all "1" bits shared by A and B (colored in green) and the denominator the sum of the weights on the "1" bits in either A or B (blue, red or green). In this example, the two hypothetical molecules become more similar when bwTc is calculated because they share a bit position that makes a significant contribution to search performance, having a relative weight of 370%.

achieved in bwTc calculations differed. For eight classes, Tc calculations failed to identify active compounds, but in all of these cases, bwTc calculations correctly recognized active molecules and achieved hit rates of up to 20% and recovery rates of up to 40% (Table 5.2). For six of the classes where Tc calculations succeeded, bwTc hit rate improvements ranged from 5% and 10% and for six other classes improvements of more than 10% were observed. In some cases, these effects were very significant. For example, for LKT and TKI, Tc calculations produced hit rates of 5% or 6% hit rate, but bwTc calculations increased these rates to 40% or more (Figure 5.6). Because these compound sets were assembled to contain only inhibitors with unique core structures (see Section 2.1 for the general calculation protocol), increasing hit rates in bwTc calculations also suggest an increase in the potential of recognizing structurally diverse compounds. Taken together, these results indicate that compound class-directed evaluation of fingerprint similarity provides a promising alternative to conventional similarity search protocols.

Although scaling calculations were also found to increase recall of ac-

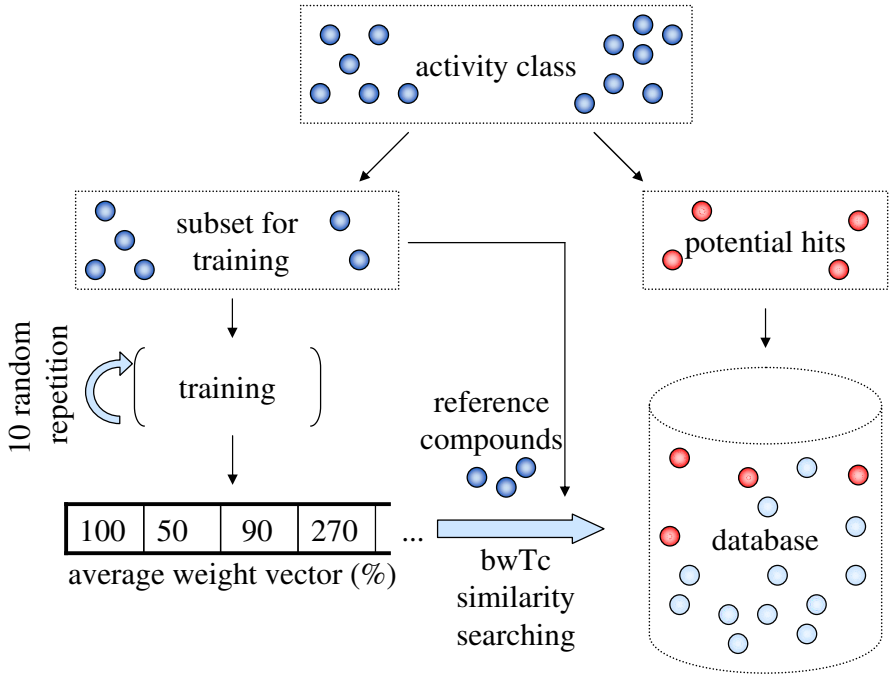


Figure 5.5: Evaluation of bwTc. The calculation protocol to systematically test bwTc is illustrated. For each activity class, the set of potential hits is independent of the training subset or the reference set. The weight vector used in bwTc similarity searching is the average result of ten independent random training experiments (shown in brackets, see Figure 5.2).

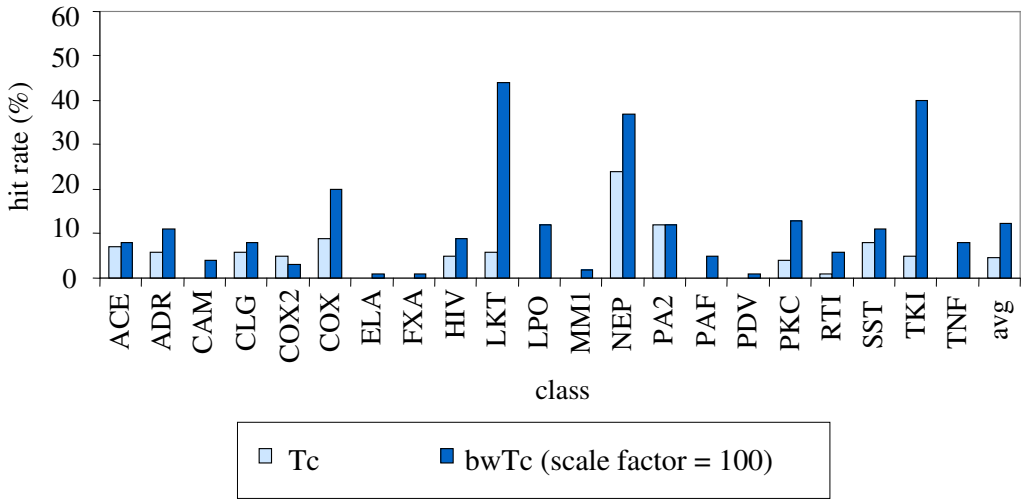


Figure 5.6: Hit rate comparison. Hit rates for 21 activity classes and the overall average ("avg") are reported for Tc (blue) and bwTc (dark blue). In bwTc calculations, a scale factor of 100 was applied.

class	Tc		bwTc $sf = 50$		bwTc $sf = 100$		bwTc $sf = 200$		FP scaling $sf = 3.0$	
	HR	RR	HR	RR	HR	RR	HR	RR	HR	RR
ACE	7	23	6	20	8	27	9	30	7	23
ADR	6	9	10	14	11	16	6	9	7	10
CAM	0	0	4	40	4	40	4	40	0	0
CLG	6	30	8	40	8	40	9	45	6	30
COX2	5	13	4	10	3	8	3	8	5	13
COX	9	6	21	15	50	14	15	11	11	8
ELA	0	0	1	10	1	10	2	20	0	0
FXA	0	0	0	0	1	3	2	5	0	0
HIV	5	10	9	18	9	18	9	18	6	12
LKT	6	5	34	28	44	37	39	33	6	5
LPO	0	0	6	9	12	17	20	29	0	0
MM1	0	0	2	10	2	10	0	0	0	0
NEP	24	40	39	65	37	62	34	57	24	40
PA2	12	12	12	12	12	12	12	12	12	12
PAF	0	0	3	6	5	10	4	8	0	0
PDV	0	0	1	10	1	10	1	10	0	0
PKC	4	6	15	21	13	19	10	14	4	6
RTI	1	1	4	4	6	6	11	11	1	1
SST	8	20	10	25	11	28	4	10	8	20
TKI	5	2	25	10	40	16	53	21	5	2
TNF	0	0	11	22	8	16	1	2	0	0
average	5	8	11	19	12	20	12	19	5	9

Table 5.2: bwTc similarity search results. Hit rates (HR) and recovery rates (RR) are reported (in %) for 21 activity classes using conventional Tc, bwTc, and fingerprint scaling ("FP scaling") calculations with different scale factors (sf).

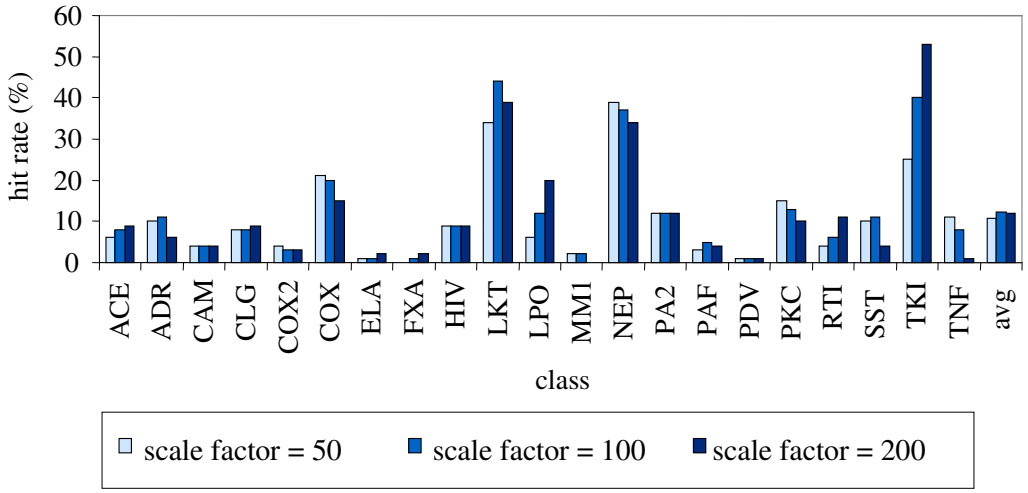


Figure 5.7: Different scale factors. Hit rates of bwTc calculations with scale factors of 50, 100, and 200 are reported and colored in light blue, blue and dark blue, respectively.

tive compounds using MACCS keys,[45] MACCS consensus bit positions for the activity classes studied here were not among the most significant bit positions for MACCS search performance. Thus, scaling of these bit positions does not emphasize the most critical bits for each activity class. The silencing method should have the principal advantage over consensus bit scaling that the most important bit positons are identified. This conclusion was confirmed by systematic bit scaling calculations using MACCS (Table 5.2).

Chemical interpretation of bit significance

With key-type fingerprints such as MACCS structural key, where bit positions can be directly mapped to substructural features, analysis of substructures corresponding to bits obtaining high or low weights in bwTc calculations makes it possible to interpret the results in a chemically intuitive manner. For example, as illustrated in Figure 5.8, substructures might be identified that are responsible for the detection of active compounds. COX inhibitors that were correctly identified using bwTc but not conventional Tc calculations are compared to ZINC compounds that were detected using Tc calculations but deselected by bwTc. A benzene moiety shared by all compounds is assigned a low bwTc weight. By contrast, two MACCS keys accounting for an "aliphatic six-membered ring containing a heteroatom" and a "N-X-O" unit detect an oxane substructure and an amide bond, respectively, that occur in the COX inhibitors but not in the ZINC compounds. These substructures were assigned high weights and help to distinguish the COX inhibitors from background database compounds.

Furthermore, in Figure 5.9 two substructural features corresponding to two top-weighted MACCS bit positions are highlighted on the structure of lisinopril. The schematic view of the structure is derived from the X-ray structure of the human angiotensin-converting enzyme–lisinopril complex.[55] This example shows the correnspondence of fingerprint bit significance as identified by bit silencing and the significance of substructures involved in interactions. Thus potential pharmacophoric groups might be selected on the basis of bit silencing and assigned high weights in similarity searching. These two examples show that structural features important for biological activity are conserved in the active compounds. In similarity searching, silencing of fingerprint bit positions that account for these features reduces search performance. However, through bit silencing they might be identified and weight vectors can be derived to emphasize significant bit positions. As a result, search performance may improve.

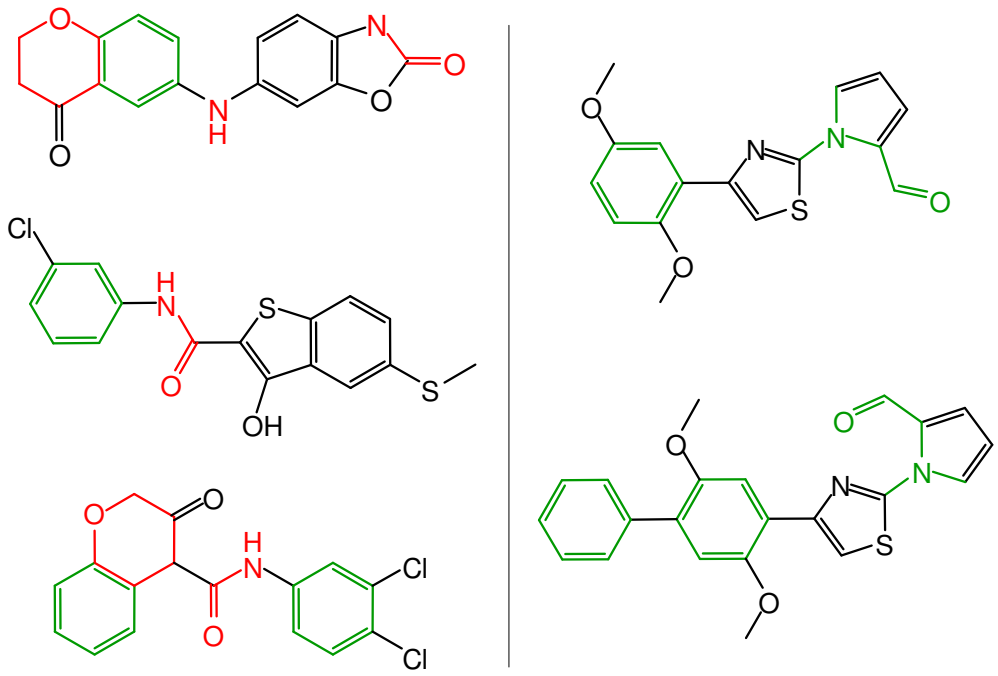


Figure 5.8: Substructures of COX inhibitors with high and low weights. Shown on the left are examples of COX inhibitors that were correctly identified using the bwTc metric but not conventional Tc calculations. On the right, ZINC compounds are shown that were found in COX compound selection sets obtained on the basis of Tc calculations but were de-selected when the bwTc metric was applied. Substructures having high and low bwTc weights are highlighted in red and green, respectively.

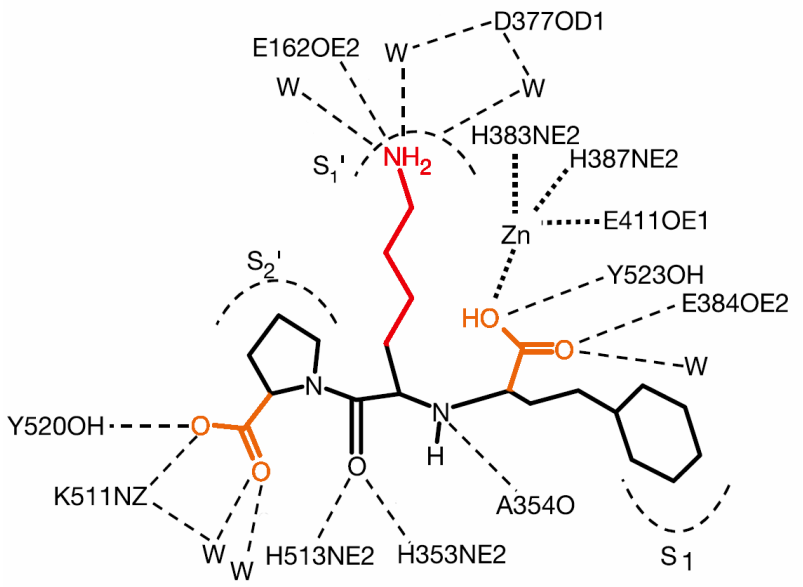


Figure 5.9: Conserved substructures of ACE inhibitors with high weights. Based on the work of Natesh et al.,[55] two substructures that correspond to the two bits having highest weights are highlighted in red (most significant feature, with bit weight 340%) and orange (300%) in a schematic view of ACE-lisinopril crystallographic complex. Dashed lines denote hydrogen bonds.[55]

5.3 Class-specific weighted Tversky similarity

In the previous section a class-directed similarity metric, bwTc, has been introduced. Emphasizing compound class-specific bit patterns in similarity calculations has been shown to improve fingerprint search performance.[9,44,45] By systematic silencing of bit positions, the contribution of each fingerprint bit to the search performance can be evaluated. A bit position is assigned a high weight in the bwTc similarity comparison if its silencing causes a reduction in the recall of active compounds; the larger the reduction, the higher the bit significance and hence the weight.

Similarly, for two fingerprints A and B the bit position weight vector can also be incorporated into Tversky coefficient in order to obtain a bit position-weighted Tv, or bwTv:

$$Tv(\boldsymbol{A},\boldsymbol{B},\alpha) = \frac{\sum_{i=1}^{N} a_i b_i}{\sum_{i=1}^{N}[\alpha(a_i^2-b_i^2)+b_i^2]} \tag{5.4}$$

$$\overset{incorporate\ \ weight\ \ vector}{\longrightarrow}$$

$$bwTv(\boldsymbol{A},\boldsymbol{B},\boldsymbol{W},\alpha) = \frac{\sum_{i=1}^{N} a_i b_i w_i}{\sum_{i=1}^{N}[(\alpha(a_i^2-b_i^2)+b_i^2]w_i} \tag{5.5}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{W}$ are defined as in Eq.(5.3) and α is the weight on unique bit settings in reference fingerprint. Analogously to Tc and bwTc, here only "1" bit positions are taken into consideration.

In order to also account for "0" bit positions, in *Chapter 3* an alternative form of the Tversky coefficient has been defined that accounts for bit positions that are set off (Eq.(3.6)):

$$\begin{aligned} Tv'(A,B,\alpha) &= \frac{c'}{\alpha(a'-c')+(1-\alpha)(b'-c')} \\ &= \frac{c'}{\alpha(a'-b')+b'} \end{aligned}$$

where a' and b' denote the number of "0" bits in A and B, respectively, and c' the number of "0" bits common to both. Alternatively, the general form of Tv' is represented as

$$Tv'(\boldsymbol{A},\boldsymbol{B},\alpha) = \frac{\sum_{i=1}^{N} a_i' b_i'}{\sum_{i=1}^{N}[\alpha(a_i'^2-b_i'^2)+b_i'^2]} \tag{5.6}$$

where a_i' and b_i' are the complements of the i-th bit element (i.e. $1-a_i$ and $1-b_i$, respectively) in fingerprint $\boldsymbol{A}$ and $\boldsymbol{B}$. Incorporating the weight vector $\boldsymbol{W}$ into this representation then produces

$$bwTv'(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{W}, \alpha) = \frac{\sum_{i=1}^{N} a_i' b_i' w_i}{\sum_{i=1}^{N} [(\alpha(a_i'^2 - b_i'^2) + b_{i}^2] w_i} \quad (5.7)$$

By combining Tv and Tv' and introducing a weighting parameter β, the relative contributions of "1" and "0" bits can be balanced (Eq.(3.7)):

$$wTv(A, B, \alpha, \beta) = \beta \frac{c}{\alpha(a-b)+b} + (1-\beta)\frac{c'}{\alpha(a'-b')+b'}$$

Accordingly, a weighted linear combination of Eq.(5.5) and Eq.(5.7) incorporating the β parameter then is

$$\begin{aligned} wbwTv(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{W}, \alpha, \beta) &= \beta \frac{\sum_{i=1}^{N} a_i b_i w_i}{\sum_{i=1}^{N} [(\alpha(a_i^2 - b_i^2) + b_i^2] w_i} \\ &+ (1-\beta)\frac{\sum_{i=1}^{N} a_i' b_i' w_i}{\sum_{i=1}^{N} [(\alpha(a_i'^2 - b_i'^2) + b_i'^2] w_i} \end{aligned} \quad (5.8)$$

It follows that this similarity metric integrates three weighting schemes: (a) relative weights on "1" bit settings of reference and database compounds, (b) relative weights on "1" and "0" bit positions, (c) compound class-specific weights on "1" bits. Thus, it is designed to balance differences in complexity between reference and database molecules and emphasize compound class-specific bit patterns in similarity calculations. In Figure 5.10, the design and calculation scheme of wbwTv is illustrated.

Modulating complexity effects with wbwTv

Extended analysis were carried out to address the two questions: (a) how can similarity metrics be combined so that molecular complexity effects are modulated and compound class-specific fingerprint features are emphasized; and (b) what are the advantages of using such similarity metrics in fingerprint-based similarity searching. Multiple compound reference sets having different complexity and screening databases of different composition were used to systematically investigate differences in search performance of alternative similarity coefficients.

For training and similarity searching, three sets of database compounds were used including a randomly collected set of 5000 ZINC[39] compounds (previously utilized in bwTc calculations), the NCI database[40] previously used in wTv calculations, and another randomly selected set of 50000 ZINC compounds that approximately matched the size of the NCI database. These screening databases

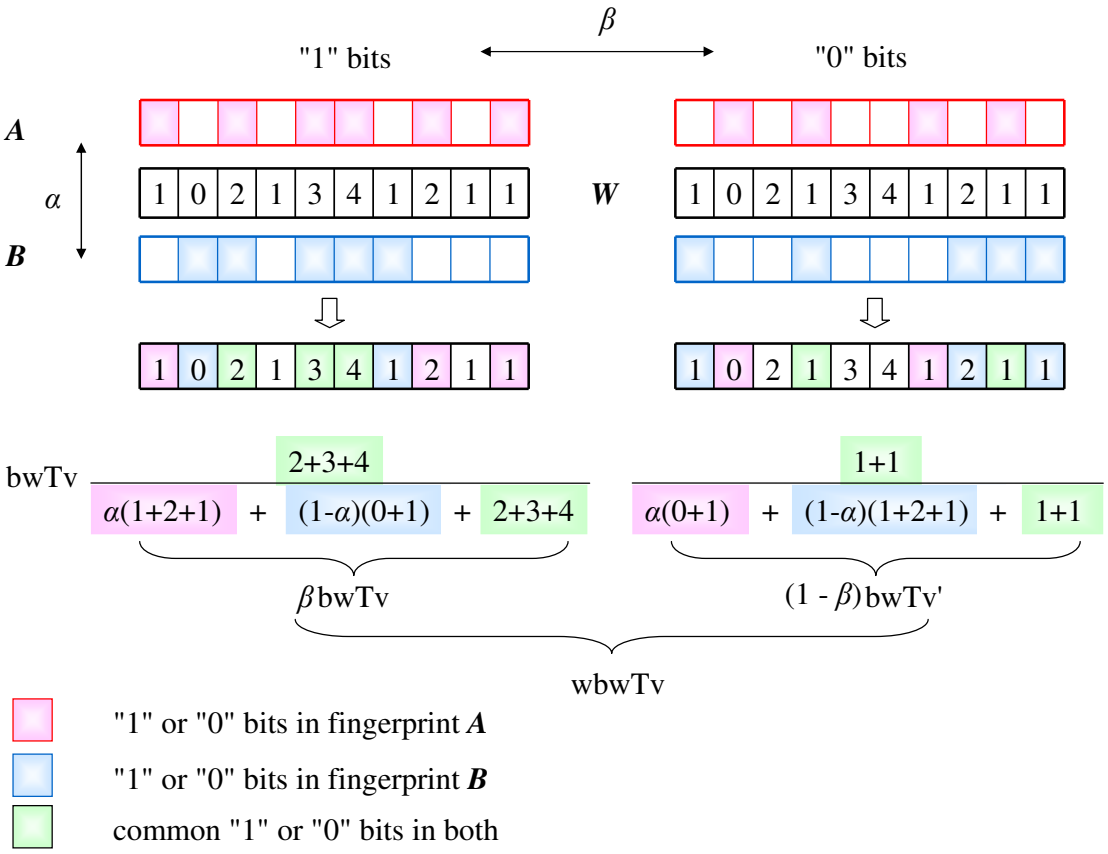


Figure 5.10: Calculation of wbwTv. Two hypothetical fingerprints ***A*** and ***B*** consisting of ten bits are compared with wbwTv using the bit position-dependent weight vector ***W*** that assigns compound class-specific weights to "1" bits. The two parameters α and β modulate the relative weights on reference vs. database compounds and on "1" vs. "0" bits, respectively. The variables a, b, and c in Tv calculations are replaced with summation of weighted terms as described in Eq.(5.5). In addition, in Tv′, a', b' and c' are modified according to Eq.(5.7). For example, the number of "1" bits shared by the two fingerprints is 3 ($c = 3$) in conventional calculations, whereas weighted calculations produce the value $2 + 3 + 4 = 9$ (highlighted in green). The weighted linear combination of Eq.(5.5) and Eq.(5.7) yields the final wbwTv similarity value.

were named ZINC5K, NCI, and ZINC50K, respectively. The ZINC5K screening set was used to derive bit weight vectors, as described in Section 5.1, and evaluate systematic parameter variations in wTv and wbwTv calculations.

For bit silencing and systematic similarity search calculations, ten activity classes out of the 21 classes used in bwTc calculations were utilized and filtered as in Section 5.1. From each activity class, a subset of potential database hits of varying size (ranging from 10-100, Table 5.1) was selected having a MACCS bit density comparable to the screening database compounds, i.e. an average bit density of 22.3% (ZINC) to 25.7% (NCI). These subsets of active molecules having comparable complexity to screening set compounds served as active database compounds (ADC) for similarity searching. The bit density requirements limited the number of active compounds that could be selected as ADC. The remaining active molecules were utilized as training compounds for bit silencing and the derivation of the weight vectors.

To derive the weight vectors, the training process as previously described was conducted. From each activity class training set, ten different subsets of 20 compounds each were randomly selected and the remaining compounds were added to ZINC5K to derive the bit weight vector. Therefore, for each of the ten reference sets, 166 bit silencing calculations were carried out (i.e. one for each bit position) in combination with 20-NN ranking, which equally takes contributions of all reference molecules into account. Hit rates were calculated for the top-ranked 100 database molecules. From these hit rates, ten individual weight vectors were calculated for each reference set with $sf = 100$ and the activity class-specific weight vector for each class was derived by averaging these reference set vectors. These ten class-specific weight vectors have been incorporated in bit position-weighted similarity calculations as illustrated in Figure 5.11.

Next, active reference compounds with different levels of complexity were selected for each activity class training set, i.e. 20 compounds with lowest bit density, 20 having average bit density, and 20 with highest bit density. These different reference sets for similarity searching were named level L (low complexity), M (moderate complexity), and H (high complexity). Level L reference compounds were comparable in complexity (i.e. bit density) to screening set compounds or slightly more complex. For these reference sets, MACCS bit densities are reported in Table 5.3. These sets were used as the reference sets to search for ADC of the corresponding activity class, as shown in Figure 5.11. Exemplary structures of reference and screening set compounds and ADC are shown in Figure 5.12.

Similarity search calculations using six similarity metrics (Tc, bwTc, wTv, wbwTv, Forbes, simple match) were carried out combined with 20-NN ranking in ZINC50K and NCI. Compound recovery rates (i.e. the percentage of correctly identified ADC relative to the total number of ADC) were calculated

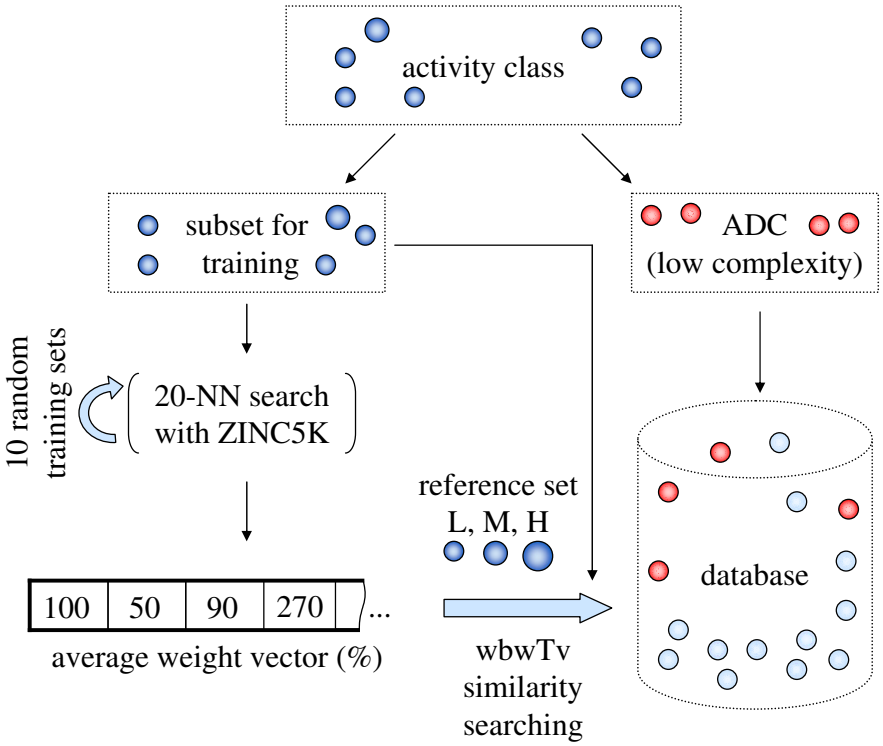


Figure 5.11: Evaluation of wbwTv. The calculation protocol to systematically test wbwTv is illustrated. For each activity class the set of potential hits is independent of the training subset or the reference set L (consisting of low-complexity compounds), M (medium complexity), or H (high complexity). The weight vector used in bwTc similarity searching is the average result of ten independent random training experiments (shown in brackets, see Figure 5.2).

for the top-ranked 100 database compounds (Table 5.4). In wTv and wbwTv test calculations, the α and β parameters were systematically and independently varied between 0 to 1 in increments of 0.1. For the resulting 121 combinations, the top recovery rate of each calculation was determined. Hence, parameter variation was not involved in the training process to derive the weight vector. In addition, different data fusion techniques were compared to wbwTv calculations. Table 5.5 reports the results for 20-NN, 1-NN, and centroid strategies and the Tc and Forbes similarity metrics on these compound test sets and the NCI database as control calculations.

Complexity effects and conventional search strategies

The influence of varying molecular complexity on MACCS Tanimoto similarity calculations is evident in Table 5.4. For all compound classes and screening

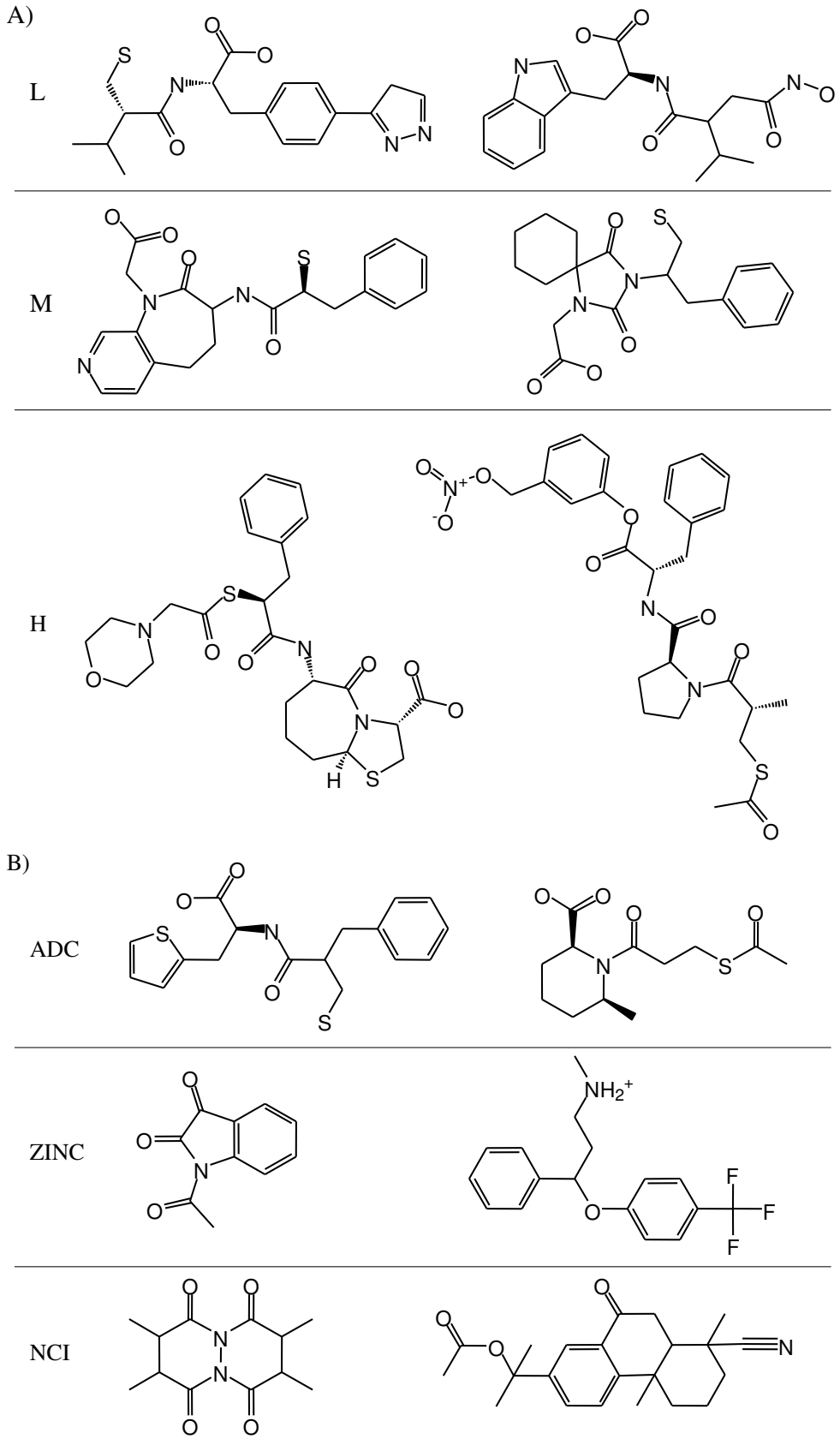


Figure 5.12: Exemplary compounds. For activity class ACE, examples are shown of (A) reference molecules of different complexity (level L, M and H) and (B) active database compounds (ADC) and screening database molecules from ZINC and NCI having comparable complexity.

class	average reference bit density (%)		
	L	**M**	**H**
ACE	27.7	32.3	37.2
ADR	27.6	33.5	41.1
CAM	29.2	36.0	41.8
CLG	28.3	35.4	40.2
FXA	27.7	38.2	49.7
MM1	27.0	33.2	38.4
PA2	31.4	36.1	37.0
PAF	27.8	35.1	43.3
PKC	30.4	35.4	40.5
SST	27.8	34.9	37.7

Table 5.3: Activity classes and complexity levels. The average MACCS bit densities for reference sets having different levels of complexity (L: low complexity, M: moderate complexity, H: high complexity) are reported. The average bit density of reference molecules having different levels of complexity (L, M, H) are reported.

databases, compound recall of Tc calculations systematically decreased with increasing fingerprint bit density of reference compounds. For the least complex reference molecules (complexity level L), active compounds were detected in standard search calculations for seven of ten classes in the ZINC and all ten classes in the NCI database. By contrast, for the most complex reference compounds (level H), Tc calculations consistently failed in ZINC and for all but one class in NCI. Thus, in the presence of significant complexity effects, standard MACCS Tc calculations essentially failed to recover any active compounds. Using complexity level M reference molecules, active compounds were also only detected for two and three classes, respectively. These results are consistent with the observation from *Chapter 3* and *Chapter 4*.

Nearest neighbor calculations produced better results than centroid searches, but were overall inferior to wbwTv calculations, as reported in Table 5.4, especially when reference compounds of high complexity were used. 1-NN Tc calculations moderately increased the search performance of 20-NN calculations by 1% to 9% for reference sets L-H, but recovery rates of wbwTv were 10% to 14% higher. A similar trend was observed for the Forbes coefficient. Overall, there were only two instances where 1-NN Tc performed better than wbwTv or wTv (PAF set L and SST set L) and two where 1-NN Forbes performed better (ADR set H and MM1 set H), but the differences were marginal. It follows that data fusion techniques were not capable of effectively balancing molecular complexity effects, as expected. By contrast, balancing complexity effects through wbwTv led to overall highest search performance.

reference set		ZINC						NCI					
		Tc	**bwTc**	**max wTv**	**max wbwTv**	**Forbes**	**simple match**	**Tc**	**bwTc**	**max wTv**	**max wbwTv**	**Forbes**	**simple match**
ACE	L	57	60	77	**83**	33	77	57	60	**83**	**83**	47	**83**
	M	3	3	**33**	30	23	10	3	3	**40**	30	27	10
	H	0	0	**27**	23	17	0	0	0	27	**30**	23	3
ADR	L	4	7	**10**	9	4	7	6	26	11	**23**	4	10
	M	0	0	**9**	1	3	6	0	0	**10**	9	1	7
	H	0	0	**6**	0	3	0	0	0	**6**	3	3	0
CAM	L	0	**30**	20	**30**	0	20	20	30	**40**	**40**	20	**40**
	M	0	**20**	**20**	**20**	**20**	0	0	20	20	**30**	20	0
	H	0	0	0	0	0	0	0	0	0	0	0	0
CLG	L	40	30	**45**	40	20	40	40	10	**45**	40	35	**45**
	M	0	15	30	**40**	25	0	0	10	**40**	**40**	30	0
	H	0	5	15	**25**	0	0	0	5	15	**25**	5	0
FXA	L	0	3	8	**25**	0	8	5	5	15	**28**	0	8
	M	0	0	0	0	0	0	0	0	0	0	0	0
	H	0	0	0	0	0	0	0	0	0	0	0	0
MM1	L	0	0	**10**	**10**	**10**	**10**	10	0	15	**20**	5	15
	M	0	0	**5**	**5**	**5**	0	0	0	**5**	**5**	**5**	0
	H	0	0	**15**	10	10	0	0	0	0	0	0	0
PA2	L	**3**	**3**	**3**	**3**	2	**3**	3	3	5	**10**	2	3
	M	0	0	**3**	**3**	0	**3**	3	2	**8**	7	0	3
	H	0	2	4	**8**	0	3	3	3	11	**12**	0	7
PAF	L	2	6	**16**	12	2	**16**	4	10	**20**	12	0	16
	M	0	0	0	**2**	0	0	0	0	0	**2**	0	0
	H	0	0	0	0	0	0	0	0	0	0	0	0
PKC	L	4	10	20	**26**	11	10	4	10	16	**21**	6	13
	M	0	4	20	**29**	13	1	0	6	17	**19**	10	6
	H	0	0	**7**	**7**	**7**	0	0	0	7	**10**	7	0
SST	L	20	23	25	**28**	20	23	20	23	23	**25**	20	23
	M	5	3	20	**25**	20	20	10	3	20	**23**	10	20
	H	0	0	18	**23**	8	0	0	0	18	**23**	0	0
avg	L	13	17	23	**27**	10	21	17	18	27	**30**	14	26
	M	1	5	14	**16**	11	4	2	4	16	**16**	10	5
	H	0	1	9	**10**	4	0	0	1	8	**10**	4	1

Table 5.4: Similarity searching using different similarity coefficients. Average recovery rates (in %) are reported for MACCS search calculations using different similarity coefficients and the ZINC50K ("ZINC") and NCI screening databases. For each class and the average ("avg") over all classes, L, M, and H report the results for reference sets of varying complexity, according to Table 5.3. In each row, the best-performing similarity coefficient is highlighted in bold.

reference set		Tc			Forbes		
		20-NN	**1-NN**	**centroid**	**20-NN**	**1-NN**	**centroid**
ACE	L	57	60	73	47	40	47
	M	3	0	3	27	3	27
	H	0	0	0	23	3	23
ADR	L	6	21	9	4	9	4
	M	0	1	4	1	4	1
	H	0	0	0	3	7	3
CAM	L	20	30	20	20	20	20
	M	0	0	0	20	20	20
	H	0	0	0	0	0	0
CLG	L	40	45	40	35	25	35
	M	0	0	0	30	10	25
	H	0	0	0	5	0	5
FXA	L	5	10	5	0	23	0
	M	0	0	0	0	0	0
	H	0	0	0	0	0	0
MM1	L	10	15	15	5	10	5
	M	0	0	0	5	5	5
	H	0	0	0	0	5	0
PA2	L	3	8	3	2	4	2
	M	3	4	3	0	2	0
	H	3	3	30	0	3	0
PAF	L	4	26	8	0	4	0
	M	0	0	0	0	0	0
	H	0	0	0	0	0	0
PKC	L	4	14	7	6	7	6
	M	0	1	0	10	9	10
	H	0	0	0	7	9	7
SST	L	20	35	23	20	23	20
	M	10	20	20	10	20	5
	H	0	8	0	0	3	0
avg	L	17	26	20	14	17	14
	M	2	3	3	10	7	9
	H	0	1	0	4	3	4

Table 5.5: Similarity searching using different data fusion strategies. Average recovery rates (in %) are reported for MACCS search calculations using two similarity coefficients, Tc and Forbes, and three data fusion techniques, 20-NN, 1-NN and centroid, are compared for the NCI database. For each class and the average ("avg") over all classes, L, M, and H report the results for reference sets of varying complexity, according to Table 5.3.

Alternative similarity coefficients

Adding compound class-specific weights to bit positions (bwTc) only marginally improved the search performance for levels H and M. For level of L (where complexity effects were essentially absent), bwTc calculations produced moderate increases in compound recall for seven of ten classes for ZINC and six for the NCI database (i.e. 3%-10%, with one exception). Thus, complexity effects severely limited the influence of compound class weight vectors and the search performance of bwTc calculations.

For the most complex reference molecules, Forbes calculations detected active compounds in five ZINC and three NCI cases where both Tc and bwTc calculations failed, whereas simple match calculations did not produce notable increases. However, Forbes calculations also frequently failed to detect active compounds on the basis of complex reference molecules and showed lower performance than Tc, bwTc, or simple match for level L reference molecules. For low-complexity reference compounds, the performance of the simple match coefficient was comparable to Tc and bwTc in ZINC but was higher for seven of ten classes in NCI.

wTv and wbwTv

Different from Tc, bwTc, Forbes, or simple match, the bit position-independent weighted Tversky coefficient (wTv) balances complexity effects by modulating relative contributions of "1" and "0" bit positions. In this case, a systematic increase in compound recovery rates was found in both screening databases. For level H and level M reference compounds, wTv calculations succeeded in seven ZINC and eight NCI instances, respectively, to recover active compound and recall rates of up to 27% (level H) and 40% (level M) were obtained. Here, the general trend was also observed that recovery rates often increased from level H to level L. For the least complex reference molecules, wTv calculations produced average hit rates over 10 classes of ∼23% in ZINC and ∼27% in NCI. Thus, directly addressing complexity effects at the level of similarity calculations clearly improved the search results.

When applying wbwTv, consistent improvements in recovery rates over all complexity levels were observed. Top recovery rates were obtained in 18 of 30 cases (i.e. of three calculations per activity class) with ZINC and in 19 cases with NCI database. Thus, despite differences in compound compositions, results obtained for the ZINC and NCI screening databases were overall similar. In many instances, wbwTv calculations produced recall rates of ∼20% or more, while other similarity coefficients (in particular, Tc) completely failed. However, wbwTv calculations were not always successful. For example, for classes CAM, FXA, or PAF, level H reference molecules presented an intractable search problem for any of the similarity coefficients. In one case, ADR level H, wTv

calculations detected a few active compounds (recovery rate 6%), but wbwTv essentially failed. In another case, PAF level M, the opposite occurred. With these minor exceptions, a clear trend was observed: when wTv was not capable of detecting active compounds, wbwTv was not either. However, when wTv calculations succeeded, an increase in recovery rates was often observed when wbwTv was applied, although the relative search performance varied in a compound class-dependent manner. For the total of 60 test calculations reported in Table 5.4, wTv and wbwTv recovery rates were the same in 19 cases and wTv and wbwTv performed best in 14 and 27 cases, respectively. Thus, taken together, these findings indicated that simultaneous balancing of complexity effects and emphasizing of class-specific bit settings yielded overall best performance in these difficult similarity search test cases.

Recovery rate distributions have been compared for the overall preferred wTv and wbwTv coefficients under systematic variation of the α and β parameters. Representative examples are shown in Figure 5.13 and Figures B.2-B.4. In these recovery rate landscapes, regions colored in red represent parameter combinations producing high recovery rates. For PKC screening in ZINC, shown in Figure 5.13, areas of high recovery rates were larger for wbwTv than for wTv. A similar trend was observed for PKC in the NCI, although recovery rates were in this case lower for both coefficients (Figure B.2). Equivalent observations were also made for MMI in ZINC (Figure B.3) and SST in NCI (Figure B.4). The recovery rate landscapes also reveal trends for preferred α and β parameter settings. For complexity level H, combinations of low α and high β or vice versa generally produced highest recovery rates, although search performance was low in these cases. Going from complexity level H to M and L combinations of increasingly larger α and β value ranges produced highest rates, while search performance was increasing.

In general, wbwTv calculations produced larger areas of high recovery rates (red in Figure 5.13) than wTv calculations and smaller areas where calculation produced only low recovery of active compounds (light blue in Figure 5.13). This means that wbwTv search calculations were less sensitive to (α, β) parameter settings than wTv calculations (i.e. more wbwTv parameter combinations produced high compound recall). Therefore, taking bit position-specific information into account made wbwTv search calculations more stable over all complexity levels, in addition to achieving net increases in recovery rates.

5.4 Summary

In this chapter, the bit silencing technique was utilized to introduce two class-specific similarity metrics, bwTc and wbwTv. Previous analyses of bit settings in keyed fingerprints have largely focused on identifying bit positions that are

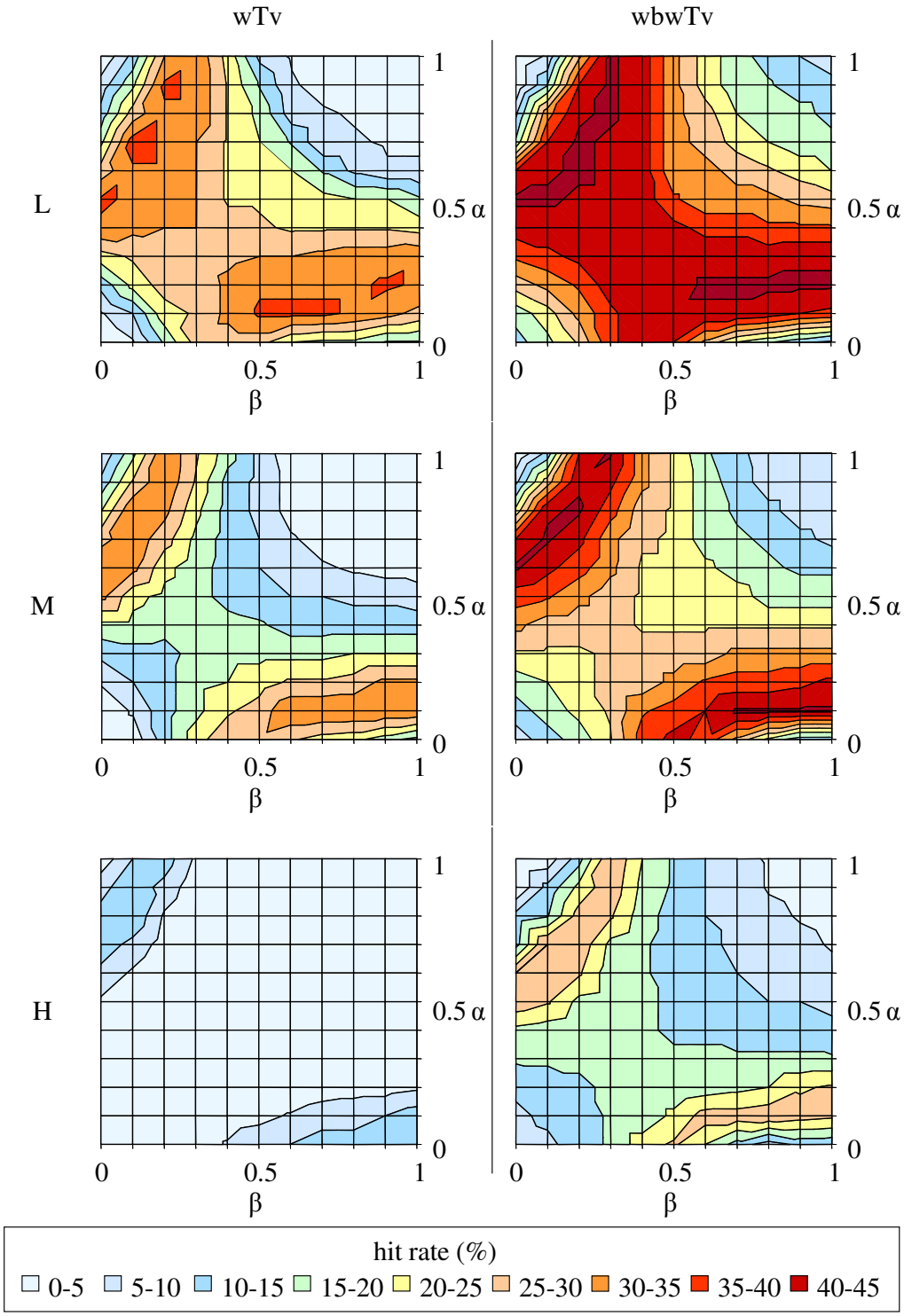


Figure 5.13: Recovery rate landscapes. Shown are maps reporting search results for wTv and wbwTv calculations under systematic parameter variation using reference sets of different complexity for class PKC against ZINC5K database.

set on with high frequency in compounds having similar activity and attempted to emphasize such positions, for example, through fingerprint scaling or calculation of consensus fingerprints for activity classes. The bit silencing technique, as introduced herein, makes it possible to systematically evaluate positive or negative contributions of all bit positions in keyed fingerprints to similarity searching.

Silencing calculations on a large number of activity classes consistently revealed differential contributions of MACCS bit positions. In many instances, individual bit settings were found to substantially increase or decrease search performance. On the basis of these observations, bit position-dependent weight vectors were derived that account for positive or negative contributions of bits and used to modify the Tanimoto coefficient and weighted Tversky coefficient described in *Chapter 3* in a compound class-specific manner.

The notion of class-specific modulation of bit position weights might be utilized as a search strategy to adjust to different similarity searching problems. For compound reference sets with varying complexity, search situations where conventional Tanimoto similarity calculations consistently failed were observed. In the presence of complexity effects, neither standard Tanimoto similarity calculations nor other conventional similarity metrics such as Forbes and simple match could achieve a high recovery rate. Furthermore, bwTc calculations, which emphasized compound class-specific bit patterns also failed to produce significant compound recall. The results discussed above mirror the crucial role of complexity effects that were only effectively balanced in wTv calculations. With wbwTv, a similarity coefficient that combines the complexity-balancing potential of wTv calculations with class-specific bit weight vectors has been derived. It is a complex similarity metric that is based on the Tversky formalism and simultaneously balances complexity effects and emphasizes class-specific bit settings. In systematic similarity searching over different compound classes and complexity levels, the wbwTv coefficient often produced significant recall in cases where standard Tanimoto similarity calculations failed and further improved the performance of the weighted Tversky coefficient that was previously introduced. Moreover, compared to the Forbes and simple match coefficients, which have been shown to be particularly suitable for searching with complex reference molecules, wbwTv achieved consistently higher recovery rates over all reference set complexity levels. In addition to practical similarity applications, wbwTv calculations can be utilized to study the relationship between molecular complexity and compound class characteristic features and further explore basic aspects of molecular similarity measures.

Chapter 6

Shannon Entropy-Based Similarity Search Strategy

In the previous chapters, several fingerprint search methods have been discussed. In this chapter, another fingerprint search strategy is discussed that also combines reference compound information prior to similarity assessment and that is based on the Shannon entropy concept.[56]

Shannon entropy (SE) was introduced in 1948 in information theory and was originally applied to assess the information content of messages transmitted through different channels.[56] In this context, messages with high information content (high SE) display few or no recognizable patterns, whereas those having low information content (low SE) exhibit regular patterns that correspond to information redundancy.[57]

The SE concept is readily transferable to molecular fingerprints when bit positions are considered to be individual channels that are capable of transmitting binary signals, i.e. by setting bit positions on (to "1") or off ("0"). Accordingly, chemical compound sets whose fingerprints share similar bit patterns produce low SE values. By contrast, if there is only little bit pattern resemblance, high SE values are obtained. Moreover, if "0" and "1" bits are randomly distributed, the SE value of the system is maximal. Accordingly, given the premise that chemically and biologically similar molecules should yield similar fingerprint bit patterns, ensembles of compounds having similar activity should produce low fingerprint SE values. Then, by adding a compound of unknown activity to the reference set and recalculating the SE for the expanded fingerprint ensemble, the similarity of a test compound to the reference set can be directly assessed. If there is only a small change in the resulting SE value, the fingerprint of the test compound is similar to the reference set and the compound is thought to have similar properties. In the following sections, the fingerprint SE approach is illustrated and systematic test calculations reported. It is shown that the performance of the fingerprint SE approach was in

general comparable to or better than k-NN (nearest neighbor) searching.

6.1 Shannon entropy of binary fingerprints

Given a compound set R and an arbitrary binary fingerprint representation X consisting of N bit positions, the SE value of a single bit position $i \in 1, ..., N$ in the set R is calculated as:[56]

$$SE_i(R) = -p_i log_2(p_i) - (1 - p_i) log_2(1 - p_i) \quad (6.1)$$

with

$$p_i = \sum_{A \in R} x_{iA}$$

Here, p_i represents the relative frequency of "1" bits at fingerprint position i in R. In the case of $p_i = 0$ or $p_i = 1$, $p_i log_2(p_i)$ or $(1 - p_i) log_2(1 - p_i)$ become 0. The Shannon entropy of the complete fingerprint of R is the sum of the individual SE_i values obtained for each bit position i:

$$SE(R) = \sum_{A \in R} SE_i(R) \quad (6.2)$$

Figure 6.1A shows an exemplary SE calculation using a hypothetical four-bit fingerprint.

6.2 Database ranking using Shannon entropy values

Given a set R of reference molecules and its calculated SE value, this value typically changes when adding another compound A to R. The magnitude (and algebraic sign) of the change indicates whether or not A matches a potential common bit pattern of R, as illustrated in Figure 6.1. Two compounds are separately added to the reference set R shown in Figure 6.1A and the SE values are recalculated. The molecule introduced in Figure 6.1B slightly decreases or increases SE_i at bit positions 1 to 2, respectively, and matches the "1" and "0" consensus bits of R at bit positions 3 to 4, respectively, so that SE_3 and SE_4 remain 0. The overall SE value only slightly increases from $SE = 1.81$ to $SE' = 1.94$. By contrast, the compound shown in Figure 6.1C does not match this pattern (SE_3 and SE_4 become 0.72) so that the overall SE value significantly increases to $SE' = 3.38$. Hence, departure from consensus bit positions and patterns in R is associated with a significant entropy penalty. Monitoring such changes in SE values when adding individual test compounds to reference

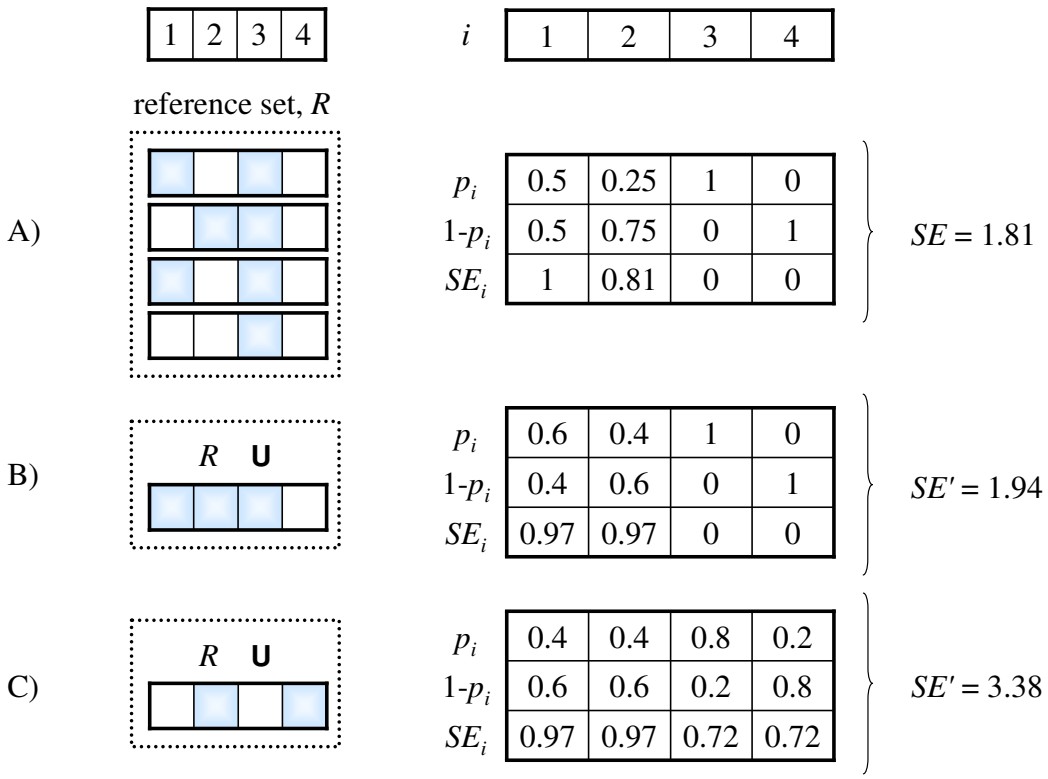


Figure 6.1: Calculation of fingerprint Shannon entropy. A hypothetical four-bit fingerprint is used to illustrate the calculation of Shannon entropy (SE) of individual bit positions and complete fingerprints for a set of molecules. "1" and "0" bits are represented using blue and white cells, respectively. In A), bit strings of a reference set R of four molecules are shown. In B) and C), an additional molecule (bit string) is added to R. For the three different compound sets, the probability p_i for a "1" bit, the probability $1 - p_i$ for a "0" bit, and the corresponding Shannon entropy (SE_i) are reported for each bit position. Resulting Shannon entropies for complete fingerprints (SE or SE') are given on the right.

sets makes it possible to sort database compounds in the order of increasing SE' values corresponding to decreasing molecular similarity and produces a database ranking. Absolute SE values depend on the bit structure of different fingerprints and the composition of the reference sets R and can thus not be transferred or interpreted *a priori*. However, irrespective of the initial SE value of a set of active compounds, similar candidate molecules generally produce less SE changes than dissimilar ones and the relative order of these candidates is only dependent on the level of similarity. Thus, for a given fingerprint and reference set, an SE' ranking of database compounds is obtained.

6.3 Fingerprint Shannon entropy of compound sets

Two databases were used for simulated similarity search calculations, the NCI anti-AIDS database,[40] and a set of 500,000 randomly selected ZINC[39] molecules. Eight compound activity classes were assembled from MDDR,[38] as reported in Table 6.1.

class	designation	number of potential hits for MACCS	number of potential hits for TGD
ACE	angiotensin-converting enzyme inhibitor	30	20
ADR	aldose reductase inhibitor	70	200
CAM	cell adhesion molecule antagonist	10	20
CLG	collagenase inhibitor	20	20
FXA	factor Xa inhibitor	40	10
PA2	phospholipase A2 inhibitor	100	100
PKC	protein kinase C inhibitor	70	100
SST	squalene synthetase inhibitor	40	100

Table 6.1: Activity classes and potential hits. For each activity class, the number of molecules extracted from the MDDR as potential database hits (active database compounds) is reported. Compound sets were specifically assembled to have MACCS or TGD fingerprint bit densities comparable to compound averages in the two test databases. For each class, 20 unique reference compounds with corresponding bit densities were also selected.

To investigate whether the SE approach can distinguish between active and inactive compounds using conventional fingerprint representations, small compound sets consisting of four reference compounds and six test molecules were analyzed. Figure 6.2 shows the molecular graphs of these compounds and reports the SE values for the MACCS fingerprint consisting of 166 bit positions.[16] The four reference molecules shown in the center belong to class ACE and produce an SE value of 41.6. Separately adding three other ACE inhibitors as candidate molecules (depicted in red boxes) changes the SE value of the expanded compound set only very little. Addition of the upper-left molecule actually leads to a small SE reduction ($SE' = 40.4$), separate addition of the compound in the middle results in $SE' = 41.8$ and of the upper-right molecule in $SE' = 42.1$, although these compounds are structurally distinct. By contrast, when separately adding three compounds randomly taken from the NCI database (in blue boxes), SE values significantly increase to 56.5, 60.4, and 62.8, respectively. Thus, in this case, the three active candidate compounds were effectively separated from three inactive ones on the basis of fingerprint SE calculations.

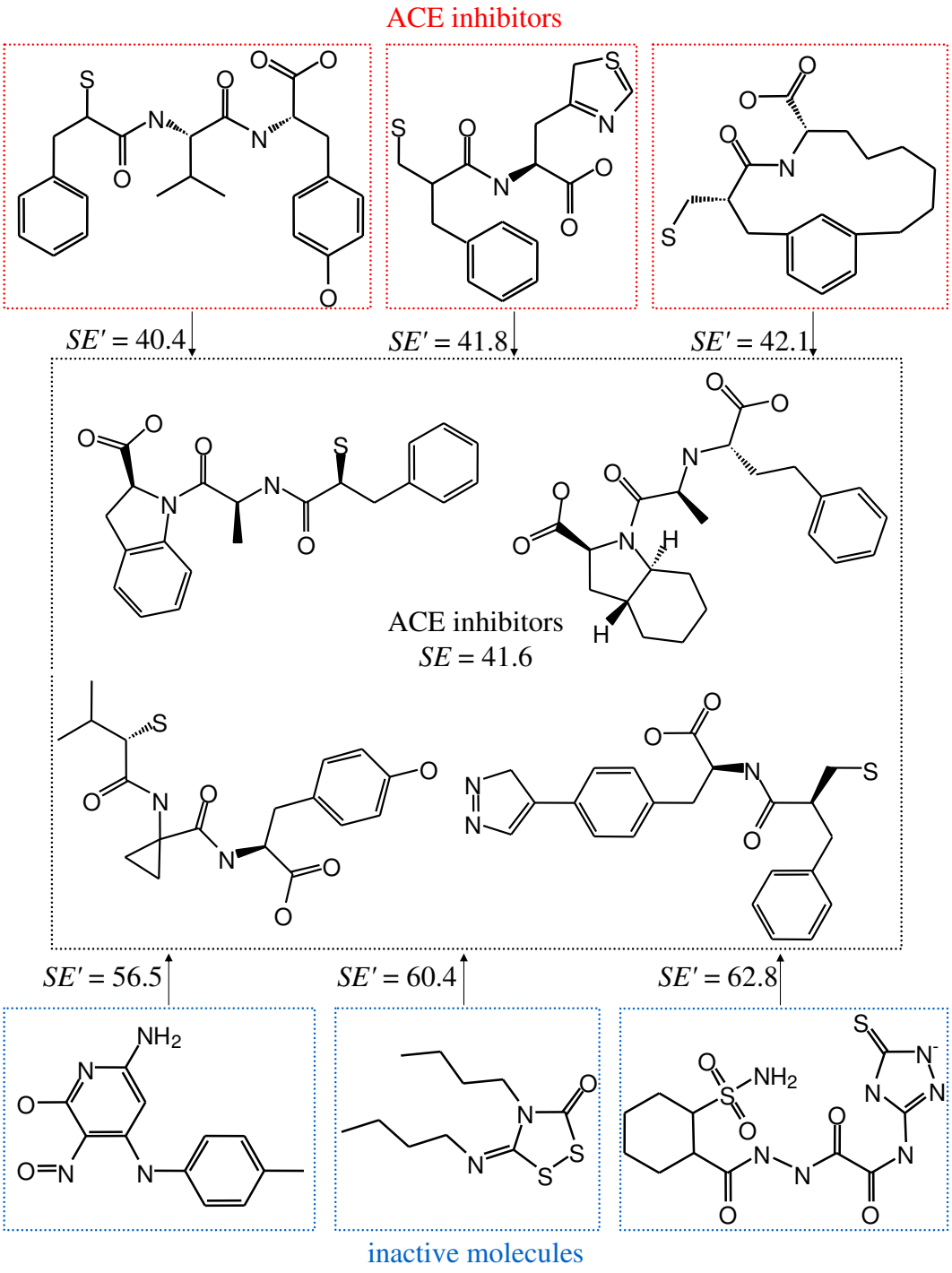


Figure 6.2: Shannon entropy-based fingerprint similarity. The Shannon entropy (SE) of a reference compound set of four ACE inhibitors (shown in the middle box) is reported for the MACCS fingerprint. Three additional ACE inhibitors (shown in red boxes) are separately added to the reference set and SE values are recalculated (SE'). For comparison, three compounds randomly selected from the NCI database (shown in blue boxes) are also separately added to the reference set and SE values are updated.

Test calculations

Two molecular fingerprints were used to test the Shannon entropy-based approach. In addition to MACCS structural keys,[16] the TGD fingerprint was also used that codes for typed graph distances and consists of 420 bit positions (Table 1.1).[19] Bit density analysis and density-based compound selection were carried out prior to similarity searching in order to balance fingerprint complexity effects that can substantially bias similarity calculations, as described in the previous chapters. From each activity class, two compound subsets were selected having MACCS or TGD fingerprint bit densities comparable to the screening databases (Table 6.1) and these compound subsets were used as potential database hits. Furthermore, for each compound class and fingerprint, reference sets of 20 active compounds were selected that also had fingerprint bit densities comparable to the screening databases.

Systematic similarity search calculations were conducted for the combination of each activity class, screening database (NCI or ZINC), and fingerprint (MACCS or TGD), resulting in a total of 32 test calculations. The recovery of active database compounds was monitored for different selection set sizes. The SE approach was compared to three standard similarity search strategies, 1-NN, 20-NN, and centroid calculations. In 20-NN calculations, the average of all 20 pairwise Tc values yielded the final similarity score and in 1-NN calculations, the largest of the 20 individual values was taken. For the centroid method, an average bit string was derived from the 20 active reference compounds and compared to database molecules in Tc calculations.

Recovery rates for selection sets of 100 and 1000 compounds are reported in Table 6.2. Results of the best-performing similarity search approach are highlighted in bold for each trial and selection set size. The results in Table 6.2 reveal that SE performed consistently better than 20-NN and centroid calculations and that it was overall comparable to or better than 1-NN. Summarizing over the 32 different trials and selection sets of 100 database compounds, SE produced highest recovery rates in 20 cases, 1-NN in ten, centroid in seven, and 20-NN in three cases. Furthermore, for a selection set size of 1000 compounds, SE performed best in 18 cases, 1-NN in 16, centroid in nine, and 20-NN in five. Figure 6.3 shows cumulative recall curves for four test calculations using the MACCS fingerprint and the NCI database. The cumulative recall curves for the other four classes are shown in Figure B.5. These curves further illustrate that SE was generally superior to centroid and 20-NN calculations and that it frequently also performed better than the 1-NN strategy.

For fingerprint similarity searching, the SE approach is computationally less complex than nearest neighbor methods. Nearest neighbor methods require the determination of pair-wise similarity values between a database molecule and each reference compound (e.g. 20 calculations per database molecule in this case). By contrast, SE (and also centroid searching) utilizes the information of

		SE		**centroid**		**20-NN**		**1-NN**	
class		**100**	**1000**	**100**	**1000**	**100**	**1000**	**100**	**1000**
MACCS and NCI	ACE	**83**	**90**	73	**90**	57	**90**	60	80
	ADR	10	39	9	17	6	17	**21**	**44**
	CAM	**40**	**40**	20	**40**	20	**40**	30	**40**
	CLG	**45**	55	40	45	40	45	**45**	**75**
	FXA	**20**	**65**	5	40	5	25	10	35
	PA2	3	14	3	12	3	12	**8**	**16**
	PKC	**16**	**47**	7	26	4	20	14	21
	SST	23	**43**	23	30	20	28	**35**	**43**
	average	**30**	**49**	23	38	19	35	28	44
MACCS and ZINC	ACE	**47**	**83**	40	73	27	57	30	57
	ADR	3	6	3	6	0	4	**13**	**26**
	CAM	**20**	**30**	0	20	0	0	**20**	**30**
	CLG	**35**	**40**	**35**	**40**	20	**40**	25	**40**
	FXA	**5**	**8**	0	5	0	0	3	3
	PA2	**3**	3	**3**	3	**3**	3	2	**4**
	PKC	**4**	**13**	1	4	0	4	3	**13**
	SST	20	20	20	20	20	20	**28**	**40**
	average	**17**	25	13	21	9	16	15	**26**
TGD and NCI	ACE	**50**	**65**	45	**65**	20	55	5	45
	ADR	**4**	**8**	3	7	2	5	**4**	**8**
	CAM	**10**	**15**	0	**15**	0	**15**	0	5
	CLG	**25**	**45**	5	35	5	30	0	25
	FXA	**10**	10	0	10	0	10	0	**20**
	PA2	12	22	**13**	19	11	17	12	**25**
	PKC	12	27	14	27	18	27	**22**	**34**
	SST	8	38	**10**	**40**	9	28	7	12
	average	**16**	**30**	12	28	9	24	9	25
TGD and ZINC	ACE	**25**	**45**	5	**45**	0	20	0	5
	ADR	**1**	**3**	**1**	**3**	**1**	2	0	1
	CAM	0	0	0	0	0	0	0	0
	CLG	0	**20**	0	5	0	5	0	0
	FXA	0	0	0	0	0	0	0	0
	PA2	**7**	8	**7**	**11**	**7**	**11**	1	6
	PKC	4	6	5	7	7	12	**10**	**17**
	SST	**6**	**13**	**6**	10	5	9	3	7
	average	**5**	**12**	3	10	3	7	2	5

Table 6.2: Recovery rates for different similarity search strategies. Recovery rates (in %) are reported for four different similarity search strategies (SE, centroid, 20-NN, 1-NN) and different combinations of fingerprints and test databases (MACCS and NCI, MACCS and ZINC, TGD and NCI, and TGD and ZINC). For each activity class, results are compared for selection sets of 100 and 1000 molecules and the search strategies producing highest recovery rates are highlighted in bold.

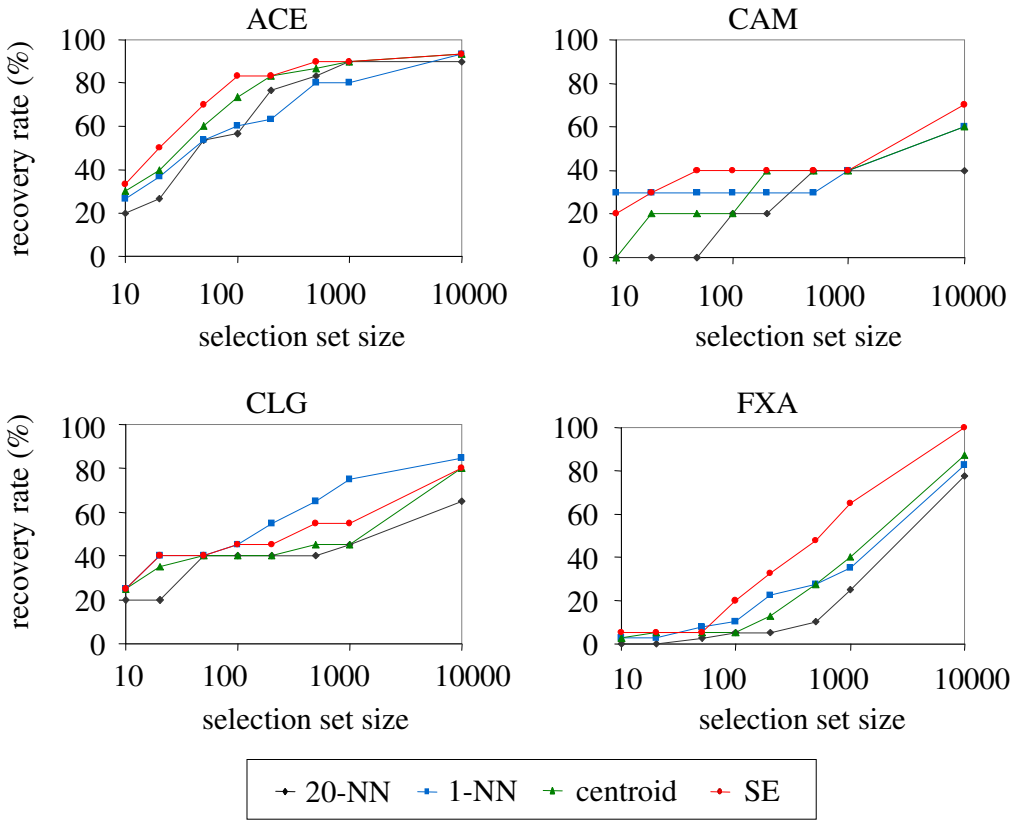


Figure 6.3: Comparison of recovery rates. Recovery rates (in %) for the four different similarity search strategies – 20-NN (black), 1-NN (blue), centroid (green) and SE (red) – using the MACCS fingerprint and NCI database are compared for selection sets of increasing size (shown on a logarithmic scale).

the whole reference set only once to generate a bit frequency profile (or centroid vector). Then, during similarity searching, a database molecule is compared to the frequency profile (or centroid vector) in a single calculation. Thus, while SE leads to comparable or better search results than nearest neighbor methods, it also accelerates similarity searching, especially when large numbers of active reference compounds are available.

6.4 Summary

In this chapter an information entropy-based similarity search strategy has been introduced for binary fingerprints that implicitly captures whether or not a database molecule shares bit patterns characteristic of a reference set. The approach conceptually differs from other search strategies and similarity metrics

and has low computational complexity.

Fingerprint-based similarity searching using sets of active reference compounds requires the application of multiple-template search strategies such as nearest neighbor methods or the centroid technique. While nearest neighbor methods rely on pair-wise compound comparisons and do not utilize the information provided by a reference set as a whole, they have often performed best in comparative benchmark studies. Both the centroid and nearest neighbor methods depend on the calculation of similarity coefficients.

Compared to nearest neighbor methods, the fingerprint Shannon entropy-based approach presented here has the computational advantage that it extracts reference set information only once prior to similarity searching. No pair-wise similarity comparison is required. Test calculations on different compound data sets, fingerprints, and screening databases reveal that the ability of this entropy-based method to detect active compounds is often superior to data fusion techniques and Tanimoto similarity calculations.

Chapter 7

Summary and Conclusions

In this thesis, a number of fingerprint-based similarity search strategies have been introduced that can be utilized to balance or eliminate complexity effects and enhance search performance.

Fingerprint search performance is dependent on intrinsic features of fingerprint descriptors, chosen search strategies, and the measurement of fingerprint similarity. Application of the Tversky similarity measure enables the calculation of molecular fingerprint similarity in a symmetric or asymmetric fashion. However, similarity calculations of molecular fingerprints have asymmetric characteristics only when they have different bit density. For conventional 2D fingerprints such as MACCS, bit density is usually correlated with molecular size and relative differences in molecular complexity influence similarity values. Yet it has been shown that for a fingerprint design with constant bit density such as PDR-FP, Tversky calculations are not affected by differences in molecular complexity. A direct relationship between fingerprint bit densities and asymmetry of Tversky similarity calculations has been revealed in this thesis. In addition, the weighted Tversky coefficient has been developed to balance such asymmetry. Systematic analysis has shown that for virtual screening applications where reference compounds are often more complex than the screening database, fingerprint-based similarity searching can be severely compromised by complexity effects.

Appart from complexity-independent fingerprint design and complexity-modulating similarity metrics, a third approach to compensate for complexity effects has been introduced. By random bit density reduction (bit silencing) of complex reference compounds, search performance can be improved despite the loss of chemical information.

Bit silencing has then been utilized to derive a bit position-dependent weight vector. Systematic bit silencing enables the assessment of the positive and negative contribution of each bit position and different weights are assigned accordingly: bits whose silencing has positive effects are assigned low weights, whereas bits whose silencing has negative effects are critical and thus assigned

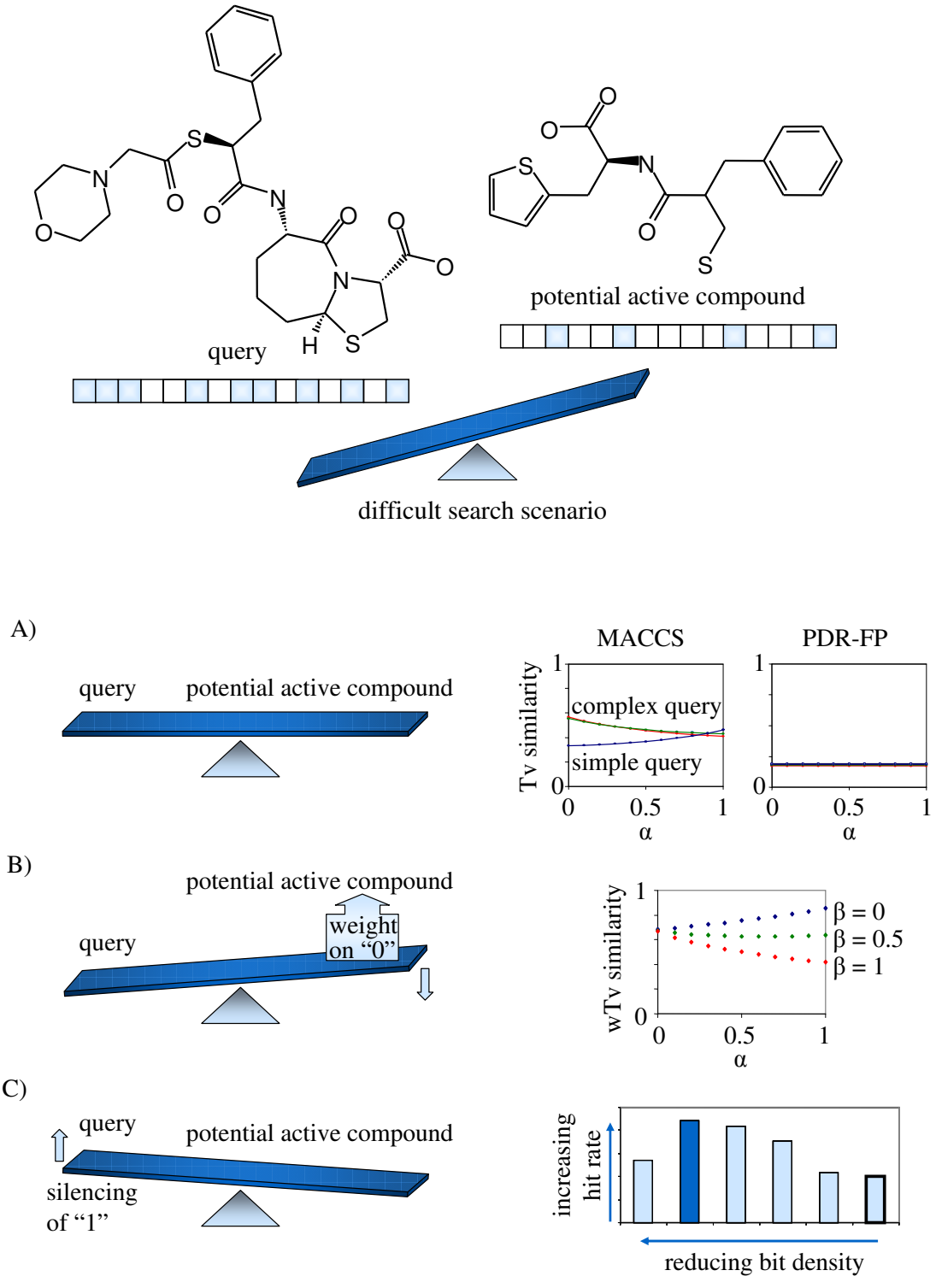


Figure 7.1: Overcoming complexity effects. Complexity effects can be overcome in three ways, A) complexity-independent fingerprint design such as PDR-FP, B) complexity-modulating similarity metric (such as wTv), and C) random fingerprint bit silencing of complex reference compounds.

high weights. These bit weights are represented in vector form, which is the *a priori* information derived from the reference set and specific to the corresponding activity class. Combining this vector with the conventional Tanimoto coefficient has yielded a novel class-specific similarity metric that showed better performance; and combining it with the weighted Tversky coefficient has produced a class-specific coefficient that modulated complexity effects.

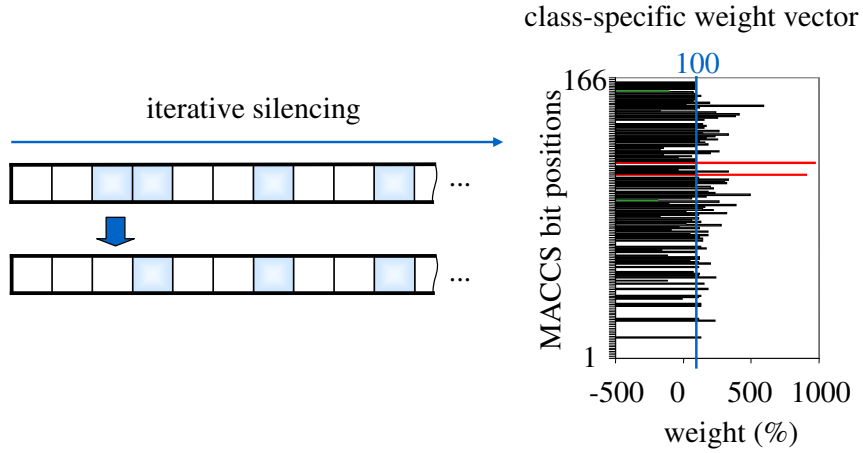


Figure 7.2: Derivation of a weight vector. A class-specific weight vector is derived from iterative silencing of individual bit positions.

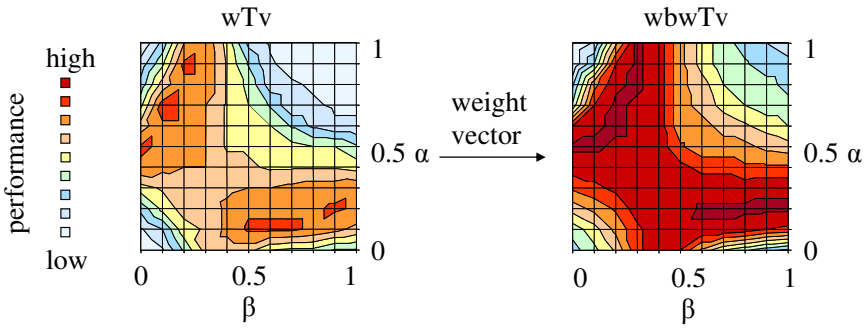


Figure 7.3: Enhanced search performance using the weight vector. Combining the class-specific weight vector with wTv calculations yields wbwTv, which shows further improved performance in similarity searching.

The chemical information of the reference fingerprints can also be transformed into Shannon entropy. In the development of a novel similarity search strategy, the frequency of each bit is derived for the reference set and the total fingerprint Shannon entropy of the set is calculated. Introduction of a database molecule to this set produces less entropy increase if the molecule is similar to the reference set compounds, and more if it is dissimilar.

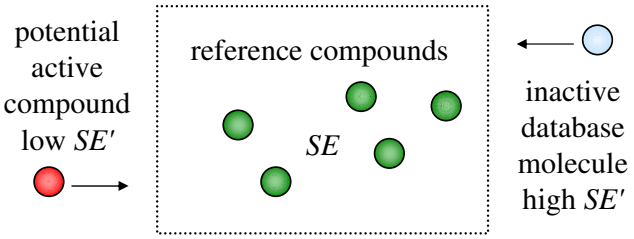


Figure 7.4: Shannon entropy-based similarity. Molecules that are similar to the reference set produce low SE' values when added to the set, whereas dissimilar molecules produce high SE' values.

In summary, taking fingerprint complexity effects into consideration increases the performance of virtual screening applications. The development of novel similarity metrics makes it possible to tailor similarity search calculations in a class-specific manner. These approaches utilize information derived from the known active compounds and modulate parametric space based on activity classes and/or relative differences in fingerprint complexity. As a result, these methods improve the search performance compared to conventional search protocols. Furthermore, systematic analysis of fingerprint properties such as bit density, bit significance, or entropy enables exploration of the chemical information contained in fingerprint descriptors.

Appendix A

Software Tools and Databases

Listed are application software and databases that are used in this thesis.

MACCS by Symyx Software: San Ramon, CA (USA). MACCS (Molecular ACCess System) structural keys represent a two-dimensional fingerprint design, consisting of 166 structural features.[16] *http://www.symyx.com*

MDDR by Symyx Software: San Ramon, CA (USA). MDDR (MDL Drug Data Report) is a molecular database having over 150,000 entries, which are biologically active compounds with annotations.[38] *http://www.symyx.com*

MOE by Chemical Computing Group Inc.: Montreal, QC (Canada). The MOE (Molecular Operating Environment) is an integrated software providing applications for fingerprint calculations such as MACCS, TGD and TGT and property descriptor calculations utilized in PDR-FP.[19,49] *http://www.chemcomp.com*

Perl by Larry Wall. Perl is a freely available programming language. *http://www.activestate.com/activeperl*

NCI by National Cancer Institute. The publicly available NCI anti-AIDS database contains structural and activity data for compounds screened by the AIDS antiviral screening program of the National Cancer Institute.[40] *http://dtp.nci.nih.gov/docs/aids/aids_data.html*

ZINC by UCSF University of California: San Francisco, CA (USA). ZINC (ZINC Is Not Commercial) is a public-domain database of compounds that are commercially available.[39] *http://zinc.docking.org*

Appendix B

Additional Data

B.1 Random reduction of fingerprint bit density

Table B.1 reports the TGD and TGT bit density distribution for two activity classes, COX and RTI, and Table B.2 the search results with random silenced reference sets RS1-RS4.

class		ADC	RS1	RS2	RS3	RS4
TGD	COX	9.9	10.1	13.5	17.3	21.6
	RTI	9.9	10.0	13.1	17.4	21.5
TGT	COX	4.1	4.0	6.3	8.1	10.9
	RTI	3.8	3.8	5.7	7.4	11.7

Table B.1: TGD and TGT bit densities before silencing. Reported are average bit densities (in %) calculated for 100 active database compounds ("ADC") and four different reference sets ("RS1" - "RS4") each consisting of 20 compounds. ADC and RS1 were selected to have bit densities comparable to the BGDB (average bit density of background database is 9.9% for TGD and 3.7% for TGT). Reference sets RS2, RS3, and RS4 were designed to contain molecules of increasing bit densities. The other three activity classes, which were used for MACCS calculations were not included in this control calculation because their bit densities were much higher than the background database and there were not sufficient ADC compounds available. (Average bit densities of TGD fingerprints for those three classes are: LKT–18.7%, PA2–16.6%, TKI–17.0%; and of TGT fingerprints: LKT–8.3%, PA2–6.9%, TKI–8.7%).

reference set (TGD)		**bit density level**					
		1-3%	**5-7%**	**9-11%**	**13-15%**	**17-19%**	**¿21%**
RS1	COX	4	13	**19**			
	RTI	4	24	**41**			
RS2	COX	2	6	9	**10**		
	RTI	1	9	18	**22**		
RS3	COX	4	**5**	2	3	1	
	RTI	**2**	**2**	**2**	**2**	1	
RS4	COX	3	**5**	3	2	1	0
	RTI	**5**	2	2	1	0	0
reference set (TGT)		**bit density level**					
		¡1%	**1-3%**	**3-5%**	**5-7%**	**7-9%**	**¿10%**
RS1	COX	4	20	**28**			
	RTI	3	31	**40**			
RS2	COX	6	6	**11**	9		
	RTI	1	7	**26**	**26**		
RS3	COX	**3**	**3**	2	1	0	
	RTI	**1**	**1**	**1**	0	0	
RS4	COX	2	**3**	2	0	0	0
	RTI	**2**	1	0	0	0	0

Table B.2: Search performance using randomly silenced TGD and TGT reference sets. Hit rates (in %) are listed for reference sets of increasing bit densities and selection sets of 100 compounds. In each block (RS1, RS2, RS3 or RS4), hit rates in the rightmost column indicate that original instead of silenced fingerprints of reference compounds are used as search templates; and bold hit rates indicate the best performance within each row. Numbers in column titles show the actual bit density of template fingerprints. In all calculations, bit strings of database compounds (and ADC hidden among them) remain unmodified.

Figure B.1 reports the average search performance of random silencing of both the template sets (RS3 and RS4) and the database.

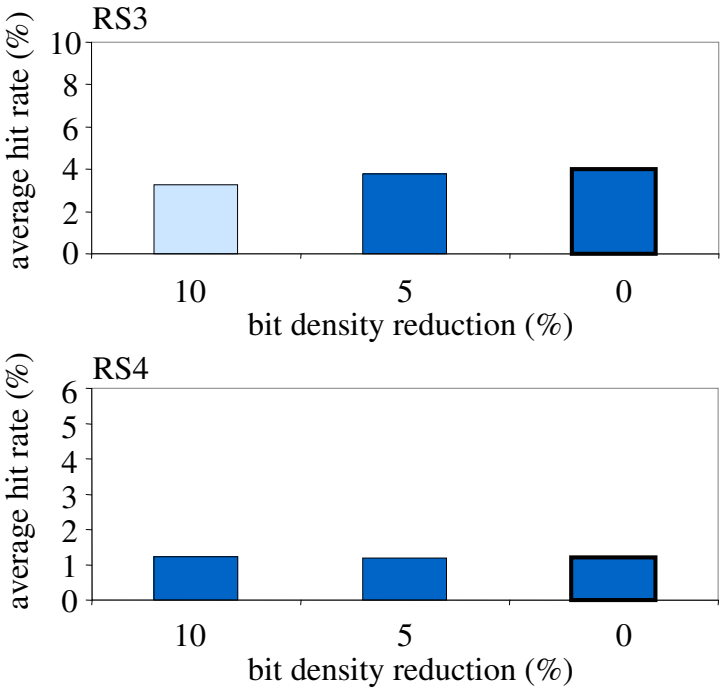


Figure B.1: Hit rates after random bit silencing of all sets. Hit rates averaged over the ten independent trials of all five activity classes are reported using reference set RS3 and RS4. For each reference set, MACCS bit density of reference and database molecules was randomly reduced at the same time to different levels. Bars with bold borders are the hit rates for unmodified fingerprints in similarity searching, while bars colored in dark blue are the optimal hit rates.

B.2 Bit position-weighted similarity metrics

Figure B.2, B.3 and B.4 report the wTv and wbwTv recovery rate landscapes of reference sets of increasing complexity from different activity classes against different databases.

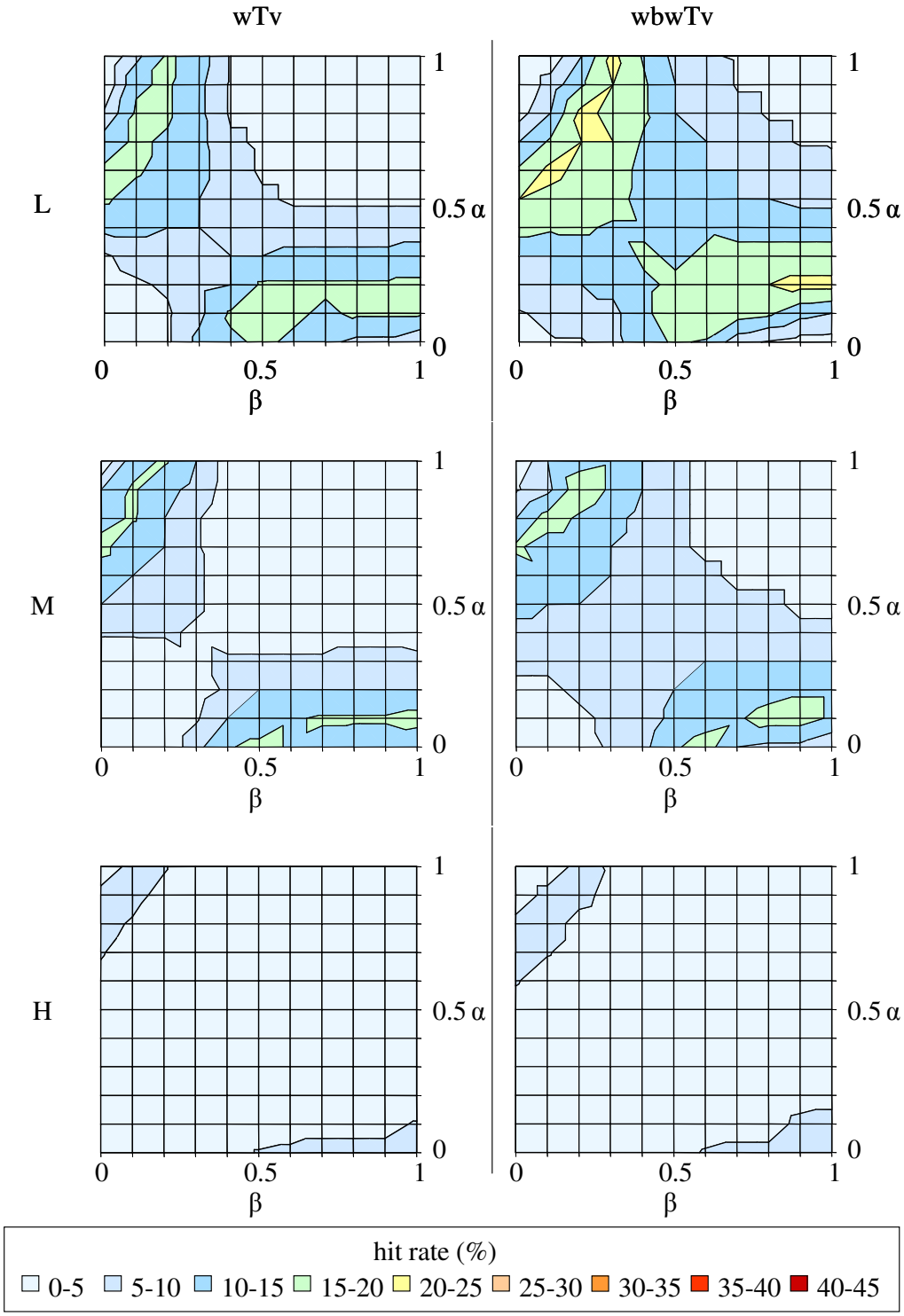


Figure B.2: Recovery rate landscapes (A). Shown are maps reporting search results for wTv and wbwTv calculations under systematic parameter variation using reference sets of different complexity for class PKC against NCI database.

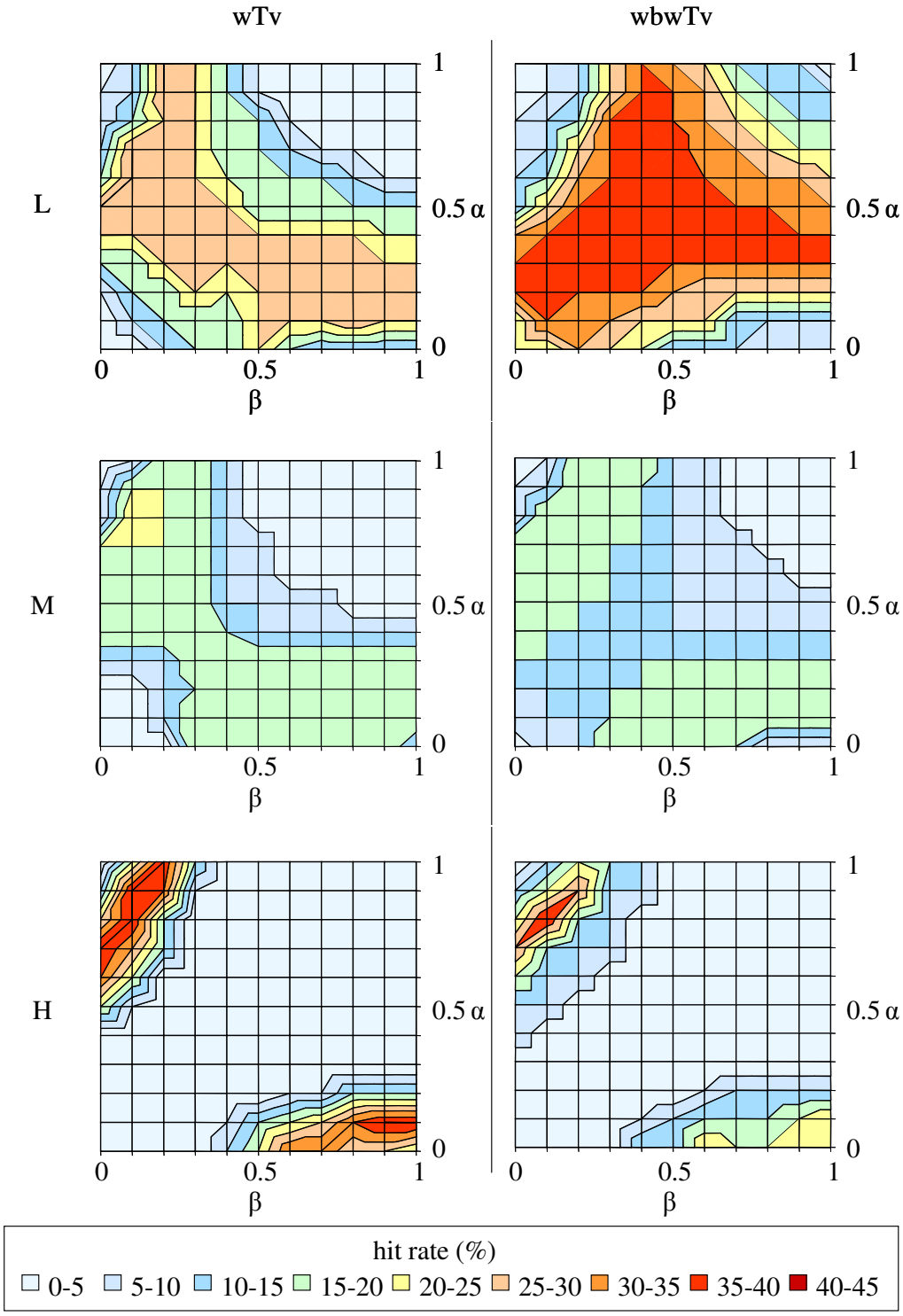


Figure B.3: Recovery rate landscapes (B). Shown are maps reporting search results for wTv and wbwTv calculations under systematic parameter variation using reference sets of different complexity for class MM1 against ZINC5K database.

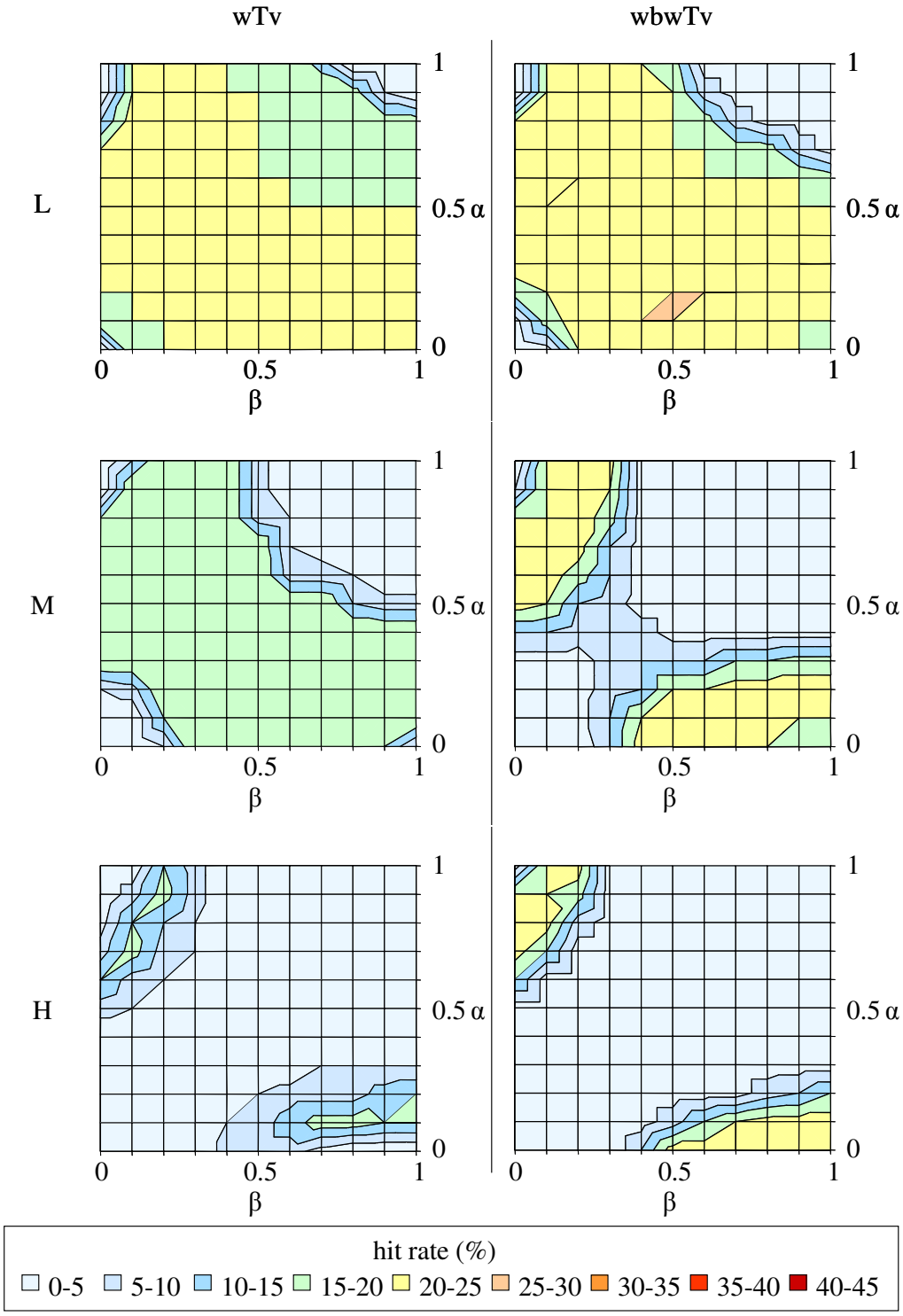


Figure B.4: Recovery rate landscapes (C). Shown are maps reporting search results for wTv and wbwTv calculations under systematic parameter variation using reference sets of different complexity for class SST against NCI database.

B.3 Shannon entropy-based similarity search strategy

Figure B.5 compares the cumulative recovery curves of four classes using Shannon entropy-based similarity search strategy and three other methods.

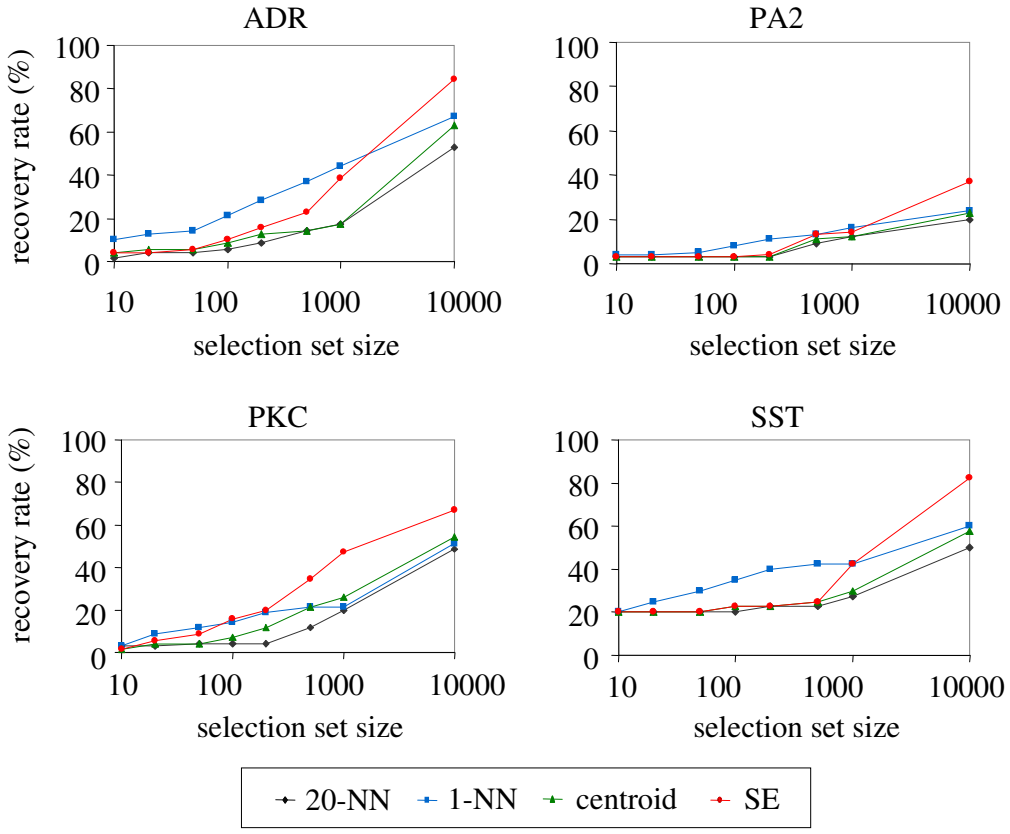


Figure B.5: Performance of Shannon entropy-based similarity searching. Recovery rates (in %) for the four different similarity search strategies – 20-NN (black), 1-NN (blue), centroid (green) and SE (red) – using the MACCS fingerprint and NCI database are compared for selection sets of increasing size (shown on a logarithmic scale).

Bibliography

[1] J. Bajorath. Integration of virtual and high-throughput screening. *Nature Reviews. Drug Discovery*, 1(11):882–894, 2002.

[2] A. R. Leach and V. J. Gillet. *An Introduction to Chemoinformatics*. Springer, October 2007. ISBN 1402062907.

[3] W. L. Jorgensen. The many roles of computation in drug discovery. *Science*, 303(5665):1813–1818, 2004.

[4] I. D. Kuntz. Structure-based strategies for drug design and discovery. *Science*, 257(5073):1078–1082, 1992.

[5] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Genetics*, 47(4):409–443, 2002.

[6] P. Willett, J. M. Barnard, and G. M. Downs. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996, 1998.

[7] P. Willett. Searching techniques for databases of two- and three-dimensional chemical structures. *Journal of Medicinal Chemistry*, 48(13): 4183–4199, 2005.

[8] J. Bajorath. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *Journal of Chemical Information and Computer Sciences*, 41(2):233–245, 2001.

[9] C. Williams. Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Molecular Diversity*, 10(3):311–332, 2006.

[10] D. R. Flower. On the properties of bit string-based measures of chemical similarity. *Journal of Chemical Information and Computer Sciences*, 38 (3):379–386, 1998.

[11] A. Schuffenhauer, P. Floersheim, P. Acklin, and E. Jacoby. Similarity metrics for ligands reflecting the similarity of the target proteins. *Journal of Chemical Information and Computer Sciences*, 43(2):391–405, 2003.

[12] R. D. Brown and Y. C. Martin. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *Journal of Chemical Information and Computer Sciences*, 36(3): 572–584, 1996.

[13] R. D. Brown and Y. C. Martin. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *Journal of Chemical Information and Computer Sciences*, 37(1):1–9, 1997.

[14] H. Eckert and J. Bajorath. Molecular similarity analysis in virtual screening: Foundations, limitations and novel approaches. *Drug Discovery Today*, 12(5-6):225–233, 2007.

[15] P. Willett. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, 11(23-24):1046 – 1053, 2006.

[16] *MACCS Structural keys*. Symyx Software, San Ramon, CA, USA, 2005. http://www.symyx.com.

[17] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, 2002.

[18] *BCI*. Digital Chemistry, Leeds, UK. http://www.digitalchemistry.co.uk.

[19] *Molecular Operating Environment*. Chemical Computing Group, Montreal, Quebec, Canada, 2007. http://www.chemcomp.com.

[20] *Daylight fingerprint*. Daylight Chemical Information Systems, Inc., Aliso Viejo, CA, USA. http://www.daylight.com.

[21] *Extended connectivity fingerprints, PiplinePilot 6.1*. Accelrys Inc., San Diego, CA, USA. http://accelrys.com/.

[22] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *Journal of Chemical Information and Modeling*, 49(1):108–119, 2009.

[23] C. M. R. Ginn, P. Willett, and J. Bradshow. Combination of molecular similarity measures using data fusion. *Perspectives in Drug Discovery and Design*, 84(4):327–352, 2000.

[24] J. D. Holliday, C-Y. Hu, and P. Willett. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Combinatorial Chemistry and High Throughput Screening*, 5: 155–166, 2002.

[25] N. Salim, J. Holliday, and P. Willett. Combination of fingerprint-based similarity coefficients using data fusion. *Journal of Chemical Information and Computer Sciences*, 43(2):435–442, 2003.

[26] J. D. Holliday, N. Salim, M. Whittle, and P. Willett. Analysis and display of the size dependence of chemical similarity coefficients. *Journal of Chemical Information and Computer Sciences*, 43(3):819–828, 2003.

[27] A. Tversky. Features of similarity. *Psychological Review*, 20(1):1–16, 1977.

[28] G. M. Maggiora and V. Shanmugasundaram. *Methods in Molecular Biology*, volume 275. Humana Press Inc. Totowa, NJ, 2004.

[29] J. Chen, J. Holliday, and J. Bradshaw. A machine learning approach to weighting schemes in the data fusion of similarity coefficients. *Journal of Chemical Information and Modeling*, 49(2):185–194, 2009.

[30] R. D. Brown and Y. C. Martin. An evaluation of structural descriptors and clustering methods for use in diversity selection. *SAR and QSAR in Environmental Research*, 8(1,2):23–39, 1998.

[31] Y. C. Martin, J. L. Kofron, and L. M. Traphagen. Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, 45(19):4350–4358, 2002.

[32] R. Taylor. Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. *Journal of Chemical Information and Computer Sciences*, 35(1):59–67, 1995.

[33] J. S. Delaney. Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Molecular Diversity*, 1(4):217–222, 1996.

[34] D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, and L. E. Weinberger. Neighborhood behavior: A useful concept for validation of molecular diversity descriptors. *Journal of Medicinal Chemistry*, 39(16): 3049–3059, 1996.

[35] S. L. Dixon and R. T. Koehler. The hidden component of size in two-dimensional fragment descriptors: Side effects on sampling in bioactive libraries. *Journal of Medicinal Chemistry*, 42(15):2887–2900, 1999.

[36] M. A. Johnson and G. M. Maggiora. *Concepts and Applications of Molecular Similarity*. Wiley-Interscience, 1st edition, 1990.

[37] F. L. Stahura and J. Bajorath. New methodologies for ligand-based virtual screening. *Current Pharmaceutical Design*, 11(9):1189–1202, 2005.

[38] *MDL Drug Data Report (MDDR)*. MDL Elsevier, San Leandro, 2005. http://www.symyx.com/.

[39] J. J. Irwin and B. K. Shoichet. ZINC – A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005.

[40] *NCI AIDS Antiviral Screen*. National Cancer Institute, 1999. http://dtp.nci.nih.gov/docs/aids/aids_data.html (accessed 01 Feb. 2007) The publicly available NCI anti-AIDS database contains structural and activity data for compounds screened by the AIDS antiviral screening program of the National Cancer Institute.

[41] L. Xue and J. Bajorath. Distribution of molecular scaffolds and R-groups isolated from large compound databases. *Journal of Molecular Modeling*, 5(5):97–102, 1999.

[42] A. Tovar, H. Eckert, and J. Bajorath. Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. *ChemMedChem*, 2(2):208–217, 2007.

[43] J. Hert, P. Willett, and D. J. Wilton. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *Journal of Chemical Information and Computer Sciences*, 44(3): 1177–1185, 2004.

[44] N. E. Shemetulskis, D. Weininger, C. J. Blankley, J. J. Yang, and C. Humblet. Stigmata: An algorithm to determine structural commonalities in diverse datasets. *Journal of Chemical Information and Computer Sciences*, 36(4):862–871, 1996.

[45] L. Xue, F. L. Stahura, J. W. Godden, and J. Bajorath. Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *Journal of Chemical Information and Computer Sciences*, 41(3):746–753, 2001.

[46] L. Xue, J. W. Godden, F. L. Stahura, and J. Bajorath. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *Journal of Chemical Information and Computer Sciences*, 43(4):1218–1225, 2003.

[47] L. Xue, F. L. Stahura, and J. Bajorath. Similarity search profiling reveals effects of fingerprint scaling in virtual screening. *Journal of Chemical Information and Computer Sciences*, 44(6):2032–2039, 2004.

[48] Y. Hu, E. Lounkine, and J. Bajorath. Improving the performance of extended connectivity fingerprints through activity-oriented feature filtering and application of a bit density-dependent similarity function. *ChemMedChem*, 4(4):540–548, 2009.

[49] H. Eckert and J. Bajorath. Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005.

[50] J. W. Godden, L. Xue, and J. Bajorath. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *Journal of Chemical Information and Computer Sciences*, 40(1):163–166, 2000.

[51] X. Chen and F. K. Brown. Asymmetry of chemical similarity. *ChemMedChem*, 2(2):180–182, 2007.

[52] M. A. Fligner, J. S. Verducci, and P. E. Blower. A modification of the Jaccard-Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics*, 44(2):110–119, 2002.

[53] M. S. Lajiness. Dissimilarity-based compound selection techniques. *Perspectives in Drug Discovery and Design*, 7(8):65–84, 1997.

[54] M. J. McGregor and P. V. Pallai. Clustering of large databases of compounds: using MDL 'keys' as structual descriptors. *Journal of Chemical Information and Computer Sciences*, 37(3):443–448, 1997.

[55] R. Natesh, S.L.U. Schwager, E.D. Sturrock, and K. R. Acharya. Crystal structure of the human angiotensin-converting enzyme–lisinopril complex. *Nature*, 421:551–554, 2003.

[56] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, 1948.

[57] G. J. Chaitin. Goedel's theorem and information. *International Journal of Theoretical Physics*, 21:941–954, 1982.

Eidesstattliche Erklärung

An Eides statt versichere ich hiermit, dass ich die Dissertation "Molecular Complexity Effects and Fingerprint-based Similarity Search Strategies" selbst und ohne jede unerlaubte Hilfe angefertigt habe, dass diese oder eine ähnliche Arbeit noch keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nachstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist:

Y. Wang, H. Eckert, and J. Bajorath. Apparent asymmetry in fingerprint similarity searching is a direct consequence of differences in bit densities and molecular size. *ChemMedChem*, 2(7):1037-1042, 2007.

Y. Wang and J. Bajorath. Balancing the influence of molecular complexity on fingerprint similarity searching. *Journal of Chemical Information and Modeling*, 48(1):75-84, 2008.

Y. Wang, H. Geppert, and J. Bajorath. Random reduction in fingerprint bit density improves compound recall in search calculations using complex reference molecules. *Chemical Biology and Drug Design*, 71(6):511-517, 2008.

Y. Wang and J. Bajorath. Bit silencing in fingerprints enables the derivation of compound class-directed similarity metrics. *Journal of Chemical Information and Modeling*, 48(9):1754-1759, 2008.

Y. Wang and J. Bajorath. Development of a compound class-directed similarity coefficient that accounts for molecular complexity effects in fingerprint searching. *Journal of Chemical Information and Modeling*, 49(6):1369-1376, 2009.

Y. Wang, H. Geppert, and J. Bajorath. Shannon entropy-based fingerprint similarity search strategy. *Journal of Chemical Information and Modeling*, 49(7):1687-1691, 2009.

Y. Wang and J. Bajorath. Advanced fingerprint methods for similarity searching: balancing molecular complexity effects. *Combinatorial Chemistry and High Throughput Screening*, in press.

Bonn, den 20 August 2009

(Yuan Wang)