

## Analysis of Neighborhood Behavior in Lead Optimization and Array Design

George Papadatos,<sup>†</sup> Anthony W. J. Cooper,<sup>‡</sup> Visakan Kadirkamanathan,<sup>§</sup> Simon J. F. Macdonald,<sup>‡</sup>  
Iain M. McLay,<sup>‡</sup> Stephen D. Pickett,<sup>‡</sup> John M. Pritchard,<sup>‡</sup> Peter Willett,<sup>†</sup> and Valerie J. Gillet<sup>\*†</sup>

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield,  
211 Portobello Street, Sheffield S1 4DP, United Kingdom GlaxoSmithKline, Medicines Research Centre,  
Gunnels Wood Road, Stevenage SG1 2NY, United Kingdom, and Department of Automatic Control and  
Systems Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, United Kingdom

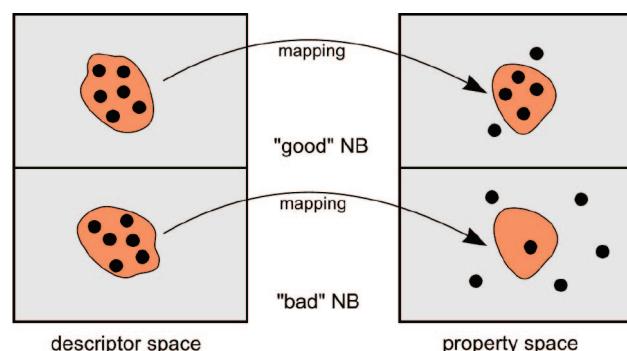
Received August 28, 2008

Neighborhood behavior describes the extent to which small structural changes defined by a molecular descriptor are likely to lead to small property changes. This study evaluates two methods for the quantification of neighborhood behavior: the optimal diagonal method of Patterson et al. and the optimality criterion method of Horvath and Jeandenans. The methods are evaluated using twelve different types of fingerprint (both 2D and 3D) with screening data derived from several lead optimization projects at GlaxoSmithKline. The principal focus of the work is the design of chemical arrays during lead optimization, and the study hence considers not only biological activity but also important drug properties such as metabolic stability, permeability, and lipophilicity. Evidence is provided to suggest that the optimality criterion method may provide a better quantitative description of neighborhood behavior than the optimal diagonal method.

### INTRODUCTION

The *similar property principle*<sup>1</sup> states that structurally similar molecules tend to exhibit similar properties and underlies many techniques in chemoinformatics such as virtual screening and QSAR and ligand binding as well as the design of corporate screening collections.<sup>2,3</sup> There are several exceptions to the Principle:<sup>4,5</sup> even so, if it was not of general applicability, then it would be very difficult to attempt the development of systematic approaches to the identification of novel bioactive molecules, and there is a large, and increasing, body of evidence to support its use in drug discovery programs.<sup>6–9</sup>

Barbosa and Horvath<sup>10</sup> note that the Principle has two major implications. First, if a molecule is known to be active against some biological target, then molecules that are structurally similar to the chosen molecule are likely to exhibit the same activity; second, it is possible to predict the properties of novel molecules, given a list of structurally similar molecules for which the requisite property data are already available. A further, closely related concept is the *neighborhood principle*,<sup>11,12</sup> which states that molecules within the same local region, or *neighborhood*, of structural space tend to include more molecules with similar values of some desired property (usually biological activity) than would be expected in any other randomly selected region of the same size. Use of the word “similar” implies some way in which structural similarity can be measured, and there are two principal components of any similarity measure: a descriptor (or set of descriptors) by which molecules may be represented and a similarity coefficient to quantify the



degree of resemblance between two such representations. In this paper, we focus on molecular descriptors, specifically on the extent to which different types of descriptor are able to satisfy the neighborhood principle, an ability that is referred to as *neighborhood behavior*.<sup>11</sup> Informally, neighborhood behavior can be regarded as the extent to which a structural space can be mapped onto the property/biological activity space in such a way that neighboring points in the former are likely to correspond to neighboring points in the latter (as shown in Figure 1). Different descriptors will exhibit different levels of neighborhood behavior, and it is thus possible to compare the effectiveness of different types of structural representation by the extent to which they are able to map successfully between structural space and property space.

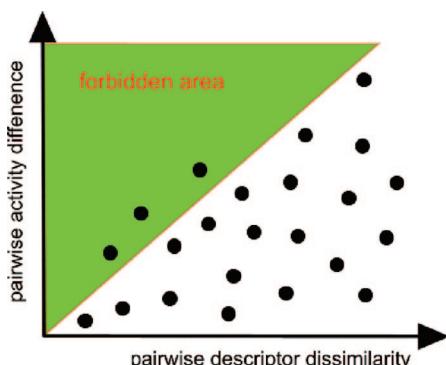
We focus here on the design of arrays to support lead optimization, an application that seems to be very well suited to the neighborhood behavior concept. Lead optimization involves synthesizing and testing hundreds or thousands of structural analogues, in an effort to find the optimal combination of desired criteria such as potency, selectivity, lipophilicity, and solubility. Optimization is increasingly

\* Corresponding author e-mail: v.gillet@sheffield.ac.uk.

<sup>†</sup> Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield.

<sup>‡</sup> GlaxoSmithKline.

<sup>§</sup> Department of Automatic Control and Systems Engineering, University of Sheffield.



**Figure 2.** The Patterson or Neighborhood plot. Adapted from ref 14.

carried out using arrays of molecules. An initial array, based on the original lead molecule, is synthesized, purified, and tested, and then the most promising member of the array (i.e., the one that best fulfills the chosen criteria) is used as a starting point, or *seed*, for the design of the next array. This seed will be structurally modified, typically by altering one or two R-groups, with the hope that at least one member of the new array will be able to serve as the seed for a subsequent array. This is a simple application of the similar property principle, with small variations in structure being carried out to enable a systematic exploration of the structural space surrounding the current seed molecule. The exploration will be effective only if the descriptor used to determine the coverage of structural space does exhibit neighborhood behavior; otherwise the systematic exploration will rapidly degenerate into a random search.

The paper is organized as follows. The next section discusses previous studies of neighborhood behavior, focusing on two methods—the optimal diagonal method of Patterson et al.<sup>11</sup> and the optimality criterion method of Horvath and Jeandenans<sup>13</sup>—that provide a quantitative basis for assessing the extent to which a descriptor does indeed exhibit neighborhood behavior. We then report the results of an extended set of experiments that use array data generated in lead optimization and array design programs at GlaxoSmithKline. The neighborhood behavior of several widely used 2D and 3D fingerprints is evaluated using the two chosen methods, and we provide evidence to suggest that the optimality criterion method may provide a better quantitative description of neighborhood behavior than the optimal diagonal method.

#### PREVIOUS STUDIES OF NEIGHBORHOOD BEHAVIOR

The neighborhood behavior concept was first presented in the much-cited paper by Patterson et al.<sup>11</sup> Here, 20 small QSAR data sets were characterized by different molecular descriptors, including 2D fingerprints, ClogP, connectivity indices, topomeric steric field-based descriptors, and random numbers (as a blind test). The dissimilarities were computed between each pair of molecules in a data set, in terms of both structure (as characterized by the Tanimoto coefficient or Euclidean distance for the chosen descriptor) and activity (as characterized by the absolute difference in logIC50). The resulting Patterson or neighborhood plot reflects the neighborhood behavior of the descriptor, as shown in Figure 2. An effective descriptor, i.e., one that exhibits neighborhood

behavior as illustrated in the upper part of Figure 1, will yield a plot in which pairs of molecules with small structural dissimilarities (i.e., similar as defined by the chosen descriptor) have small differences in activity; an ineffective descriptor will have many data points in the *forbidden area*, corresponding to pairs of molecules that have large activity differences but small structural dissimilarities. Patterson et al. defined the *neighborhood enhancement* (or NBE) of a descriptor, this quantifying the extent of the neighborhood behavior exhibited by that descriptor, and described a procedure based on  $\chi^2$  to test the statistical significance of the resulting NBE (as described in detail in the section Experimental Methods). A modification of this test was used in a study of ACE inhibitors characterized by UNITY fingerprints, molecular steric fields, WHIM indices, pharmacophore atom-pairs, spatial autocorrelation functions, and, for comparison, molecular weight and random numbers.<sup>15</sup> The study concluded that UNITY fingerprints and the pharmacophore atom-pairs exhibited the best neighborhood behavior.

The studies of Patterson et al. and Matter suggested that a good descriptor would be one that had most of the data points in the lower-right part of a Patterson plot, i.e., below the diagonal in Figure 2. Subsequently, Dixon and Merz suggested that neighborhood behavior should be associated with a distinct pattern of progressively wider ranges in the biological activity differences at increasing dissimilarity scores, i.e., a “gradual fanning out” of points as one moves to the right of a Patterson plot.<sup>16</sup> They also showed that the  $\chi^2$  test used by Patterson et al. tends to overestimate any genuine neighborhood behavior, markedly in some cases, and proposed an adjusted  $\chi^2$  test based on repeated randomization of the activity values in the data set being analyzed. Finally, Dixon and Merz also advocated the use of the correlation coefficient *r* as an alternative indicator of the strength of the relationship between intermolecular activity differences and structural dissimilarity.

More recently, Horvath and Jeandenans have discussed neighborhood behavior in the context of *activity profiles*, where molecules have multiple activity values associated with them.<sup>13,17</sup> Specifically, an activity profile is a consistent set of experimental activity values for a molecule when measured against a panel of different biological assays. A profile is hence a vector that describes the location of a molecule in activity space, and the dissimilarity between pairs of these vectors can be measured using Euclidean or Dice distances (see also work by Fliri et al.<sup>18</sup> and by Schuffenhauer et al.<sup>19</sup>). Horvath and Jeandenans described two parameters that quantify the neighborhood behavior of a descriptor. The *consistency criterion* ( $\chi$ ) is the propensity of a similarity measure to selectively rank pairs of activity-related molecules among the structurally most similar pairs; the *overall optimality criterion* ( $\Omega$ ) is the fraction of activity-related pairs that rank among the structurally most similar pairs. These two parameters, consistency and optimality, are analogous to the precision and recall parameters that are used to evaluate text-retrieval systems and that have been studied for chemoinformatics applications by Edgar et al.<sup>20</sup>

The performance of different similarity measures can be assessed and compared by generating  $\Omega-\chi$  plots, where good neighborhood behavior is associated with high optimality values at high consistency values, and this approach was

followed in a study of 584 commercially available drugs and reference compounds screened against a panel of 42 assays.<sup>17</sup> A large number of 2D and 3D topological and structural descriptors were generated, with the Euclidean and Dice distance coefficients being used to measure structural dissimilarity. FBPA (Fuzzy Bipolar Pharmacophore Autocorrelogram) descriptors were found to exhibit the best neighborhood behavior, closely followed by global 2D and 3D descriptors, with the Dice coefficient performing noticeably better than the Euclidean distance at measuring intermolecular dissimilarity. A subsequent study of FBPA descriptors compared the extent of the relationship between neighborhood behavior and clustering behavior.<sup>21</sup>

Most neighborhood behavior studies have used relatively small, and often homogeneous, data sets. However, Perekhodtsev used four large, structurally diverse sets of serine protease inhibitors in a neighborhood behavior study to validate the use of standard 2D similarities (Accelrys Accord fingerprints and the Tanimoto coefficient) for the prediction of binding free energies (both measured and computed).<sup>22</sup> Analyses based on Patterson plots and on the correlation coefficient *r* (following Dixon and Merz<sup>16</sup>) indicated that all four data sets demonstrate neighborhood behavior, i.e., that structurally similar compounds tend to exhibit more similar binding affinities than do those that are less structurally similar. Perekhodtsev's focus on the data set, rather than on the descriptors used to characterize molecules within it, mirrors recent work that seeks to quantify and classify the nature of the underlying structure–activity relationships. Methods such as the SAR Index (SARI)<sup>23</sup> and the Structure–Activity Landscape Index (SALI)<sup>24</sup> are based on simple formulas that involve the direct calculation of activity distances  $\Delta A$  and structural dissimilarities  $\Delta S$  among molecule-pairs and are thus directly related to neighborhood behavior; for example, SALI (eq 1) is the slope of the diagonal between a given point (i,j) and the origin of a Patterson plot (Figure 2).

$$\text{SALI}_{ij} = \frac{|\Delta A_{ij}|}{\Delta S_{ij}} \quad (1)$$

## EXPERIMENTAL METHODS

**Data Sets and Descriptors.** Data from three GSK lead optimization projects (Projects I–III) were used for our experiments. The Project I data set contained data for 2331 molecules, with IC<sub>50</sub> values against two protein targets (Target 1 and Target 2). The Project II data set contained data for 2971 molecules, including IC<sub>50</sub> values against three protein targets (Targets 3–5) as well as measurements of lipophilicity, membrane permeability, and metabolic stability. Finally, the Project III data set contained 3286 molecules with IC<sub>50</sub> values against one target (Target 6).

An extensive series of preprocessing routines was applied to the raw project data. Duplicate entries were merged, and the associated property values averaged. Entries with obsolete or obviously mistaken values or values with modifiers (i.e., '<' or '>') were excluded, as were singleton compounds that had not been synthesized via an array and salts, mixtures, and stereoisomers. Low-potency compounds (*pIC*<sub>50</sub> < 5.5) were excluded from the analyses involving bioactivity. The remaining compounds were then passed through substructure

**Table 1.** Numbers of Molecules in the Data Sets and Chemotypes Used in the Experiments

chemotype	Project I		
	Target 1	Target 2	
1	355	488	
2	183	149	
3	162	159	
4	35	124	
total	735	920	
Project II			
chemotype	Target 3	Target 4	Target 5
	175	179	146
5	322	320	288
6	892	901	655
total	1389	1400	1089
Project III			
chemotype	permeability	metabolic stability	lipophilicity
	32	87	43
5	53	233	147
6	177	482	183
total	262	802	373
chemotype	Project III Target 6		
	8	222	
9	146	368	
total			

filters that classified them into intuitive chemical families/chemotypes that could be used in SAR tables. Each such chemotype subset has a common core with variations in up to three attached R-groups and contains molecules from between five and 90 different arrays. The makeup of the processed data sets is summarized in Table 1, where it will be seen that the data sets range in size from 32 (Project II/Permeability/Chemotype 5) to 901 (Project II/Target 4/Chemotype 7) molecules.

The molecules in the various data sets were characterized by the twelve different fingerprints summarized in Table 2. These are of three major types: structural keys, hashed substructures, and pharmacophore fingerprints.<sup>28</sup>

Dictionary-based fingerprints were exemplified by the 166-bit MDL Public Keys, which were generated using the Pipeline Pilot software package.

Hashed fingerprints were exemplified by the use of both path and circular fragment definitions. Daylight fingerprints were generated using paths of length 0–7 bonds, and the resulting substructures were hashed into a bit-string of length 2048 bits. SciTegic's path fingerprints (EPFP) were generated using paths of length 0–8 bonds. EPFPs are similar to Daylight fingerprints but use a different coding scheme for the constituent atoms: the Daylight paths use elemental types, whereas the SciTegic paths use the same codes as the ECFP circular substructures described below. Three types of circular fragment were used for fingerprint generation: the SciTegic Extended Connectivity Fingerprints (ECFP) and Functional Connectivity Fingerprints (FCFP) and the MOLPRINT 2D Atom Environments.<sup>29</sup> The initial code assigned to an atom for ECFPs is based on the number of connections, the element type, the charge, and the mass; in the case of the FCFPs, a more functional atom code definition is used, namely hydrogen-bond donor (HBD), hydrogen-bond ac-

**Table 2.** Summary of Descriptors and Programs Used To Generate Them

type of fingerprint	name	abbreviation	program used
2D structural keys	MDL Public Keys	MDLPublicKeys	Pipeline Pilot <sup>25</sup>
2D path substructures	Daylight	Daylight	Daylight <sup>26</sup>
	EPFP	EPFP_8	Pipeline Pilot
2D circular substructures	Extended Connectivity	ECFP_4 and _6	Pipeline Pilot
	Functional Class	FCFP_4 and _6	Pipeline Pilot
	Atom Environments	SEFP_4 and _6	Pipeline Pilot
2D pharmacophores	GSK Topological Pharmacophores	TopPharm	in-house
3D pharmacophores	3-point	piDAPH3	MOE <sup>27</sup>
	4-point	piDAPH4	MOE

ceptor (HBA), positively ionizable, negatively ionizable, aromatic, and halogen. The experiments here used the ECFP\_4 ECFP\_6, FCFP\_4 and FCFP\_6 fingerprints, where the numeric code denotes the diameter in bonds up to which features are generated. Atom Environments are also based on circular substructures but differ from their SciTegic counterparts in two major ways: they use force field (mol2) atom types and encode fragments of a given diameter, whereas SciTegic fingerprints include features up to a given diameter. The Atom Environments used here were implemented in Pipeline Pilot as SEFP\_4 and SEFP\_6. ECFP, FCFP, and SEFP fingerprints are hashed by SciTegic's hashing algorithm to create a fingerprint of a virtual size of  $2^{32}$  features.

Pharmacophore fingerprints were exemplified by the use of both 2D and 3D definitions. Topological pharmacophores were generated using a GSK in-house program called TopPharm. The method is similar to the CATS approach,<sup>30</sup> atoms are encoded as pharmacophore features (aromatic HBA, aliphatic HBA, HBD, aromatic, positively and negatively ionizable, and linker) with the distance between pairs of features given as the separation in terms of numbers of bonds. The resulting keys are then mapped into a bit-string of length 1024 bits. The 3D fingerprints were generated from single low-energy conformations produced by the MOE software package, with atoms encoded by one of eight atom types computed from three atomic properties ("in pi system", "is donor", "is acceptor"). 3-Point and 4-point pharmacophores were then generated using sets of atoms and the associated interatomic distances, with the resulting keys being hashed into a bit-string of virtual sizes of  $2^{18}$  and  $2^{31}$  features, respectively.

We have analyzed our experiments in two ways: using the optimal diagonal (Patterson plot) approach first described by Patterson et al.<sup>11</sup> and then amended by Dixon and Merz<sup>16</sup> and using the optimality criterion method first described by Horvath and Jeandenans.<sup>13</sup> These two approaches are detailed below.

**The Optimal Diagonal Method.** The Patterson plot provides a simple visual representation of the neighborhood behavior of a descriptor when applied to a particular data set. The plot can be analyzed to provide a quantitative representation of the neighborhood behavior by means of the following construction.

Given a descriptor and a data set, the dissimilarity values  $D(i,j)$  between every unique pair of molecules i and j in the data set ( $x$ -axis) are calculated and plotted against the pairwise absolute differences in biological activities ( $y$ -axis). The structural dissimilarity  $D$  is defined as the complement to the Tanimoto similarity value, i.e.,  $D(i,j)=1-T_c(i,j)$ . For a

data set of  $N$  compounds, the Patterson plot will hence contain  $N(N-1)/2$  data points.

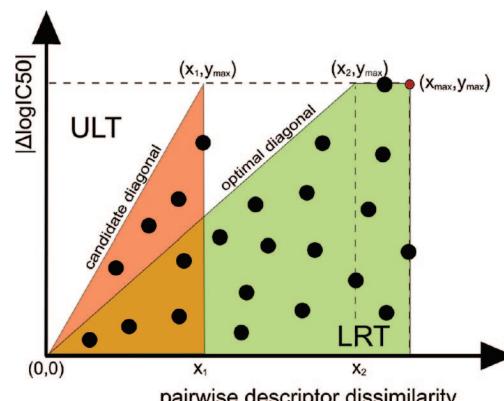
Each data point  $(x,y)$  on the Patterson plot defines a candidate triangle with vertices  $(0,0)$ ,  $(x,0)$ , and  $(x,y_{\max})$  and a candidate diagonal with vertices  $(0,0)$  and  $(x,y_{\max})$ , where  $y_{\max}$  is the maximum ordinate value, as shown in Figure 3. The *optimal diagonal* is taken to be that diagonal which maximizes the density of points PD below it, i.e., the number of points  $N_{LRT}$ , divided by the area A that contains them.

$$PD_{LRT} = \frac{N_{LRT}}{A} \quad (2)$$

The NBE is then defined as the ratio of the point density PD under the optimal diagonal (Lower Right Trapezoid, LRT) to the point density of the whole rectangular plot area (LRT+ULT) (see eq 3). The maximum score is observed when all the data points lie in the lower right triangular half of the plot, in which case NBE=2.0.

$$NBE = \frac{PD_{LRT}}{PD_{LRT+ULT}} \quad (3)$$

The validity of the optimal diagonal method has been assessed using a  $\chi^2$  test. The activity values are scrambled among the data set members 200 times (y-randomization), a Patterson plot is generated each time, and a count made of the number of scrambled points that fall under the optimal diagonal ( $\hat{N}_{LRT}$ ), this serving as the expected value for the statistical test. This value is then compared to the number of points that actually lie in the LRT ( $N_{LRT}$ , observed value) and  $\chi^2$  calculated as



**Figure 3.** Computing the diagonal in a Patterson plot (ULT: upper left triangle; LRT: lower right trapezoid).

$$\chi^2 = \frac{(N_{LRT} - \hat{N}_{LRT})^2}{\hat{N}_{LRT}} \quad (4)$$

The computed value is significant at the 0.05 level of statistical significance if  $\chi^2 \geq 3.84$  (one degree of freedom).

It should also be noted here that, contrary to Dixon and Merz, the use of the correlation coefficient  $r$  as an additional indicator for neighborhood behavior was not considered since it is simply an indication of linearity and not of actual neighborhood behavior.

**The Optimality Criterion Method.** The optimality criterion method due to Horvath and Jeandenans also employs the sets of  $N(N-1)/2$  pairs of structural and property dissimilarities used in the optimal diagonal approach. Each such pair is then assigned to one of four categories as follows:

- Pairs that are both structurally similar (at a user-defined similarity threshold  $s$ ) and activity-related (at a user-defined activity difference threshold  $l$ ) confirm the neighborhood principle and are called “true similar” (TS) pairs.

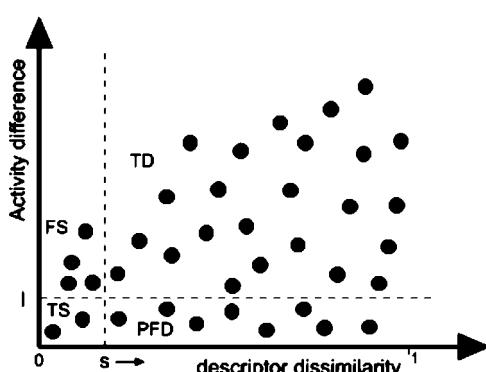
- Pairs that are structurally similar but not activity-related violate, by definition, the neighborhood principle and are called “false similar” (FS) pairs.

- Pairs that are neither structurally similar nor activity-related are denoted as “true dissimilar” (TD) pairs.

- Pairs that are structurally dissimilar but activity-related are called “potentially false dissimilar” (PFD) pairs. This includes both genuinely dissimilar but activity-related pairs as well as activity-pairs for which the descriptor/similarity metric overestimated their dissimilarity.

The application of the thresholds  $l$  and  $s$  hence divides the plot into four rectangular subsections, one for each different type of pair, as shown in Figure 4 with the analogous contingency table shown in Figure 5.

As the dissimilarity threshold  $s$  moves from 0 to 1, consistency decreases (more FS pairs) in favor of increasing completeness (more TS pairs). There is hence a tradeoff between consistency and completeness, and the role of the optimality



**Figure 4.** The optimality criterion classification scheme.  $l$  and  $s$  are the activity and similarity thresholds, respectively.

		Structural similarity	
		Similar ( $\Delta S < s$ )	Dissimilar ( $\Delta S > s$ )
Activity similarity	Similar ( $\Delta Act < l$ )	$N_{TS}$	$N_{PFD}$
	Dissimilar ( $\Delta Act > l$ )	$N_{FS}$	$N_{TD}$

**Figure 5.** Contingency table, illustrating the concordant (TS and TD) and nonconcordant molecule-pairs (FS and PFD).

coefficient is to identify the best experimental compromise between these two conflicting criteria. The tradeoff is defined by monitoring the sum of the number of FS pairs ( $N_{FS}$ ) and of the number of PFD pairs ( $N_{PFD}$ ) at various similarity cut-offs,  $s$ , as compared to the sum that would be expected by mere chance. Specifically, the optimality coefficient is defined by eq 5, where  $k$  is a constant penalty weight

$$\Omega(s) = \frac{kN_{FS} + N_{PFD}}{kN_{FS}^{\text{expected}} + N_{PFD}^{\text{expected}}} \quad (5)$$

In an operational environment, the FS pairs in a neighborhood plot are far more important than the PFD pairs since the former can lead to a waste of time and resources as false positives are synthesized and tested. It is for this reason that the penalty weight  $k$  is included in eq 5.

The expected values can be computed from the contingency table. Thus

$$N_{FS}^{\text{expected}} = \frac{(N_{TS} + N_{FS})(N_{TD} + N_{FS})}{N} \quad (6)$$

and

$$N_{PFD}^{\text{expected}} = \frac{(N_{TS} + N_{PFD})(N_{TD} + N_{PFD})}{N} \quad (7)$$

(where  $N$  is the total number of molecules, i.e.,  $N_{TS} + N_{PFD} + N_{FS} + N_{TD}$ ).

The lower-bound for  $\Omega$  (best possible) occurs when both  $N_{FS}$  and  $N_{PFD}$  are zero, in which case  $\Omega$  is zero. The upper-bound (worst possible) occurs when both  $N_{TS}$  and  $N_{TD}$  are zero, in which case  $\Omega$  is given by

$$\Omega(s) = \frac{N(kN_{FS} + N_{PFD})}{kN_{FS}^2 + N_{PFD}^2} \quad (8)$$

In practice, the minimal value will be greater than the lower-bound of zero; this minimal value can be identified by varying the similarity threshold  $s$  and observing the effect on  $\Omega$ . The value of  $s$  that minimizes  $\Omega$ ,  $s^*$ , then defines the *neighborhood radius*. In similar vein, the maximal value will, in practice, be less than 1 (which would occur if the observed distribution of FS and PFD pairs was that expected by chance, i.e., if the descriptor's neighborhood behavior was no better than random).

## RESULTS AND DISCUSSION

**Optimal Diagonal Method.** The optimal diagonal method has two measures to quantify the appropriateness of the Patterson plots: the NBE score and  $\chi^2$ .

If a high value for the NBE is to be achieved, then the data points in the Patterson plot should fall predominantly into a trapezoidal or (preferably) a triangular shape; conversely, a steep diagonal and/or the existence of data points outside the ULT area will generally result in low values for the NBE. This is illustrated by the NBE scores in the set of Patterson plots shown in Figure 6 for the Project II/Target 3/Chemotype 7 data. In this figure (and subsequently), the  $N(N-1)/2$  dissimilarity points are shown in green, and the optimal diagonal (computed using MATLAB) is shown in red. The optimal diagonal algorithm has resulted in relatively low NBE scores for several of the descriptors, e.g., MDL

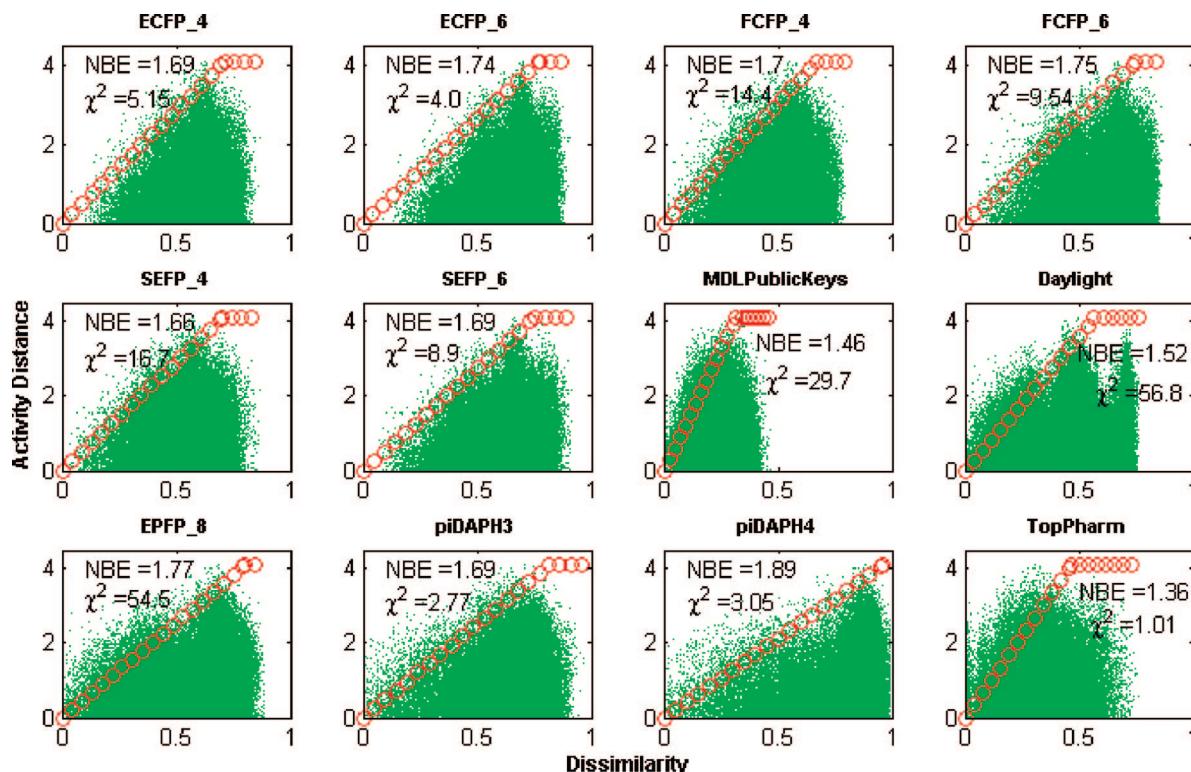


Figure 6. Optimal diagonal plots for the Project II/Target 3/Chemotype 7 data.

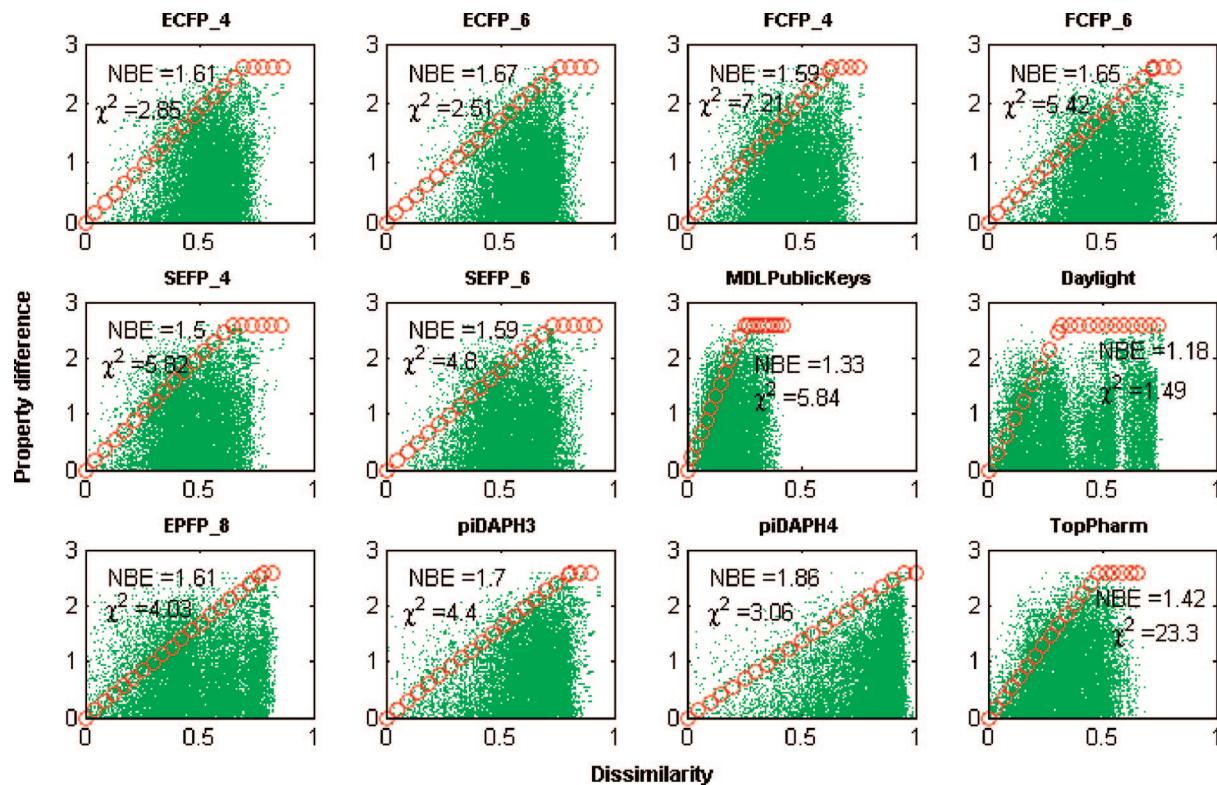


Figure 7. Optimal diagonal plots for the Project II/Permeability/Chemotype 7 data.

Public Keys, TopPharm, and Daylight, with scores of 1.46, 1.36, and 1.52, respectively. This mirrors the plots directly, since they are more bell-shaped than the favored triangular or trapezoidal shape. Conversely, the shapes of, e.g., the ECFP\_6, FCFP\_6, and piDAPH4 plots result in higher NBE scores (1.74, 1.75, and 1.89, respectively). Similar comments apply to plots for the Project II/Permeability/Chemotype 7 data shown in Figure 7 where, e.g., the near-rectangular shape

of the Daylight plot results in a (near-random) NBE score of 1.18.

If a high value for  $\chi^2$  is to be achieved, then the Patterson plot needs data points at both the lower left and upper right corners of the plot, with a gradual fanning out,<sup>16</sup> conversely, significant point densities in the upper left and/or lower right corners of the plot will generally result in low values for  $\chi^2$ . Two extreme cases of this behavior are shown in Figures 8

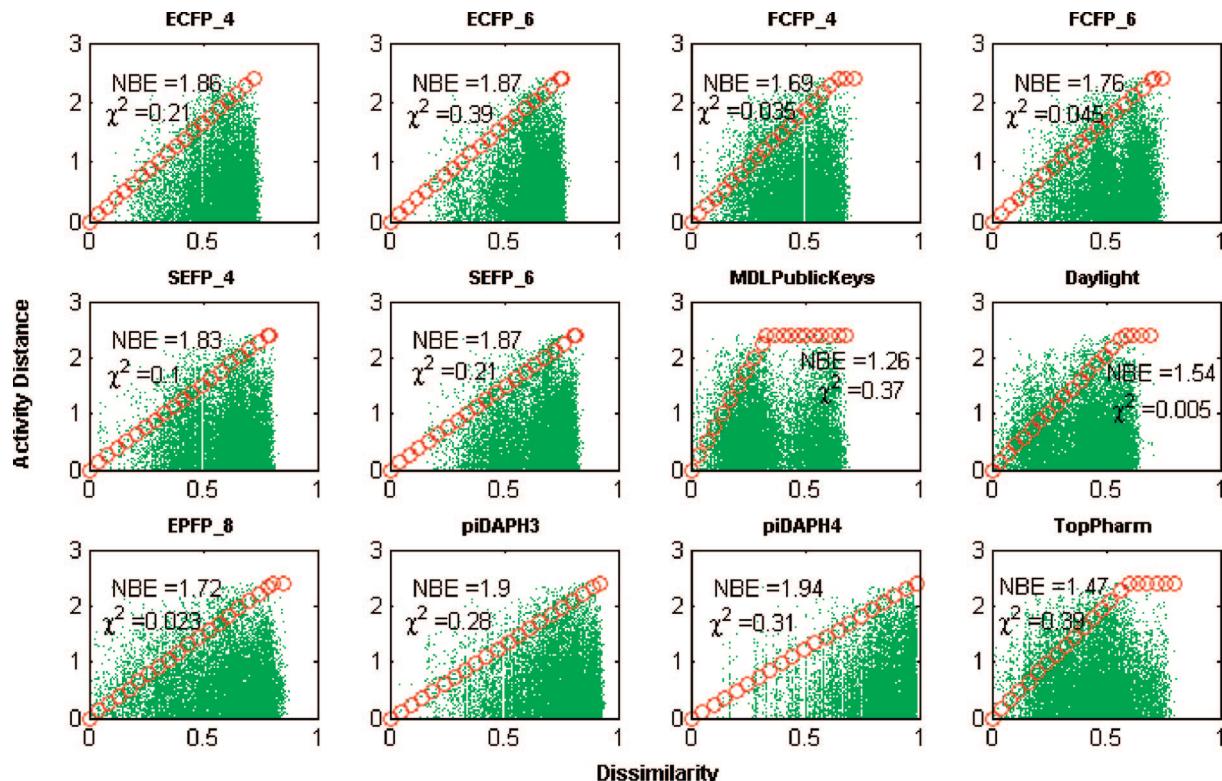


Figure 8. Optimal diagonal plots for the Project I/Target 1/Chemotype 3 data.

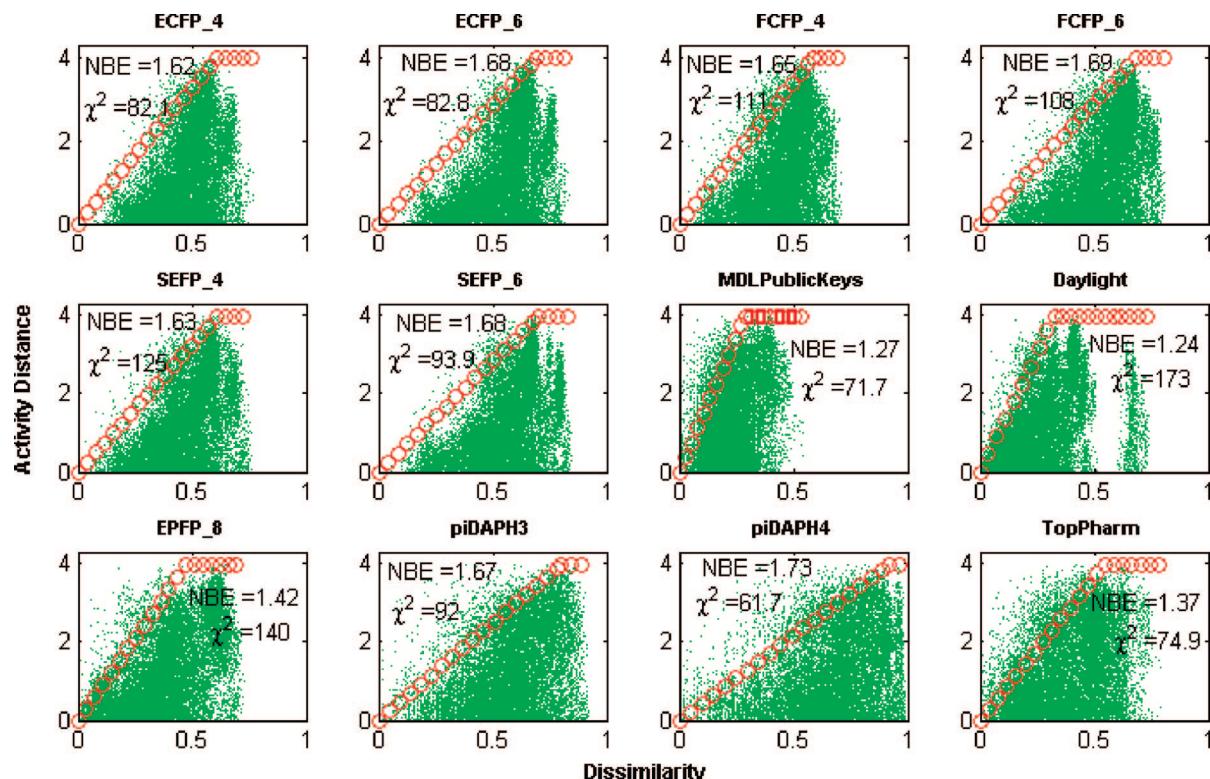


Figure 9. Optimal diagonal plots for the Project II/Target 3/Chemotype 5 data.

(Project I/Target 1/Chemotype 3) and Figure 9 (Project II/Target 3/Chemotype 5). All 12 descriptors in Figure 8 achieve low  $\chi^2$  scores (Daylight, EPFP\_8, FCFP\_4, and FCFP\_6 particularly so), reflecting the absence of significant point density in the lower-left corner of the plot, and/or the existence of a large point density in the lower-right corner of the plot. Conversely, all 12 plots in Figure 9 have high

$\chi^2$  scores, owing to the presence of the large density of points in both the lower-left and upper-right corners of the plots (and also sometimes to a low point density in the lower-right corner).

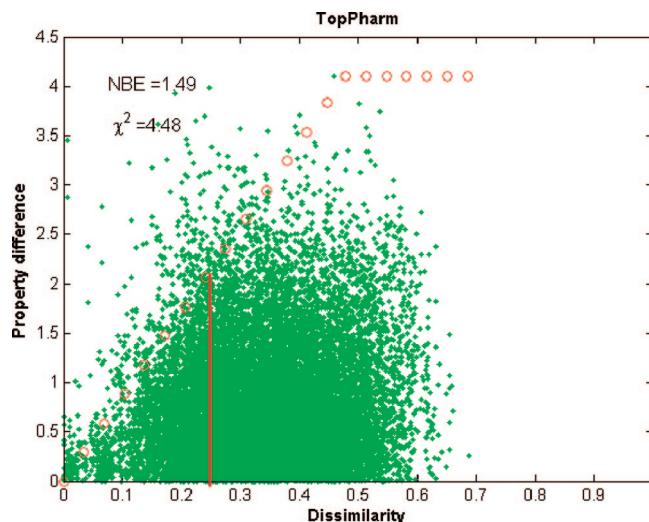
When presenting their modified version of the test, Dixon and Merz note that “even a moderate level of statistical significance still implies an advantage over random screen-

ing". This is surely correct, but is it also correct then to use the magnitude of significance values to compare the performance of different descriptors? For example, in Figure 6, do the FCFP\_4 fingerprints, with  $\chi^2=14.4$ , exhibit better neighborhood behavior than the FCFP\_6 fingerprints, with  $\chi^2=9.54$ ? Furthermore, the latter plot is better (by eye) than that for the piDAPH4 fingerprints, with  $\chi^2=3.05$  which is less than the critical value of 3.84 for the 0.05 level of statistical significance: we can hence conclude that there is no significant neighborhood behavior for piDAPH4. This being so, it is rather disconcerting to find that the piDAPH4 fingerprints have a higher NBE than for any other descriptor in Figure 6.

Using the terminology of Horvath and Jeandennans,  $\chi^2$  will give higher scores if there is a significant fraction of true similar and/or true dissimilar molecule-pairs in the Patterson plot, i.e., molecule-pairs that obey the similar property principle and hence the neighborhood principle. Conversely,  $\chi^2$  will give lower scores if there is a significant fraction of false similar or false dissimilar molecule-pairs in the plot, i.e., molecule-pairs that do not obey the similar property principle and hence the neighborhood principle. Thus, the consistently low  $\chi^2$  values in Figure 8 could indicate that the underlying SAR for this chemotype resembles the "the sharp cliffs of Bryce Canyon", to use Maggiora's phrase, whereas the consistently high  $\chi^2$  values in Figure 9 could indicate a "smooth hills of Kansas" SAR.<sup>31</sup>

The optimal diagonal method does have some limitations. Some of these were first discussed by Clark et al.<sup>32</sup> in a follow-up study to the original Patterson paper. In an attempt to overcome misleading results, the authors proposed three key modifications. These were (1) the use in the algorithm of the rank transformation and the lower right triangle rather than the trapezoid, (2) the substitution of the NBE statistic with an enrichment ratio, and (3) the use of bootstrapping to calculate the expected values. Here we decided to implement the last of these modifications, but in order to maintain comparability with earlier published studies the first two modifications were not implemented. Rather, the commonly used protocols of Patterson and Dixon and Merz were employed.

Regarding the limitations of optimal diagonal method, first, we emphasize the fact that  $\chi^2$  and NBE scores are not necessarily in accord with each other; indeed, Figures 8 and 9 show that they can yield diametrically opposed conclusions as to the extent of the neighborhood behavior in a plot. Thus, with the sole exception of the MDL Public Keys (where the two values are almost identical), all of the NBE scores in Figure 8 are greater than those in Figure 9, whereas all of the  $\chi^2$  scores are (very much) lower: indeed, none are statistically significant at the 0.05 level. Furthermore, no significant correlation was found between the average NBE and  $\chi^2$  scores across all the 12 descriptors. While this is not necessarily surprising, the use of both  $\chi^2$  and NBE scores introduces an extra ambiguity when it comes to ranking the descriptors based on their performance, as discussed earlier. Second, the method considers all of the LRT points when computing the NBE score. This can underestimate the true neighborhood behavior performance of a descriptor in the case of a large data set containing mainly dissimilar pairs of molecules. For example, the plot in Figure 10 yields a moderate score of 1.49, despite the strong neighborhood behavior up to a dissimilarity of about 0.25: in such cases,



**Figure 10.** Optimal diagonal plot for the TopPharm descriptor on the Project II/Lipophilicity/Chemotype 7 data.

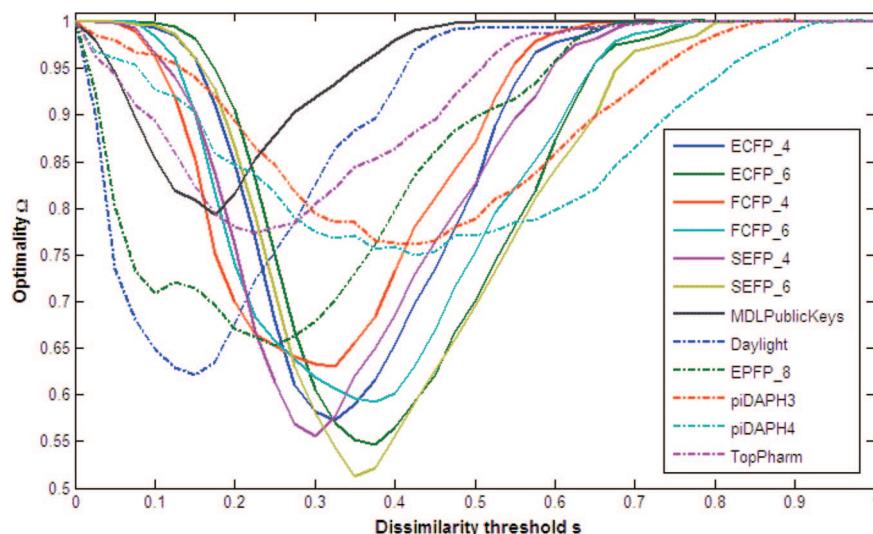
**Table 3.** Endpoints and the Associated Distance Thresholds Used in the Optimality Criterion Method

property	unit	distance	threshold l
potency	pIC50	Euclidean $ \Delta pIC50 $	0.5
lipophilicity	$\log D_{pH7.4}$	Euclidean $ \Delta \log D $	0.5
permeability	nm/s	Euclidean $ \Delta \log Perml $	0.5
metabolic stability	% remaining	Euclidean $ \Delta \log Stabl $	0.3

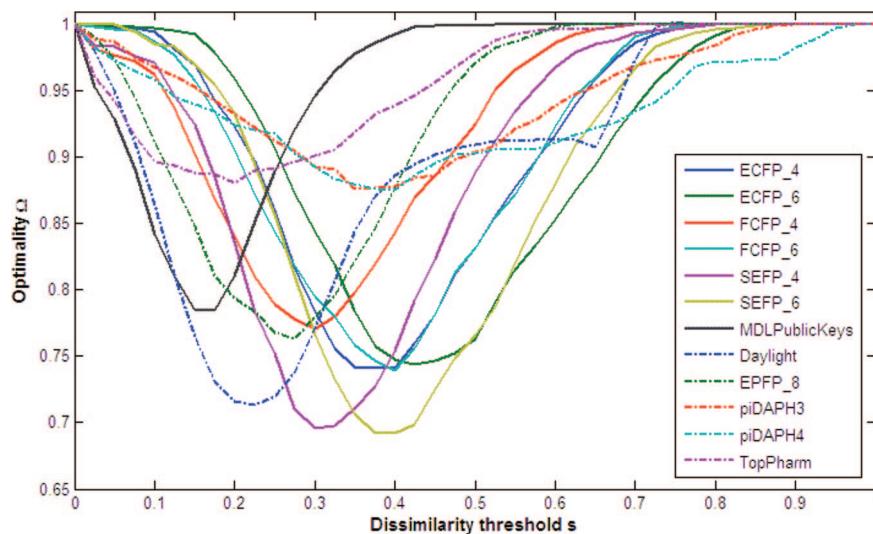
it might have been better to ignore the points in the rectangular part of the trapezoid and consider only those in the triangle, following the modification suggested previously by Clark et al.<sup>32</sup> and mentioned above. Third, a visual inspection of the trends on a Patterson plot is almost impossible given the vast amount of data points on it. This of course could be partly dealt with by generating the same plot for a much smaller random subsample of the data set. Fourth, the speed of computation. Computing the optimal diagonal is straightforward for the small QSAR data sets studied by Patterson et al. but becomes quite time-consuming for the data sets studied here, with our optimized MATLAB routines taking about four hours on a standard PC to identify the optimal diagonal for a single descriptor on a 700-compound data set.

**Optimality Criterion Method.** The implementation of the optimality criterion method requires the specification of an activity dissimilarity threshold ( $l$ ) and the weighting factor in eq 5 (k). The values for  $l$  are listed in Table 3, while k was set to 5 in all experiments, reflecting the much greater effect of FS (false similar), as against FD (potentially false dissimilar), molecule-pairs on the time and resources required in an optimization project. The values for  $l$  of 0.5 in Table 3 are greater than the threshold of 0.2 used by Horvath and Mao.<sup>21</sup> The larger values adopted here reflect the error ranges associated with most biological assays, with consecutive testing of the same molecule often yielding a  $|\Delta pIC50|$  of 0.3 or 0.4. A lower threshold of 0.3 was selected for metabolic stability, in order to cope with the smaller maximum pairwise property distance of 1 log unit.

The  $\Omega$ -s plots for different descriptors give information about their neighborhood behavior as well as their optimal similarity cutoff ( $s^*$ ) and hence allow a direct comparison between descriptors. A typical plot, that for the Project II/



**Figure 11.** Optimality plot for the Project II/Target 3/Chemotype 5 data.



**Figure 12.** Optimality plot for the Project II/Target 4/Chemotype 6 data.

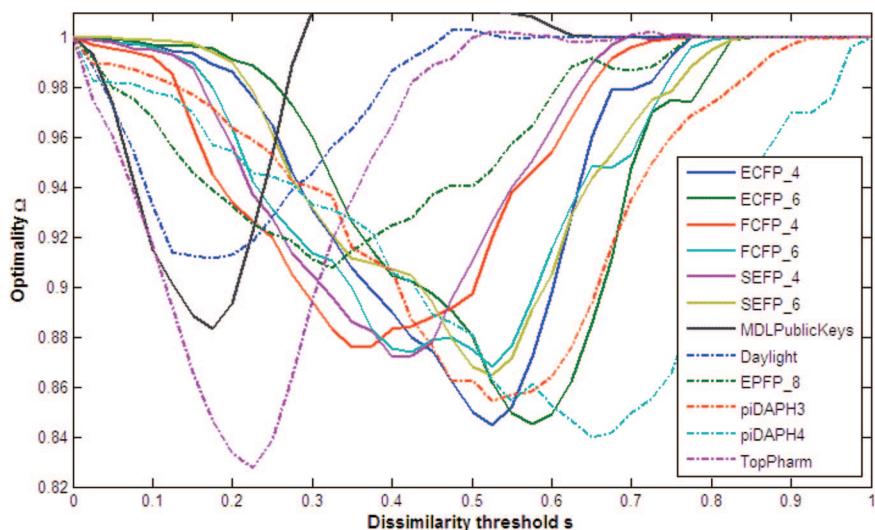
Target 3/Chemotype 5 data, is shown in Figure 11. The plot shows clearly the circular substructures, e.g., the ECFP and SEFP families of fingerprints, performing better for these data than, e.g., the Daylight or EPFP<sub>8</sub> fingerprints, where the best-performing descriptors are those with the lowest minimum value of  $\Omega$  (i.e., the deepest dip in the curve). Fingerprints such as MDL Public Keys, TopPharm, piDAPH3, and piDAPH4 show intermediate neighborhood behavior in the plot, as described by their minimal  $\Omega$  values. The circular substructure fingerprints exhibit the best neighborhood behavior at a dissimilarity cutoff significantly lower than that of the other descriptors, with the neighborhood radius,  $s^*$ , ranging between 0.3 and 0.4: this appears to be strongly characteristic of these fingerprints since the same behavior was observed consistently across all of our data sets.

In Figure 12 (for the Project II/Target 4/Chemotype 6 data) it is the Daylight and two SEFP fingerprints that perform the best, although the three descriptors have noticeably different values for the neighborhood radius. The three pharmacophore fingerprints are noticeably inferior to the other descriptors in Figure 12 but do much better when applied to the Project I/Target 1/Chemotype 1 data as shown in Figure 13. Here, both 2D and 3D pharmacophore fingerprints perform at least as well as their circular

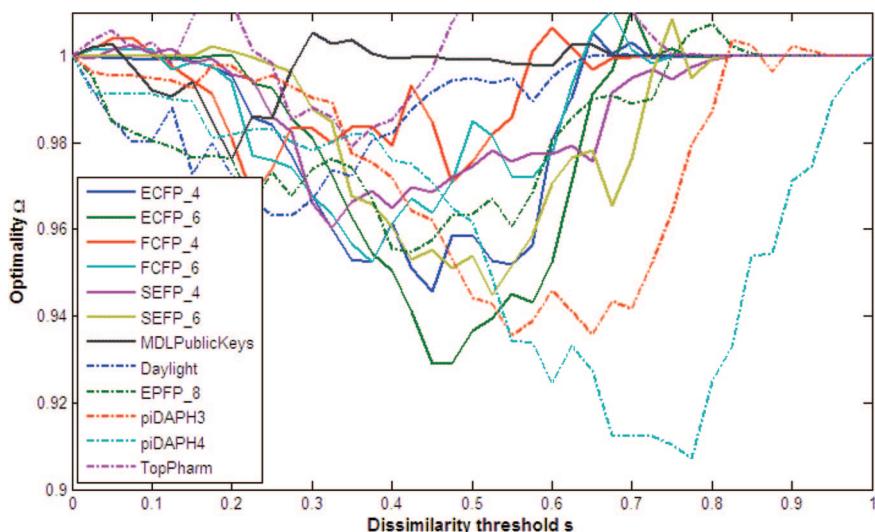
counterparts, indicating that for this combination of target and chemotype there may be pharmacophoric features that are important for potency. However, it should be noted that even the lowest minimum  $\Omega$  value is at about 0.83, much higher than for the plots in Figures 11 and 12 where the minimum values are about 0.51 and 0.69, respectively. The lowest minimum  $\Omega$  value is still worse for the third chemotype in this project/target combination, as shown in Figure 14 where the lowest value is about 0.91 (and this is for the piDAPH4 fingerprint with an impractical (too broad) neighborhood radius of 0.78). This implies very poor neighborhood behavior and is confirmed by the consistently low  $\chi^2$  values in the corresponding Patterson plot (see Figure 8).

Our final example of the application of the optimality criterion method is shown in Figure 15, for the Project II/Permeability/Chemotype 7 data. Here, the TopPharm fingerprints are very markedly superior to all of the other descriptors (as was also the case with the  $\chi^2$  scores in the corresponding optimal diagonal analysis (Figure 7)); indeed some of the other descriptors have minimal  $\Omega$  values as high as 0.94, indicating little or no neighborhood behavior.

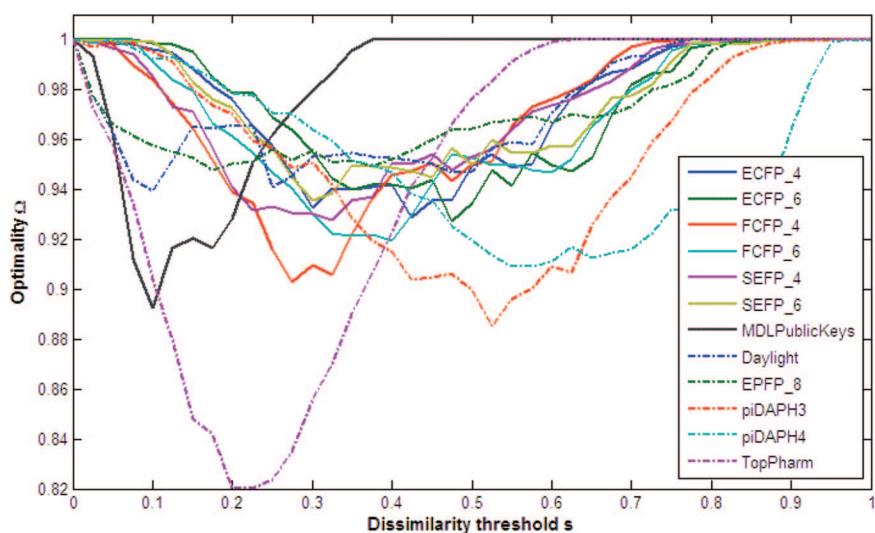
In general, it was found that the data sets with the lowest optimality scores were generally those with the highest  $\chi^2$  scores in the optimal diagonal method. To quantify this



**Figure 13.** Optimality plot for the Project I/Target 1/Chemotype 1 data.



**Figure 14.** Optimality plot for the Project I/Target 1/Chemotype 3 data.



**Figure 15.** Optimality plot for Project II/Permeability/Chemotype 7 data.

relationship, we averaged the optimality  $\Omega$ , NBE, and  $\chi^2$  values for each descriptor across all data sets and calculated the Pearson's correlation coefficient  $r$  between them. Table 4 illustrates the almost linear relationship, between the average  $\Omega$  and  $\chi^2$  values, across the 12 descriptors. However,

and disappointingly, no significant correlation was found between the average NBE and  $\Omega$  values.

We can obtain an overall view of the performance of the various fingerprints by averaging the values of the minimal optimality and the neighborhood radius across all the potency

**Table 4.** Pearson's Correlation Coefficient Scores between Average  $\Omega$  and  $\chi^2$  for the 12 Descriptors

descriptor	$r(\Omega, \chi^2)$	$r(\Omega, \text{NBE})$	descriptor	$r(\Omega, \chi^2)$	$r(\Omega, \text{NBE})$
ECFP_4	-0.905	0.44	MDLPublicKeys	-0.824	0.106
ECFP_6	-0.902	0.230	Daylight	-0.844	0.204
FCFP_4	-0.867	0.038	EPFP_8	-0.731	0.460
FCFP_6	-0.877	-0.050	piDAPH3	-0.830	-0.191
SEFP_4	-0.883	0.617	piDAPH4	-0.784	-0.258
SEFP_6	-0.893	0.488	TopPharm	-0.857	0.247

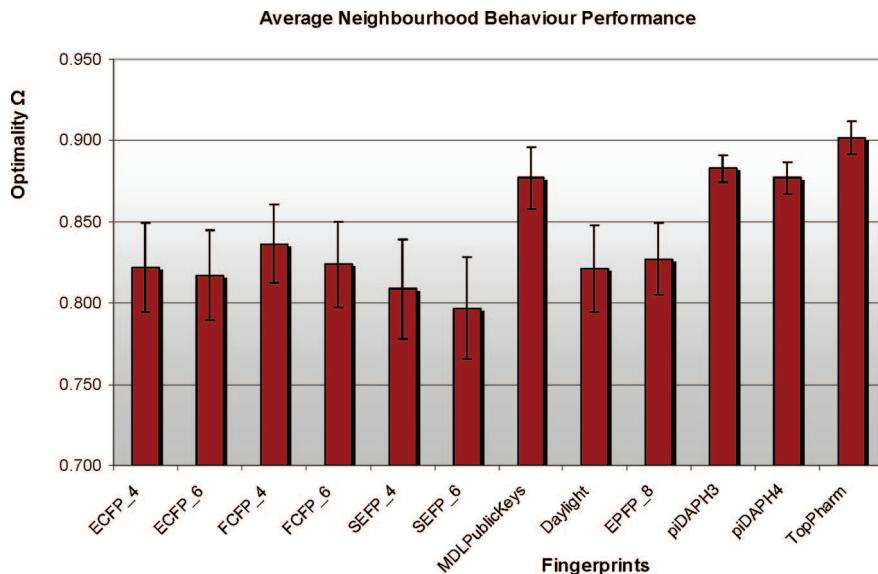
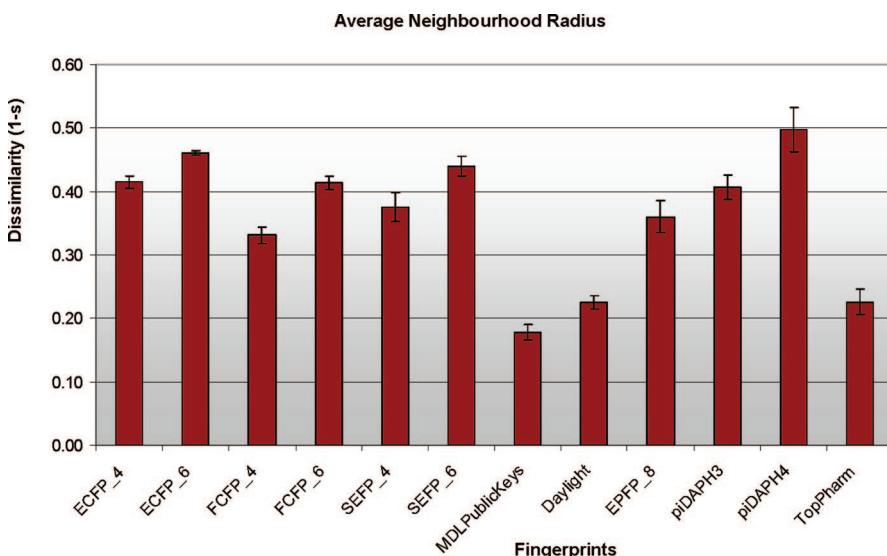
data sets for each descriptor. This analysis is shown in Figures 16 and 17.

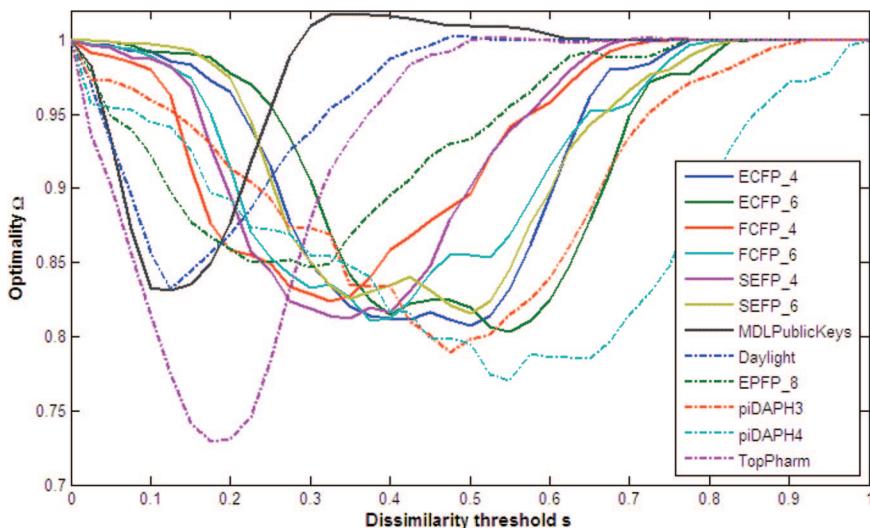
The principal observable trend is that the circular fingerprints (particularly ECFP\_6, SEFP\_4, and SEFP\_6) performed consistently well, with one of these being among the top three descriptors for every data set, a finding that is in agreement with previous studies of similarity searching studies.<sup>29,33</sup> The structural key (i.e., MDL Public Keys) and pharmacophore (i.e., piDAPH3, piDAPH4, and TopPharm) fingerprints performed poorly overall, although TopPharm

occasionally achieved a very high optimality score (e.g., the data shown in Figure 15). Of the path fingerprints, EPFP\_8 was of moderate effectiveness, being slightly worse than Daylight, which in turn was slightly worse than the circular substructure fingerprints such as ECFP\_4.

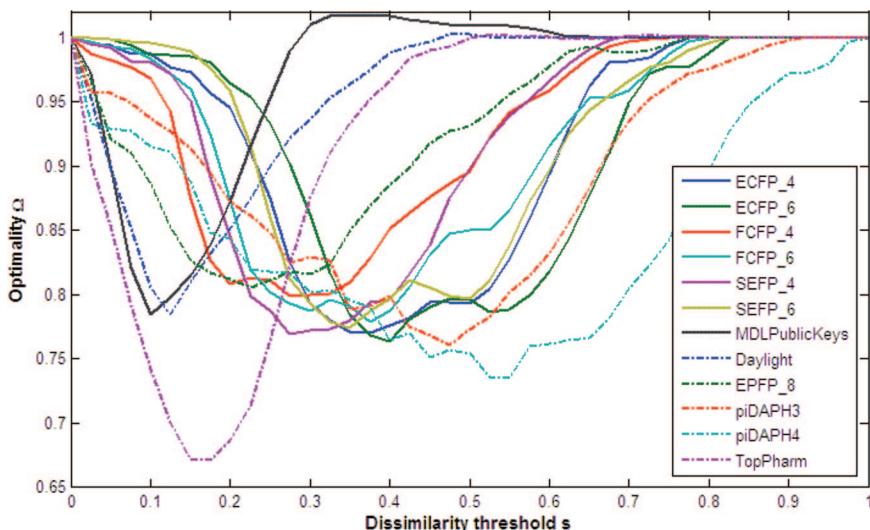
The neighborhood radius for the circular fingerprints is much lower than for the structural key or path-based fingerprints. Specifically, circular fingerprints achieve their optimal neighborhood behavior at a similarity threshold  $s^*$  of approximately 0.55. This, of course, can be attributed to the fact that they represent a huge set of features (2<sup>32</sup>) which results in much sparser bit-strings than most of the other fingerprints used in this study. The same applies to 3D pharmacophoric fingerprints. On the other hand, the well-established Daylight fingerprints have an average similarity threshold of slightly lower than 0.8. These findings are in direct agreement with the recent paper of Muchmore et al.<sup>34</sup> as well as with other 2D similarity studies.<sup>35</sup>

We have also investigated the effect of the weight  $k$  on the optimality scores. For the Project I/Target 1/Chemotype

**Figure 16.** Average neighborhood behavior performance and the associated standard errors according to the optimality criterion method.**Figure 17.** Average optimal similarity cutoff (neighborhood radius) and the associated standard errors according to the optimality criterion method.



**Figure 18.** Optimality plot for the Project I/Target 1/Chemotype 1 data, with  $k=15$ .



**Figure 19.** Optimality plot for the Project I/Target 1/Chemotype 1 data, with  $k=25$ .

1 data set (Figure 13 where  $k$  was set to 5), plots were generated with the value of  $k$  set to 15 and 25 (Figures 18 and 19, respectively).

Two main trends can be detected here. First, neighborhood behavior increases ( $\Omega$  is lower) as  $k$  increases. For  $k=5$ , the optimal neighborhood behavior is approximately 0.83, whereas for  $k=25$  the optimal neighborhood behavior is 0.67. Second, the similarity threshold  $s^*$  at which the descriptors achieve their best neighborhood behavior shifts slightly to the left, i.e. toward bigger values. Both observations can be explained by the fact that an increase in  $k$  favors descriptors with higher consistency (i.e., fewer false similar pairs among the structurally most similar ones) and hence the separation among them becomes clearer. In general, the increase in  $k$  did not dramatically affect the relative rank of the descriptors, with the TopPharm fingerprints coming first in all three cases. These results are consistent across all the data sets (data not shown).

Finally, we have already noted that the consistency and optimality criteria in the analysis of Horvath and Jeandanans are closely related to parameters that are used to measure the effectiveness of information retrieval systems. It can be shown that the optimality is also closely related to the long-

established kappa ( $\kappa$ ) coefficient of Cohen.<sup>36</sup> This coefficient provides a simple way to assess the degree of concordance between two methods (or judges or raters) that classify objects into two or more mutually exclusive categories. The coefficient has been extensively used in fields such as ecology and medicine<sup>37</sup> and a variant of it, the Rand coefficient,<sup>38</sup> in chemoinformatics to compare different clustering methods.

Cohen's  $\kappa$  statistic is defined as

$$\kappa = \frac{O - E}{1 - E} \quad (9)$$

where  $O$  and  $E$  are the observed and the expected accuracies of classification. Using the contingency table in Figure 5

$$O = \frac{N_{TS} + N_{TD}}{N} \quad (10)$$

and

$$E = \frac{(N_{TS} + N_{FS})(N_{TS} + N_{PFD}) + (N_{PFD} + N_{TD})(N_{FS} + N_{TD})}{N^2} \quad (11)$$

from which we can compute kappa as

$$\kappa = \frac{N(N_{TS} + N_{TD}) - (N_{TS} + N_{FS})(N_{TS} + N_{PFD}) - (N_{PFD} + N_{TD})(N_{FS} + N_{TD})}{N^2 - (N_{TS} + N_{FS})(N_{TS} + N_{PFD}) - (N_{PFD} + N_{TD})(N_{FS} + N_{TD})} \quad (12)$$

If we set  $k=1$  in the definition of  $\Omega$  (eq 5), so that

$$\Omega = \frac{N_{FS} + N_{PFD}}{N_{FS}^{\text{expected}} + N_{PFD}^{\text{expected}}} \quad (13)$$

and substitute for the expected values of  $N_{FS}$  and  $N_{PFD}$ , from eqs 6 and 7 then we get

$$\Omega = \frac{N(N_{FS} + N_{PFD})}{(N_{TS} + N_{FS})(N_{TD} + N_{FS}) + (N_{TS} + N_{PFD})(N_{TD} + N_{PFD})} \quad (14)$$

A comparison of eqs 12 and 14 reveals that  $\Omega = 1 - \kappa$ , i.e., that kappa is simply the complement of optimality when (5) is unweighted.

## CONCLUSIONS

This paper has compared the use of two methods for assessing the neighborhood behavior of molecular descriptors using data from three array-based lead optimization projects at GlaxoSmithKline. We have shown that both methods—the optimal diagonal method and the optimality criterion method—are applicable to such data and that they can provide useful information as to the merits of different molecular descriptors, information that can then be used to assist in the design of new arrays. That said, we have found some limitations with the optimal diagonal method, which are magnified when it comes to large data sets: the two scoring mechanisms associated with it (NBE and  $\chi^2$ ) can often yield misleading and confusing results; NBE can underestimate the true neighborhood behavior performance of a descriptor in the case of a large data set containing mainly dissimilar pairs of molecules; the resulting Patterson plots are difficult to interpret and the definition of the neighborhood radius/dissimilarity threshold is rather subjective; and it is time-consuming when (as here) large data sets need to be processed. Conversely, the generation of  $\Omega$ -s plots is rapid, and their nature allows for a direct and clear comparison of the neighborhood behavior performance across descriptors as well as their optimal similarity threshold. In conclusion, the optimality criterion method, which is essentially a reformulation of the widely used Cohen's kappa statistic, provides a simple, objective, and robust framework for evaluating the neighborhood behavior of molecular descriptors (and presumably of similarity coefficients, although we have not studied these here). We hence recommend this as an effective way of selecting descriptors for array-based lead optimization programs.

## ACKNOWLEDGMENT

We thank the Engineering and Physical Sciences Research Council and GlaxoSmithKline for funding and the Royal Society and the Wolfson Foundation for laboratory support. G.P. would also like to thank the Alexander S. Onassis Public Benefit Foundation in Greece for their financial support. SciTegic/Accelrys are gratefully acknowledged for provision of Pipeline Pilot.

**Supporting Information Available:** Data tables for all data sets including pairwise similarity values for all descriptors and pairwise property differences. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## REFERENCES AND NOTES

- (1) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley: New York, 1990.
- (2) Willett, P. Similarity methods in chemoinformatics. *Annu. Rev. Inf. Sci. Tech.* **2009**, *43*, 3–71.
- (3) Harper, G.; Pickett, S. D.; Green, D. V. S. Design of a compound screening collection for use in High Throughput Screening. *Comb. Chem. High Throughput Screening* **2004**, *7*, 63–70.
- (4) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
- (5) Kubinyi, H. Similarity and dissimilarity: a medicinal chemist's view. *Perspect. Drug Discovery Des.* **1998**, *9–11*, 225–232.
- (6) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activities. *J. Med. Chem.* **2002**, *45*, 4350–4358.
- (7) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.
- (8) Bostrom, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion. *J. Med. Chem.* **2006**, *49*, 6716–6725.
- (9) Oprea, T. I.; Tropsha, A. Target, chemical and bioactivity databases—integration is key. *Drug Discovery Today: Technol.* **2006**, *3*, 357–365.
- (10) Barbosa, F.; Horvath, D. Molecular similarity and property similarity. *Curr. Top. Med. Chem.* **2004**, *4*, 589–600.
- (11) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighbourhood behaviour: a useful concept for validation of “molecular diversity” descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.
- (12) Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity – a review. *QSAR Comb. Sci.* **2003**, *22*, 1006–1026.
- (13) Horvath, D.; Jeandenans, C. Neighborhood Behavior of in silico structural spaces with respect to in vitro activity spaces – A novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680–690.
- (14) Schuffenhauer, A.; Brown, N. Chemical diversity and biological activity. *Drug Discovery Today: Technol.* **2006**, *3*, 387–395.
- (15) Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219–1229.
- (16) Dixon, S. L.; Merz, K. M. One-dimensional molecular representations and similarity calculations: methodology and validation. *J. Med. Chem.* **2001**, *44*, 3795–3809.
- (17) Horvath, D.; Jeandenans, C. Neighborhood Behavior of in silico structural spaces with respect to in vitro activity spaces – A benchmark for Neighborhood Behavior assessment of different in silico similarity metrics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 691–698.
- (18) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261–266.
- (19) Schuffenhauer, A.; Brown, N.; Ertl, P.; Jenkins, J. L.; Selzer, P.; Hamon, J. Clustering and rule-based classifications of chemical structures evaluated in the biological activity space. *J. Chem. Inf. Model.* **2007**, *47*, 325–336.
- (20) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of retrieval in similarity searches of chemical databases: A review of performance measures. *J. Mol. Graphics Modell.* **2000**, *18*, 343–357.
- (21) Horvath, D.; Mao, B. Neighborhood Behavior: Fuzzy molecular descriptors and their influence on the relationship between structural

- similarity and property similarity. *QSAR Comb. Sci.* **2003**, *22*, 498–509.
- (22) Perekhdtsev, G. Neighborhood behavior: Validation of two-dimensional molecular similarity as a predictor of similar biological activities and docking scores. *QSAR Comb. Sci.* **2007**, *26*, 346–351.
- (23) Peltason, L.; Bajorath, J. SAR index: quantifying the nature of structure-activity relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.
- (24) Guha, R.; van Drie, J. H. Structure-Activity Landscape Index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.
- (25) Scitegic Pipeline Pilot, 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123–1365, U.S.A. Available from SciTegic Inc. at <http://www.scitegic.com> (accessed Oct 15, 2008).
- (26) Daylight Daylight Toolkit, Daylight Chemical Information Systems, Inc., 120 Vantis - Suite 550, Aliso Viejo, CA 92656, U.S.A. Available at <http://www.daylight.com> (accessed Oct 15, 2008).
- (27) Chemical Computing Group *Molecular Operating Environment (MOE)*, Chemical Computing Group. 1010 Sherbrooke St. W, Suite 910 Montreal, Canada H3A 2R7. Available from Chemical Computing Group at <http://www.chemcomp.com> (accessed Oct 15, 2008).
- (28) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics. Revised Edition*; Springer: Dordrecht, 2007.
- (29) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments: information-based feature selection and a naive Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.
- (30) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. “Scaffold-hopping” by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.
- (31) Maggiola, G. M. On outliers and activity cliffs - why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
- (32) Clark, R. D.; Brusati, M.; Jilek, R.; Heritage, T.; Cramer, R. D. Validating novel QSAR descriptors for use in diversity analysis. In *Molecular Modeling and Prediction of Bioactivity*; Gundertofte, K., Jorgensen, W., Eds.; Kluwer Academic/Plenum Publishers: New York, 2000; pp 95–100.
- (33) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.
- (34) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941–948.
- (35) Hert, J. Two-dimensional similarity-based methods for virtual screening, Ph.D. thesis, The University of Sheffield: Sheffield, 2005.
- (36) Cohen, J. A coefficient of agreement for nominal scales. *Educational Psychological Meas.* **1960**, *20*, 37–46.
- (37) Manel, S.; Williams, H. C.; Ormerod, S. J. Evaluating presence-absence models in ecology: the need to account for prevalence. *J. Appl. Ecol.* **2001**, *38*, 921–931.
- (38) Rand, W. M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **1971**, *66*, 846–850.

CI800302G