

# Structural Similarity Based Kriging for Quantitative Structure Activity and Property Relationship Modeling

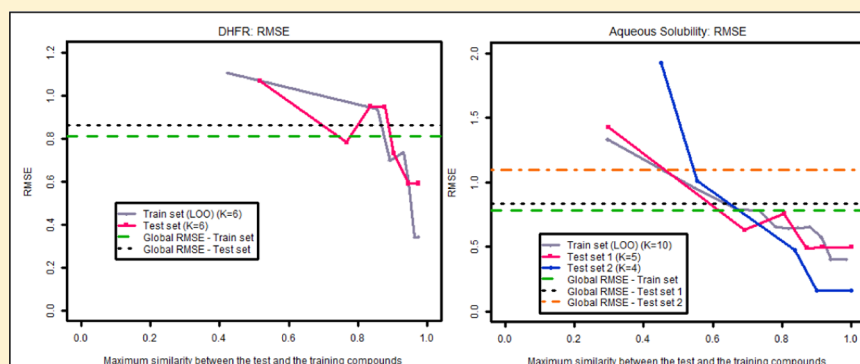
Ana L. Teixeira<sup>\*,†,‡</sup> and Andre O. Falcao<sup>†,¶</sup>

<sup>†</sup>LaSIGE, Faculty of Sciences, University of Lisbon, 1749-016 Lisbon, Portugal

<sup>‡</sup>CQB - Centro de Química e Bioquímica, Faculty of Sciences, University of Lisbon, 1749-016 Lisbon, Portugal

<sup>¶</sup>Department of Informatics, Faculty of Sciences, University of Lisbon, 1749-016 Lisbon, Portugal

## S Supporting Information



**ABSTRACT:** Structurally similar molecules tend to have similar properties, i.e. closer molecules in the molecular space are more likely to yield similar property values while distant molecules are more likely to yield different values. Based on this principle, we propose the use of a new method that takes into account the high dimensionality of the molecular space, predicting chemical, physical, or biological properties based on the most similar compounds with measured properties. This methodology uses ordinary kriging coupled with three different molecular similarity approaches (based on molecular descriptors, fingerprints, and atom matching) which creates an interpolation map over the molecular space that is capable of predicting properties/activities for diverse chemical data sets. The proposed method was tested in two data sets of diverse chemical compounds collected from the literature and preprocessed. One of the data sets contained dihydrofolate reductase inhibition activity data, and the second molecules for which aqueous solubility was known. The overall predictive results using kriging for both data sets comply with the results obtained in the literature using typical QSPR/QSAR approaches. However, the procedure did not involve any type of descriptor selection or even minimal information about each problem, suggesting that this approach is directly applicable to a large spectrum of problems in QSAR/QSPR. Furthermore, the predictive results improve significantly with the similarity threshold between the training and testing compounds, allowing the definition of a confidence threshold of similarity and error estimation for each case inferred. The use of kriging for interpolation over the molecular metric space is independent of the training data set size, and no reparametrizations are necessary when more compounds are added or removed from the set, and increasing the size of the database will consequentially improve the quality of the estimations. Finally it is shown that this model can be used for checking the consistency of measured data and for guiding an extension of the training set by determining the regions of the molecular space for which new experimental measurements could be used to maximize the model's predictive performance.

## 1. INTRODUCTION

The number of compounds discovered each day continues to grow at an exponential rate due to constantly refined and optimized experimental technologies.<sup>1</sup> However, the experimental determination of the chemical, physical, and biological properties of compounds is often expensive, time-consuming, and in many cases impossible. According to George Hammond in the 1968 Norris Award Lecture, "the most fundamental and lasting objective of synthesis is not production of new compounds, but production of properties", thus it is evident that there is a great need to apply property prediction methods with a good

predictive performance when experimental values are not available.

A commonly used approach to predict chemical, physical, and/or biological properties of chemical compounds resorts to the structure of the molecule using data mining methods through the quantitative structure–property/activity relationships (QSPR/QSAR).<sup>2–4</sup> The three major difficulties in the development of QSPR/QSAR models are (1) quantifying the

**Received:** December 17, 2013

inherently abstract molecular structure, (2) determining which structural features most influence the given property (representation problem),<sup>5–7</sup> and (3) establishing and validating the functional relationships that most accurately describe the relationship between structural descriptors and the property/activity data (mapping problem).<sup>8–11</sup> Furthermore, it is acknowledged that it is not possible to develop a model providing reliable predictions for all possible compounds (chemical space).<sup>12</sup> Classical QSPR/QSAR approaches have several shortcomings, namely (1) the predictive power of the model is highly dependent on the selection of predictor variables and on the presence of correlation between these variables, (2) the prediction capacity of the model is limited by the molecular diversity and distribution of the molecules in the training set,<sup>13</sup> (3) the models need to be retrained every time new compounds are added or removed, and (4) usually only the uncertainty of the model is assessed and reported.<sup>14</sup>

For this study we propose the use of a method that, in light of the structural similarity principle,<sup>15</sup> takes into account the high dimensionality of the chemical space, predicting chemical, physical, or biological properties based on the most structurally similar compounds in the molecular space, consequently avoiding the selection of descriptors. Another aspect we will address is the assessment of the reliability and the uncertainty of each estimation based on the structural similarity level. However, the definition of structural similarity is not trivial, since the concept of similarity is subjective and even chemists are not consistent when comparing molecules.<sup>16</sup> The definition of structural similarity for molecules consists of mapping the chemical space, specifically representing the molecules and quantifying the similarity between them, enabling, in light of the similarity principle, the use of the derived similarity measures in the prediction context. Various methods to define structural similarity between molecules are available in the literature,<sup>17,18</sup> and they can be divided in three broad categories, each with its own specificities - approaches based on the following: (1) structural descriptors (two- and three-dimensional),<sup>19–22</sup> (2) molecular fragments (such as fingerprints),<sup>23</sup> and (3) graph matching/descriptor-independent methods (such as the noncontiguous atom matching function (NAMS)<sup>24</sup>).

Making predictions out of similarity or distance metrics is a known problem in several areas of science. One of the most known used methods for quantitative estimation based on topological distances is kriging, a method traditionally used in geostatistics which makes use of Tobler's first law of geography:<sup>25</sup> *"Everything is related to everything else, but near things are more related than distant things"*, meaning that a spatial dependence in the data is considered contrary to traditional statistical methods which assume that all data are independent. This method involves the estimation of a regionalized variable at a particular unsampled location by the weighted combination of the values of the neighboring locations.<sup>26–28</sup> The use of this method has several advantages, namely the following: (1) estimates the estimation error along with the estimate of the property for each compound and this estimation error is minimized, therefore it is expected to be zero at the locations where experiments are performed and to grow with distance from these; (2) easier to comprehend than a black box model; (3) makes use of the distance/similarity between the compounds and it is not dependent on the selection of molecular descriptors; (4) fast enough to apply to a large data set; (5) searches for the relationship among

measured properties rather than approximate the modeled system by fitting the parameters of the selected basis functions.

Kriging models are not new in chemoinformatics. Pioneer work was developed by Burden<sup>29</sup> which demonstrated the applications of kriging in QSAR modeling for three data sets. Fang et al.<sup>30</sup> used this technique for predicting boiling points of hydrocarbons and showed that kriging models could significantly improve the performances of the models by other existing methods. Obrezanova et al.<sup>31</sup> applied kriging for the prediction of absorption, distribution, metabolism, and excretion properties. Hawe et al.<sup>32</sup> used kriging to predict the basicities of pyridines. Sun et al.<sup>33</sup> showed that kriging models were able to outperform other methods in the development of predictive models for skin absorption. However, in all of these studies there was always an explicit use of chemical descriptors arbitrarily chosen according to the nature of the problems. To the best of our knowledge, none of the above approaches combined structural similarity with kriging methods for property/activity prediction of chemical compounds. In this study we intend to demonstrate the application of kriging for molecular property estimation coupled with different similarity metrics based solely on the structure of the compound.

The main objectives of this work are (1) to demonstrate that structural similarity functions can be useful to define the chemical space that is used to accurately predict properties/activities for diverse chemical compounds as yet unmeasured or even not synthesized, (2) to assess the extent to which kriging can be used to predict unmeasured properties of chemical compounds that were selected randomly or based on temporal characteristics using solely the metric map defined by structural similarity, (3) to determine the uncertainty of each estimation, and (4) to determine the effect of the training set size on the predictive results of the method. Further potential applications of this methodology are illustrated by using three different structural similarity approaches based on molecular descriptors, fragments, and graph matching to predict aqueous solubility and inhibition activity using two data sets of compounds with different structural characteristics.

## 2. METHODS

This section presents the modeling methodology which is based on the kriging algorithm that requires the use of the chemical space based on the distance between molecules. Three different ways to represent molecules based solely on their structure are studied in order to determine if it is possible to use their structural distance to estimate properties and which is the best way to calculate it in order to maximize the predictive power and minimize the number of neighbor compounds needed. In order to ensure minimal bias in evaluating the results an internal and external validation procedure was followed and is described, both for model selection as well as for final model assessment.

**2.1. Modeling Methodology.** The estimation of property values for which their properties were not experimentally determined and based solely on the structural similarity between the molecules is not sufficient. It is necessary to take into account the irregularities in the property values, i.e. if the response variable surface has some spatial correlation then it is possible to infer the response in the immediate environment. One method of incorporating these concepts in the estimation model is to use kriging. Kriging is a family of estimators generally used in geostatistics for the interpolation of spatial data, i.e. to estimate variables at unobserved locations based on

observed points at nearby locations.<sup>26,27,34</sup> The kriging interpolation method seems to be a promising approach, as based on values measured in points from a certain range, it allows making predictions and the uncertainty of each prediction knowing just the distances to the known instances. The most widely used method is ordinary kriging, which was also selected for this study as it is the simplest model, makes no assumption on the nature or properties of the metric space, and uses only distances between instances and measured values for inference.

**Ordinary Kriging.** The definition of Ordinary Kriging (OK) is often associated with the acronym "BLUE", for "best linear unbiased estimator". "Best" because OK aims at minimizing the variance of the errors, "linear" since its estimates are weighted linear combinations of the available data, and "unbiased" because it attempts to reduce the mean residual error to zero. These goals are ambitious since the mean residual error and the variances of the errors are unknown for the data points to be predicted. When using other modeling techniques, the usual procedure involves building a model of the data and work with the average error and the error variance of the model. OK, on the other hand, uses a probabilistic model in which the bias and the error variance can be calculated in order to choose weights for the nearby sample which ensures that the average error of the model is zero and that the modeled variance is minimized. To estimate the error, its mean value and its variance a random function model can be used, since it takes into account the uncertainty of what happens at unsampled points. This allows the construction of a map of both predicted values and level of uncertainty about the predicted values. To estimate unsampled points ( $\hat{v}$ ) a weighted linear combination of the available samples can be used as in eq 1, where  $n$  is the number of compounds with known property/activity in the set,  $v$  are the values of the property/activity, and  $w_j$  are the weights assigned to each known compound. The set of weights can change as the location of the unknown points change.

$$\hat{v} = \sum_{j=1}^n w_j \cdot v \quad (1)$$

The error of the  $i$ -th estimate ( $r_i$ ) can then be defined as the difference between the estimated value ( $\hat{v}_i$ ) and the true value at the same location ( $v_i$ ) (eq 2).

$$r_i = \hat{v}_i - v_i \quad (2)$$

The average error ( $m_R$ ) of a set of  $k$  estimates can then be defined as in eq 3.

$$m_R = \frac{1}{k} \cdot \sum_{i=1}^k r_i \quad (3)$$

However, in practical situations the true value ( $v_i$ ) is not known; therefore, as mentioned above, a probabilistic approach allows the calculation of unknown values as the outcome of a random process. For that purpose a random variable  $V(x_0)$  with an expected value of  $E^2$  is assigned to the unknown value to be estimated. The pairs of random variables have a distribution that depends only on their distance and not on their locations. The covariance between pairs of random variables separated by a distance  $h$  is  $\tilde{C}_v(h)$ . The predicted estimate and the estimation error are also random variables since these are the outcome of a weighted linear combination on the random variables at the available sample location as described in eq 1 with  $v_i = V(x_0)$

and eq 2 with  $r_i = R(x_0)$ . For an unbiased estimation it is important to take into account that  $E\{R(x_0)\}$  should be equal to zero, which means that the sum of weights has to be equal to one.

The error variance  $\sigma_R^2$  of a set of  $k$  estimates can be expressed as eq 4, and it represents the function that OK aims to minimize.

$$\sigma_R^2 = \frac{1}{k} \cdot \sum_{i=1}^k (r_i - m_R)^2 \quad (4)$$

The average error ( $m_R$ ) is assumed to be zero, and therefore it can be eliminated from the equation. As already mentioned the true values are not known; therefore the same random function models are needed to minimize the variance of the modeled error  $R(x_0)$ . For that purpose, the variance of the error can be expressed as the variance of a weighted linear combination of random variables (eq 5).

$$\text{Var}\left\{\sum_{i=1}^n w_i \cdot v_i\right\} = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \cdot \text{Cov}\{v_i v_j\} \quad (5)$$

Minimizing the variance of the error requires setting the  $n$  partial derivatives, namely the weights  $w_1, \dots, w_n$  to zero. This produces a system of  $n$  simultaneous linear equations with  $n$  unknowns for the  $n$  sample locations, having in mind the unbiasedness condition that the sum of weights has to be equal to one. The solution for this  $n + 1$  system of equations is not straightforward; however, it can be solved using Lagrange multipliers (eq 6).

$$\begin{aligned} \frac{\partial \sigma_R^2}{\partial w_i} = 0 &\Rightarrow \sum_{j=1}^n w_j \tilde{C}_{ij} + \mu = \tilde{C}_{i0} \quad \forall i = 1, \dots, n \\ \frac{\partial \sigma_R^2}{\partial \mu} = 0 &\Rightarrow \sum_{i=1}^n w_i = 1 \end{aligned} \quad (6)$$

This solution will provide the set of weights and the mean value (the Lagrange parameter  $\mu$ ) that minimizes the modeled error variance (eq 7) under the constraint that weights sum to one.

$$\tilde{\sigma}_R^2 = \tilde{\sigma}^2 + \sum_{i=1}^n \sum_{j=1}^n w_i w_j \tilde{C}_{ij} - 2 \cdot \sum_{i=1}^n w_i \tilde{C}_{i0} + 2 \cdot \mu \left( \sum_{i=1}^n w_i - 1 \right) \quad (7)$$

The system of equations represented in eq 6 can be expressed in matrix notation which is usually known as the OK system (eq 8).

$$\begin{aligned} \begin{bmatrix} \tilde{C}_{11} & \dots & \tilde{C}_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \tilde{C}_{n1} & \dots & \tilde{C}_{nn} & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \mu \end{bmatrix} &= \begin{bmatrix} \tilde{C}_{01} \\ \vdots \\ \tilde{C}_{n0} \\ 1 \end{bmatrix} \\ C \cdot w &= D \\ w &= C^{-1} \cdot D \end{aligned} \quad (8)$$

The vector  $D$  provides a weighting scheme of the distances using the covariances between all the sample locations (denoting  $i = 1, 2, \dots, n$ ) and locations where an estimation is needed (denoting 0). The higher the covariance between a sample and the location being estimated, the more that sample



contributes to the estimation. The matrix  $C$  describes the covariances between all the sample pairs, bestowing information about the distribution of the available sample data. Therefore,  $C$  matrix readjusts the sample weight according to their clustering. Alternatively to the covariance between a sample and the locations being estimated, a closely related measure can be used to give the same information - the semivariance.

In terms of the matrices defined in 8, the minimized error variance (eq 7) can be expressed as eq 9, which is usually referred to as the kriging estimated variance.

$$\hat{\sigma}_R^2 = \hat{\sigma}^2 - w \cdot D \quad (9)$$

The kriging estimated variance takes into account four important factors and the interactions between them: (1) the number of samples used to make the estimation, it is expected that estimates based on many samples will be more reliable than those based on just a few; (2) proximity of the samples, as the average distance increases, the estimate becomes less reliable; (3) the spatial arrangement (clustering) of the samples around the test compound; and (4) the nature of the problem in terms of spatial continuity, smoothness and well-behaved variables will have better estimates than very erratic variables.

In summary, to minimize the modeled error variance, it is necessary to choose the  $(n+1)^2$  covariances that will describe the spatial distribution of the random function model. The set of weights that produce an unbiased estimate with a minimum error variance (eq 9) can be simply calculated using the system of eqs 8. The choice of the covariance or semivariance model to describe the spatial continuity is then a prerequisite to apply OK.

**Semivariogram.** A semivariogram describes how the spatial continuity changes with distance between all pairs of sample locations, quantifying the spatial correlation. In practice, OK is usually implemented using the semivariogram rather than the covariogram because it has better statistical properties.<sup>35</sup> The construction of a semivariogram consists of two parts: an empirical semivariogram and a model semivariogram that will extend the estimations to locations where there are possibly no sample locations by fitting a function to the empirical semivariogram. The empirical semivariogram is constructed by calculating the semivariance ( $\gamma(h)$ ) of each point in the set with respect to each of the other points, using eq 10 in relation to the distance between these points.

$$\gamma(h) = \frac{1}{2N(h)} \sum_{N(h)} (z_i - z_j) \quad (10)$$

$N(h)$  is the set of all pairwise distances ( $h$ ), and  $z_i$  and  $z_j$  are data values at spatial locations  $i$  and  $j$  separated by  $h$ . Using all pairs of compounds on the semivariogram may complicate its interpretation, and therefore it is usual to apply a binning process i.e., average semivariance data by distance intervals. An appropriate parametric model is then typically fitted into the empirical semivariogram and utilized to calculate distance weights for interpolation. Identifying the optimal model may involve running and evaluating a large number of models. Usually the model includes three parameters: (1) the "nugget" which represents the semivariance at distance zero due to microscale variations or low accuracy of the measurement; (2) the "range" which represents the distance at which semivariance levels off, that is to say the spatially correlated portion of the data; (3) the "sill" which represents the semivariance at which

the mentioned leveling takes place. After a suitable semivariogram model has been selected, the kriging process is able to define a continuous surface for the entire study area using weights calculated with the semivariogram model, as well as values and locations of the measured points. It is also possible to adjust the distance or number of measured points that are used for making predictions for each unknown value.

**2.2. Implementation of Ordinary Kriging.** *CoordKrig - Coordinate Based Kriging.* The R<sup>36</sup> package *geoR*<sup>37,38</sup> has an efficient implementation of OK. However, this package requires the coordinates of the data points instead of their distances, since this package was designed for geostatistical data analysis in which typical data inputs are the coordinates of data locations and the data values. For that purpose, multidimensional scaling (through the function *isoMDS* of R package *MASS*<sup>39</sup>) was used to transform the distances between the molecules into (XY) coordinates. These coordinates were then jittered uniformly on the regions around points with very similar coordinates using the function *jitter2D* of the *geoR* package. To fit a model to the semivariogram a spherical function was considered adequate given data distribution and the method *variofit* of the package *geoR* was used to estimate its parameters (sill and range) that give the smallest value of the summation and use them as initial values for the minimization of the loss function using *cressie* weights. The nugget was fixed at zero, and the spherical function was used to model the semivariance. Preliminary tests with several data sets showed that this function provided consistently good results.

*DistKrig - Distance Based Kriging.* Alternatively, and since OK derives predicted values based on the distance between points in space and the variation between measurements as a function of distance, an OK algorithm was implemented in R using as input a distance matrix between the molecules and following all the steps presented in the modeling methodology. However, for every molecule that we aim to predict the property value, a neighborhood will be demarcated by a predefined number (*neighs*) of molecules in the training set that are nearest to the test molecule. To fit a model to the semivariogram, linear regression was chosen, since this function shown to be suitable to model the data and simple to automate the process, using the R package *lm*<sup>40</sup> and defining that the regression line will pass through the origin (defined in geostatistics as the nugget), since it is assumed that if two molecules are 100% similar, then it is expected that they share the same property value.

**2.3. Molecular Representation.** The use of kriging for nonspatial problems requires working on a metric space, where the distances between all existing elements can be computed. In this context, a requirement to apply kriging is that the molecules need to be represented in a map based on their dissimilarity. As presented above, various methods to define structural similarity are available in the literature<sup>17,41</sup> and can be divided in three board categories: approaches based on structural descriptors, molecular fragments and graph matching. In the conducted study these three approaches to quantify the structural similarity of molecules, which are posteriorly transformed into distances, will be explored:

**A. Structural Similarity Based on Molecular Descriptors.** Molecular descriptors can be computed from the molecular structure encoding in numerical form chemical information contained in the molecule. In this work a set of 1666 molecular descriptors (2- and 3-Dimensional) was generated for each data set using *e-DRAGON*,<sup>42,43</sup> a free online version of *DRAGON*.

The 3D atomic coordinates of the lower energy conformation for the provided molecules were calculated using CORINA.<sup>44</sup> A preprocessing step was carried out where all zero variance variables (i.e., all the observations are the same) were removed and standardization was applied to transform each descriptor values to have zero mean and unit variance according to eq 11 where  $x$  represents the values of a molecular descriptor,  $\bar{x}$  is the mean value for descriptor  $x$ , and  $\sigma$  is its standard deviation. Each descriptor positions each abstract molecular representation in the descriptor space, and the molecular dissimilarity can be measured calculating the Euclidean distance between all dimensions of the chemical space.

$$x_{\text{standardized}} = \frac{x - \bar{x}}{\sigma} \quad (11)$$

**B. Structural Similarity Based on Molecular Fragments.** In this work the structural similarity score based on molecular fragments is obtained by comparing path-based fingerprints (FP2) calculated by openbabel<sup>45</sup> using the Tanimoto coefficient.<sup>23</sup> The FP2 binary fingerprints are bit strings that encode the presence or absence of topological patterns up to 7 atoms in a molecule and map them onto a bit-string of length 1024 using a hash function (similar to the Daylight fingerprints).

The degree of similarity given by the Tanimoto coefficient ( $s(x, y)$ ) was converted to a degree of dissimilarity ( $d(x, y)$ ) applying a monotonically decreasing transformation using the natural logarithm (eq 12).

$$d(x, y) = -\ln(s(x, y)) \quad (12)$$

**C. Structural Similarity Based on Graph Matching.** A molecule can also be represented, using graph theory, as a labeled graph whose vertices correspond to the atoms and edges correspond to the covalent bonds. The representation of molecules using graphs has some advantages, namely, graphs are intuitive when representing a molecule since they are close to our understanding of a molecule, it is a descriptor-independent approach, and they have a solid mathematical background with different existing techniques to compare labeled graphs.<sup>46</sup> In this study we will use the noncontiguous atom matching structural similarity method (NAMS)<sup>24,47</sup> which has proven to be useful for comparing molecular structures. NAMS is defined on the annotated molecular graph, based on a recursive concept of graph similarity and an optimal alignment between atoms using a heuristic and a penalty function to account for the differences in both topological profiles and atoms/bonds characteristics (namely (1) the nature of the atomic elements; (2) whether the atom is chiral and its orientation; (3) whether the atom is part of at least one ring; (4) the bond order; (5) whether the bond is part of at least one ring; (6) whether the bond is aromatic; (7) whether a double bond has E-Z stereoisomerism). Again, the degree of similarity given by NAMS ( $s(x, y)$ ) was converted to a degree of dissimilarity ( $d(x, y)$ ) applying a monotonically decreasing transformation using the natural logarithm (eq 12).

**2.4. Model Validation.** The described approach requires a parametrization step in order to predict properties of new compounds by selecting the most similar compounds from the training set. For that purpose, a leave-one-out (LOO) cross validation approach was followed which comprises for  $n$  samples, the creation of  $n$  different learning sets and  $n$  different test sets by taking all the samples except one as learning set and the sample left out as test set. The goal of this cross-validation

is not only to select the best parameters but also to estimate the expected level of fit of the approach to new data that is not used in the training set and to statistically ensure that the approach is sound. The cross-validated correlation coefficient ( $q^2$ ), the percentage of compounds for which the estimation error is between an acceptable interval, and the root mean squared error (RMSE) are performed to determine the goodness of fit of the model.

Yet to adequately assess the validity of each model,<sup>11</sup> an external validation set is used to predict the properties of a set of instances, never used in the model development so as to adequately evaluate how well the model generalizes in a real world scenario. To assess the external predictive ability of each model, three statistics are used, namely the following: the predictive correlation coefficient ( $Q^2$ ), the percentage of compounds for which the estimation error is between an acceptable interval and the RMSE.

### 3. DATA

For this study two data sets of diverse chemical compounds were collected from the literature and preprocessed, one to predict biological activity, the dihydrofolate reductase inhibition activity, and another one to predict a physical property, the aqueous solubility.

**3.1. Data Set A - Dihydrofolate Reductase (DHFR) Inhibitors Activity.** Dihydrofolate reductase (DHFR) is an enzyme that catalyzes NADPH-dependent reduction of dihydrofolic acid to tetrahydrofolic acid. Inhibition of DHFR activity leads to a deficiency of thymidylate (dTMP), thus causing inhibition of cell growth.<sup>48</sup>

The data set of DHFR inhibitors activity for rat liver has been taken from the study of Sutherland et al.,<sup>49</sup> in which it has been used to access the predictive accuracy of various methods encoding the molecular structure and using five different machine learning algorithms. In this study, the half maximal inhibitory concentration ( $IC_{50}$ ) values are converted to the  $pIC_{50}$  scale ( $pIC_{50} = -\log_{10}(IC_{50})$ ) in terms of molar concentration (mol/L). This data set contains the  $pIC_{50}$  for 397 compounds (Supporting Information 1) which are divided in 237 compounds for the training set, 124 compounds for the test set, and 36 inactive compounds with indeterminate activities ( $IC_{50} > 10 \mu M$ ) and that are used to verify if the models can correctly identify inactive compounds. In this data set each experimental measurement is associated with a reference from 1991 to 2002 making it possible to perform a temporal selection of training and test data. Thus, simulating a real-world scenario taking into account the appearance of new chemical series since earlier data will be used to predict later data. For that purpose, the data set (including inactive compounds) was divided based on the reference year of the property experimental measurement for each compound: 313 measurements obtained from 1991 to 1998 as training data to predict 84 measurements obtained from 1999 to 2002.

**3.2. Data Set B - Aqueous Solubility.** Aqueous solubility is an important physical property of small organic molecules with pharmaceutical, environmental, and industrial applications.<sup>50</sup> It represents the maximum concentration of a chemical that will dissolve in pure water at a specified temperature.<sup>51</sup> Several computational methods have been used to predict aqueous solubility using the structure of the molecules, including group contribution methods (e.g. ref 52), thermodynamic calculations (e.g. ref 53), and quantitative structure–property relationships (e.g. refs 54 and 55).

**Table 1. Summary of the Best Results (Leave-One-Out Cross Validation) Obtained for Training and Testing Sets (Data Set A) Using Each Dissimilarity Matrix**

molecular representation	method	<i>neighs</i> <sup>a</sup>	type	<i>n</i> <sup>b</sup>	RMSE	%[−1.0, 1.0] <sup>c</sup>	<i>q</i> <sub>LOO</sub> <sup>2</sup> / <i>Q</i> <sup>2</sup>
molecular descriptors	<i>DistKrig</i>	236	train <sup>d</sup>	237	0.8430	77.87	0.5564
			test	124	0.9584	67.52	0.5043
fingerprints	<i>CoordKrig</i>	10	train <sup>d</sup>	237	0.8273	77.84	0.5728
			test	124	0.9696	69.20	0.4926
NAMS	<i>DistKrig</i>	5	train <sup>d</sup>	237	<b>0.8105</b>	<b>79.76</b>	<b>0.5900</b>
			test	124	<b>0.8609</b>	<b>73.41</b>	<b>0.6000</b>
NAMS (temporal)	<i>DistKrig</i>	5	train <sup>d,e</sup>	313	0.8738	76.39	0.6535
			test <sup>f</sup>	84	0.8940	72.35	0.6163

<sup>a</sup>Number of selected neighboring molecules for each prediction. <sup>b</sup>Total number of compounds in the set. <sup>c</sup>% of predictions with error between −1.0 and 1.0. <sup>d</sup>Leave-one-out cross-validated results. <sup>e</sup>Property measurements published between 1991 and 1998. <sup>f</sup>Property measurements published between 1999 and 2002.

The experimental aqueous solubility values used in this study for a total of 1291 diverse compounds (Supporting Information 2) were obtained from the literature,<sup>54,56</sup> which have been used for the development of several models (e.g. refs 54 and 57–60) and were divided into a training set of 1033 compounds and a test set of 258 compounds (by selecting every fifth compound into the test set) as suggested by other authors.<sup>57</sup> The model was also externally tested on a small set of 21 compounds that has been extensively used for solubility prediction method validation since its introduction by Yalkowsky.<sup>61</sup> The aqueous solubility experimental measurements are reported as the negative logarithm of the molar solubility in water (mol/L) at temperatures between 20 and 25 °C.

#### 4. RESULTS

In order to validate the hypothesis that it is possible to predict a property of interest based on structural similarity/dissimilarity between the molecules, as described above, the kriging algorithm was tested in data sets A and B with different parametrizations coupled with three distance matrices (based on molecular descriptors, fingerprints, and NAMS). The results obtained for the best parametrizations are summarized below; however, fully detailed results are provided in Supporting Information 3.

**4.1. Data Set A - Dihydrofolate Reductase (DHFR) Inhibitors Activity.** As a preliminary step it was analyzed, in light of the structural similarity principle, the capacity of each of the structural similarity methods in study to discriminate molecules with different activity value solely based on their structural distance. The results of this analysis are provided in Supporting Information 4, and, in general, it is possible to verify that for both Fingerprints and especially for NAMS there is a high probability for compounds that are very close to each other to have a small difference in the *pIC*<sub>50</sub> value, verifying the similarity principle. When using molecular descriptors, the relationship between the pairwise distance of the compounds and their difference in the *pIC*<sub>50</sub> is not as clear, even though there is a tendency for pairs of very dissimilar compounds to have higher differences in the *pIC*<sub>50</sub> value. Therefore, Fingerprints and NAMS are more likely to obtain a discriminating metric space to be interpolated by kriging.

One problem in retrospective QSAR studies is that data is randomly sampled so that virtually all scaffolds are represented in training and test sets, being limited when new compound scaffolds appear or when the structure of a test compound varies significantly. To overcome this situation, we also performed a validation of the method in a real-world context

of drug discovery using a temporal data selection, as explained above.

Table 1 summarizes the best results for the training (obtained with leave-one-out cross-validation) and testing sets with each dissimilarity matrix, selected based on the RMSE as well as results for temporal data selection using the best model settings.

The best model for the training set was obtained using NAMS to calculate the distance between molecules coupled with the method *DistKrig* and reached a leave-one-out cross-validated RMSE of 0.8105 which corresponds to a *q*<sub>LOO</sub><sup>2</sup> of 0.5900 with 79.76% of the compounds being predicted with an absolute error smaller than ±1 (Table 1). These results were obtained excluding one compound at the time for testing and using the most similar 5 compounds in the training set to predict its property. The results using NAMS to calculate the distance between the compounds tend to decrease with the increase of the number of compounds used to predict the property which complies with the similarity principle and also due to the high redundancy between the selected compounds interfering in the spatial correlation needed to construct the semivariogram. For the testing set and using the most similar 5 compounds in the training set to predict the property value, a RMSE of 0.8609 was obtained which corresponds to a *Q*<sup>2</sup> of 0.6000 with 73.41% of the compounds being predicted with an absolute error smaller than ±1 (Table 1).

The prediction results using molecular descriptors to calculate the distance between the compounds tend to improve with the number of compounds used to predict the property, which complies with the preliminary analysis (Supporting Information 4), since there is not a clear relationship between the pairwise distance and the *pIC*<sub>50</sub> difference, especially for smaller distances considering that there is no guarantee that molecular descriptors that contribute most to the calculation of the distance between molecules represent the substructures that most influence the property value, thus justifying the fact that using only the most similar compounds would not be enough, leading to semivariograms that are too irregular to be fitted with a linear function by not showing any spatial correlation.

The prediction results using fingerprints to calculate the distance between the compounds has a tendency to improve until 10 compounds are used, yet, from this point on this tendency reverts due to the high redundancy between the selected compounds interfering in the spatial correlation needed to construct the semivariogram.

For this molecular representation, the implementation *CoordKrig* that preliminarily transforms the distances between



**Table 2. Comparison of the Predictive Power of the Model Developed in This Study with Other Published Models with the Best Results (Selected by the Performance on the Training Set) Based on the Same Data Set (with Different Partitions of the Data into Training and Testing) by Comparative Molecular Field Analysis (CoMFA) with Partial Charge Calculation Method MMFF94, 3D Pharmacophores QSAR with Self-Consistent Atomic Property Fields by Optimization (SCAPFold), Hologram QSAR Coupled with Partial Least Squares (PLS) and 2.5D Descriptors Coupled with Neural Network (NN) Models**

reference	model	train set			test set			inactive set	
		$n^a$	$q^2$	RMSE	$n^a$	$Q^2$	RMSE	$n^a$	% inactives
our model	Kriging	237	0.59 <sup>b</sup>	0.81 <sup>b</sup>	124	0.60	0.86	36	97
Mittal 2009 <sup>63</sup>	CoMFA	397	0.69 <sup>c</sup>						
Totrov 2008 <sup>64</sup>	SCAPFold				124	0.64	0.84		
	HQSAR-PLS	237	0.69 <sup>d</sup>	0.71 <sup>d</sup>	124	0.63	0.84	36	92
Sutherland 2004 <sup>49</sup>	2.5D Descp-NN	237	0.61 <sup>d</sup>	0.79 <sup>d</sup>	124	0.42	1.05	36	83

<sup>a</sup>Total number of compounds in the set. <sup>b</sup>Leave-one-out cross-validated results. <sup>c</sup>Leave-several-out cross-validated results. <sup>d</sup>10-fold cross-validated results.

**Table 3. Summary of the Best Results (Leave-One-Out Cross Validation) Obtained for Training and Testing Sets (Data Set B) Using Each Dissimilarity Matrix**

molecular representation	method	neighs <sup>a</sup>	type	$n^b$	RMSE	%[−1.0, 1.0] <sup>c</sup>	$q^2_{LOO}/Q^2$
molecular descriptors	DistKrig	300	train <sup>d</sup>	1033	0.9475	76.09	0.7840
			test 1	258	0.8678	78.21	0.8143
			test 2	21	0.9105	72.33	0.7496
			train <sup>d</sup>	1033	1.2407	65.27	0.6296
fingerprints	CoordKrig	8	train <sup>d</sup>	1033	1.2407	65.27	0.6296
			test 1	258	1.1161	68.34	0.6929
			test 2	21	0.7871	77.48	0.8129
			train <sup>d</sup>	1033	0.7793	82.44	0.8537
NAMS	DistKrig	5	train <sup>d</sup>	1033	0.7793	82.44	0.8537
			test 1	258	0.8332	81.55	0.8288
			test 2	21	1.0941	73.18	0.6384
			train <sup>d</sup>	1033	0.7793	82.44	0.8537

<sup>a</sup>Number of selected neighboring molecules for each prediction. <sup>b</sup>Total number of compounds in the set. <sup>c</sup>% of predictions with error between −1.0 and 1.0. <sup>d</sup>Leave-one-out cross-validated results.

molecules into 2D coordinates, jitters duplicated coordinates, and uses a spherical function to fit the semivariogram showed advantages, which may be related to the existence of several pairs of structurally different compounds in the fingerprint-distance matrix with a score of zero (corresponding to 100% structurally similar compounds).

As already mentioned in the data description there were 36 inactive compounds with  $IC_{50}$  that have not been experimentally determined ( $>10 \mu M$  and artificially labeled with an observed value of 3.30) that were not included in the training or testing sets. The settings of the best model (NAMS coupled with *DistKrig* selecting the 5 most similar molecules) were used for predicting the  $pIC_{50}$  of these inactive compounds considering a threshold of 6.0 for discrimination between highly active and inactive compounds as previously used in other studies.<sup>49,62</sup> The mean prediction property for the inactive set is  $4.22 \pm 0.99$ , and only one inactive compound (id: 1-127977) is predicted with a higher value (7.7) than the defined threshold.

Table 1 also shows the results obtained applying temporal selection to divide the compounds in training and testing sets using the best model settings obtained with random data selection. This model used NAMS to calculate the distance between molecules coupled with the method *DistKrig* and reached for the training set (experimental measurements obtained between 1991 and 1998) a leave-one-out cross-validated RMSE of 0.8738 which corresponds to a  $q^2_{LOO}$  of 0.6535 with 76.39% of the compounds being predicted with an absolute error smaller than  $\pm 1$ . For the testing set and using the most similar 5 compounds in the training set to predict the property value, a RMSE of 0.8940 was obtained which

corresponds to a  $Q^2$  of 0.6163 with 72.35% of the compounds being predicted with an absolute error smaller than  $\pm 1$ . Although the results obtained using temporal selection cannot be directly compared with the previously obtained using random selection, it is possible to observe that these are of comparable quality in terms of predictive performance.

Although the main objective of this study is not to compare the predictive performance of the proposed methodology with the state-of-art QSPR/QSAR approaches, it is important to have a general idea of the best results obtained for the data set. Table 2 shows a summary of the best models found in the literature using data set A, which cannot be directly compared due to the fact that different partitions of the data and different validation methods are used in the different studies. However, it is possible to observe that the results of other authors for the test set are of comparable quality to the use of kriging coupled with NAMS.

**4.2. Data Set B - Aqueous Solubility.** As a preliminary step it was also analyzed in light of the structural similarity principle the capacity of each of the structural similarity methods in study to discriminate molecules with different aqueous solubility value solely based in their structural distance. The results of this analysis are provided in Supporting Information 4 and, in general, are similar to the results obtained for data set A, showing that Fingerprints and NAMS are more likely to obtain a discriminating metric space to be interpolated by kriging.

Table 3 summarizes the best results for the training (obtained with leave-one-out cross-validation) and testing sets with each dissimilarity matrix selected based on the RMSE.

**Table 4.** Comparison of the Predictive Power of the Model Developed in This Study with Other Published Models with the Best Results (Selected by the Performance on the Training Set) Based on Same Data Set (with Different Partitions of the Data into Training and Testing) by Multilinear Regression (MLR) Analysis and Artificial Neural Network (ANN) Models and Using Different Selections of Molecular Descriptors

reference	model	train set			test set 1			test set 2		
		$n^a$	$q^2/R^2$	RMSE	$n^a$	$Q^2$	RMSE	$n^a$	$Q^2$	RMSE
our model	Kriging	1033	0.85 <sup>b</sup>	0.78 <sup>b</sup>	258	0.83	0.83	21	0.64	1.09
Hou 2003 <sup>60</sup>	MLR	878	0.92	0.59	412	0.90	0.63	21	0.88	0.84
Yan 2003 <sup>59</sup>	MLR	797	0.79	0.93	496	0.82	0.79	21	0.56	1.20
	ANN	797	0.93	0.5	496	0.92	0.59	21	0.85	0.77
Liu 2001 <sup>57</sup>	ANN	1033	0.86 <sup>b</sup>	0.70 <sup>b</sup>	258	0.86	0.70	21	0.79	0.91
Tetko 2001 <sup>58</sup>	MLR	879	0.86	0.75	412	0.85	0.81	21	0.77	0.99
	ANN	879	0.93	0.53	412	0.90	0.66	21	0.89	0.67
Huuskonen 2000 <sup>54</sup>	MLR	884	0.89	0.67	413	0.88	0.71	21	0.83	0.88
	ANN	884	0.94	0.47	413	0.92	0.60	21	0.91	0.63

<sup>a</sup>Total number of compounds in the set. <sup>b</sup>Leave-one-out cross-validated results.

The best model for the training set was obtained using *DistKrig* coupled with NAMS to calculate the distance between molecules and reached a RMSE of 0.7796 which corresponds to a  $q^2$  of 0.8537 with 82.44% of the compounds being predicted with an absolute error smaller than  $\pm 1$  (Table 3). These results were obtained excluding one compound at the time for testing and using the most similar 5 compounds in the training set to predict its property. The results using NAMS to calculate the distance between the compounds tend to decrease with the increase of the number of compounds used to predict the property which complies with the similarity principle and also due to the high redundancy between the selected compounds interfering in the spatial correlation needed to construct the semivariogram. For testing set 1 and using the most similar 5 compounds in the training set to predict the property value, a RMSE of 0.8332 was obtained which corresponds to a  $Q^2$  of 0.8288 with 81.55% of the compounds being predicted with an absolute error smaller than  $\pm 1$  (Table 3). When evaluating the model with testing set 2 and using the most similar 5 compounds in the training set to predict the property value, a RMSE of 1.0941 was obtained which corresponds to a  $Q^2$  of 0.6384 with 73.18% of the compounds being predicted with an absolute error smaller than  $\pm 1$  (Table 3). For test set 2, there are two compounds with large errors which heavily penalize the predictive scores of this model due to the small number of compounds in the set.

Similarly to data set A, the results using *DistKrig* coupled with molecular descriptors to calculate the distance between the compounds tend to improve with the number of compounds used to predict the property, which complies with the preliminary analysis (Supporting Information 4), since there is an approximately inverse relationship between the pairwise distance and the aqueous solubility difference, especially for smaller distances since there is no guarantee that molecular descriptors that contribute most to the calculation of the distance between molecules represent the substructures that most influence the property value, thus justifying the fact that using only the most similar compounds would not be enough, leading to semivariograms that are too irregular to be fitted with a linear function by not showing any spatial correlation.

The prediction results using *CoordKrig* coupled with fingerprints to calculate the distance between the compounds tend to improve until 8 compounds are used; from this point on this tendency reverts due to the high redundancy and distance between the selected compounds. Due to the high

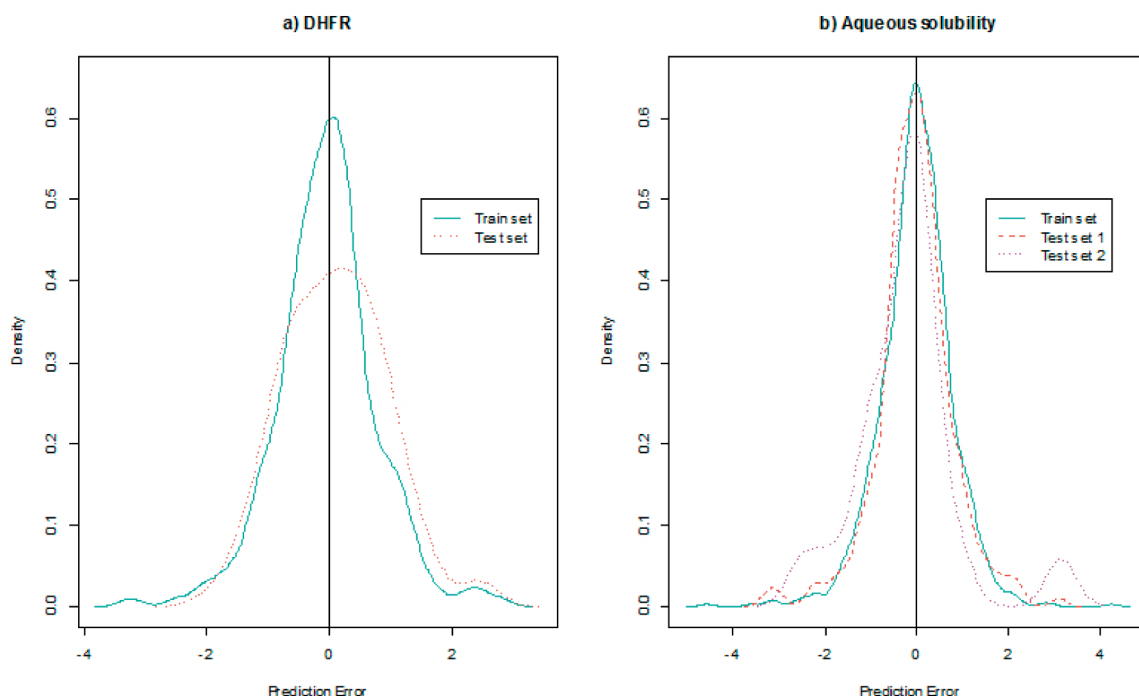
redundancy between the compounds in the data set when compared using fingerprints and to the existence of several pairs of structurally different compounds in the fingerprint-distance matrix with a score of zero, contrarily to the other distance methods *DistKrig* obtains low predictive power, and in some cases it is even impossible to apply it due to ill-conditioned distance matrices with a large condition number which means that such a matrix is almost singular and the computation of its inverse is not possible or it is prone to large numerical errors. The best predictive performance for test set 2 was obtained with fingerprints (Table 3), since there are some compounds in the testing set that are significantly different from all compounds in the training set and which benefit with a less granular similarity score.

Table 4 shows a summary of the best models found in the literature using data set B, which cannot be directly compared due to the fact that different partitions of the data and different validation methods are used in each study. However, it is possible to observe that the predictive performance of *DistKrig* coupled with NAMS is within the range of the performances obtained by the best models in the literature for this data set, especially when compared with Liu et al.<sup>57</sup> which uses the same partitions in training and testing sets and the same validation method.

## 5. DISCUSSION

Based on the results for both data sets, *DistKrig* coupled with NAMS for calculating the distance between the molecules is the natural choice for a final model, since for these data sets it is able to produce models with the best predictive performance with the smallest number of compounds needed to predict the property of interest. The obtained models are robust since similar predictive performances were obtained for both training and test sets. Also, the predictive performance of the models complies with the results obtained in the literature using typical QSPR/QSAR approaches; however, two of the great advantages of kriging are the exploration of a richer hypothesis space by creating local approximations for each test instance and the estimation of the prediction error for each predicted value in contrast to typical model-based approaches that commit to a single global hypothesis that covers the entire instance space and estimate only a global model prediction error. In this section the prediction error, its relationship with the similarity between the compounds, the kriging estimated variance and its relationship with the true prediction error, and





**Figure 1.** Density plots of the differences between the observed values and the predicted values for the train and test sets compounds using *DistKrig* coupled with NAMS to calculate the distance between compounds for the data sets a) DHFR and b) aqueous solubility.

the effect of the size of the training set in the predictive performance of the method will be analyzed in detail for the best model obtained for each data set.

**5.1. Prediction Error Analysis.** The prediction errors obtained for the train and test sets using NAMS to calculate the similarity between the compounds for both data sets A and B were further analyzed and are represented in Figure 1.

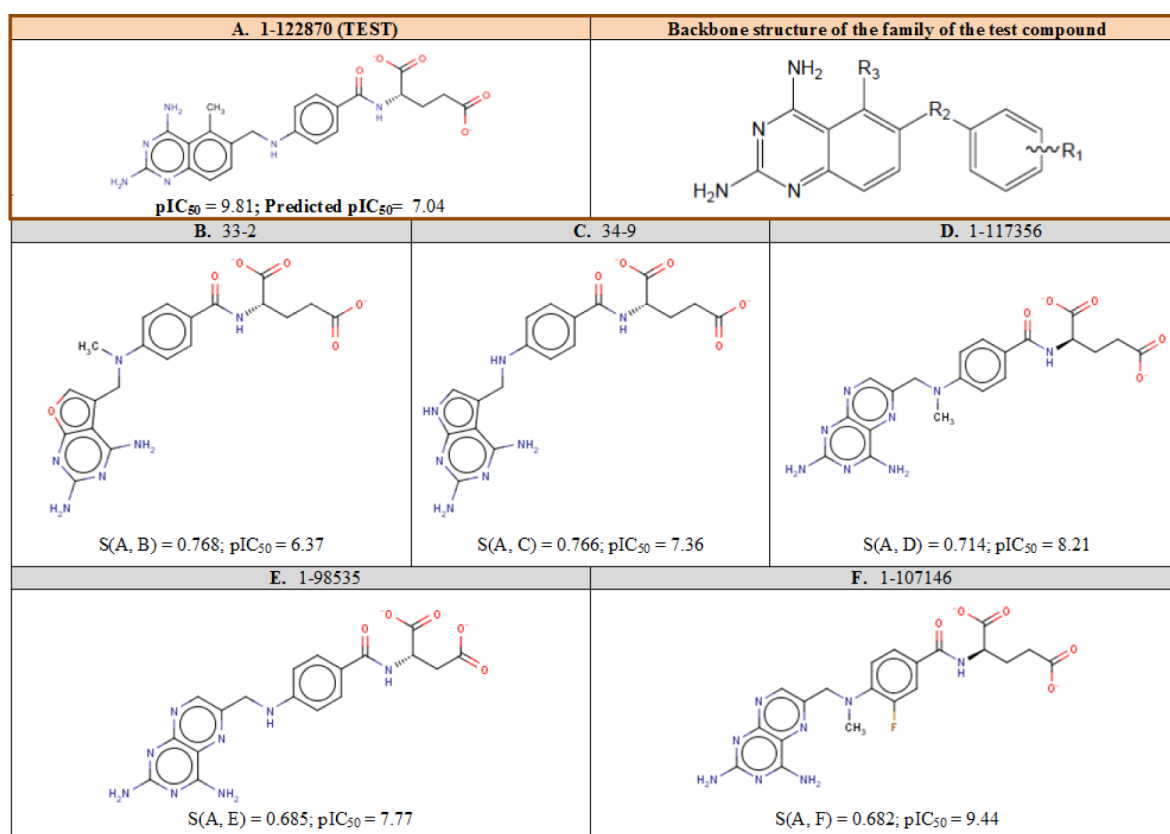
For data set A - DHFR, the model is predicting the  $pIC_{50}$  with a slight right bias (higher values than expected), and the most probable error for both training and testing sets is 0.196. Although for both, train and test sets, most errors (79.8% and 73.4%, respectively) are within the range of error values between  $-1.0$  and  $1.0$ , there are some higher errors that have a heavy weight in the RMSE calculation. The compound (ID: 1-233903) with the highest difference ( $-3.241$ ) between the observed and the predicted property value was inspected. Using the original ref 65 from which the value of the property for this compound was compiled,<sup>49</sup> it was found out that there was a mistake since the original value of the  $IC_{50}$  is  $220$  nM and not  $220$   $\mu$ M, corresponding then to a  $pIC_{50}$  of  $6.657$  instead of  $3.658$ . Therefore, the real difference between the observed and the predicted property value should be  $-0.242$  instead of  $-3.241$ . Considering the heavy weight of this prediction error and all predictions that were affected with the wrong value of the property of this compound, we can advocate that the RMSE would be significantly lower ( $0.773$  instead of  $0.811$  for the training set and  $0.859$  instead of  $0.861$  for the test set). We can also underpin the use of this method in curating data sets by the analysis of the prediction error in comparison with the most similar compounds.

The compound (ID: 1-122870) with the second higher difference ( $2.763$ ) between the observed and predicted property value was then selected and analyzed. Figure 2 depicts the structure of the test compound 1-122870 and the selected 5 training compounds. The  $pIC_{50}$  values of all training

compounds are lower than the observed  $pIC_{50}$  of the test compound, and, as expected, the predicted property is closer to the property of the most similar compounds. Figure 2 also shows the structure of the family of the test compound, which compared to the structures of the training compounds leads us to the conclusion that the high similarity scores between training and test compounds are due to a high similarity in the radicals R1 and R2 and in the pyrimidine ring; however, all these compounds lack the core quinazoline substructure (two fused six-membered simple aromatic rings: a benzene ring and a pyrimidine ring) which is determinant for the high potency of the test compound 1-122870.

The results of the study of the temporal data selection for this data set have confirmed the ability of the method demonstrated when applying random data selection. The training set was built using 26 different references, while the test set was built using 11 different references; however, the prediction errors do not show any trends based on reference or publication year. Most compounds with higher prediction errors are common using both random and temporal data selection.

For data set B - aqueous solubility (Figure 1 - b)), *DistKrig* coupled with NAMS model is predicting the aqueous solubility for training set and test sets 1 and 2 with the most probable errors  $-0.266$ ,  $-0.008$ , and  $-0.070$ , respectively. For the training and test 1 sets, the prediction error has a narrowed density curve, condensing 82.4% and 81.6% of the errors between  $-1.0$  and  $1.0$ , explaining its performance in relation to the distance matrices. However, for test set 2 only 73.2% of the prediction errors are between  $-1.0$  and  $1.0$ , and there are two peaks in the density curve on both sides representing high positive or negative errors, demonstrating that their decrease in the predictive performance in relation to test set 1 is due to few compounds that are predicted with a high error.



**Figure 2.** Structure of the test compound 1-122870 (A), the structure of the family of this compound and the 5 compounds selected for training (B–F). The similarity scores between the test and training compounds as well as the  $pIC_{50}$  values of each structure are also presented.

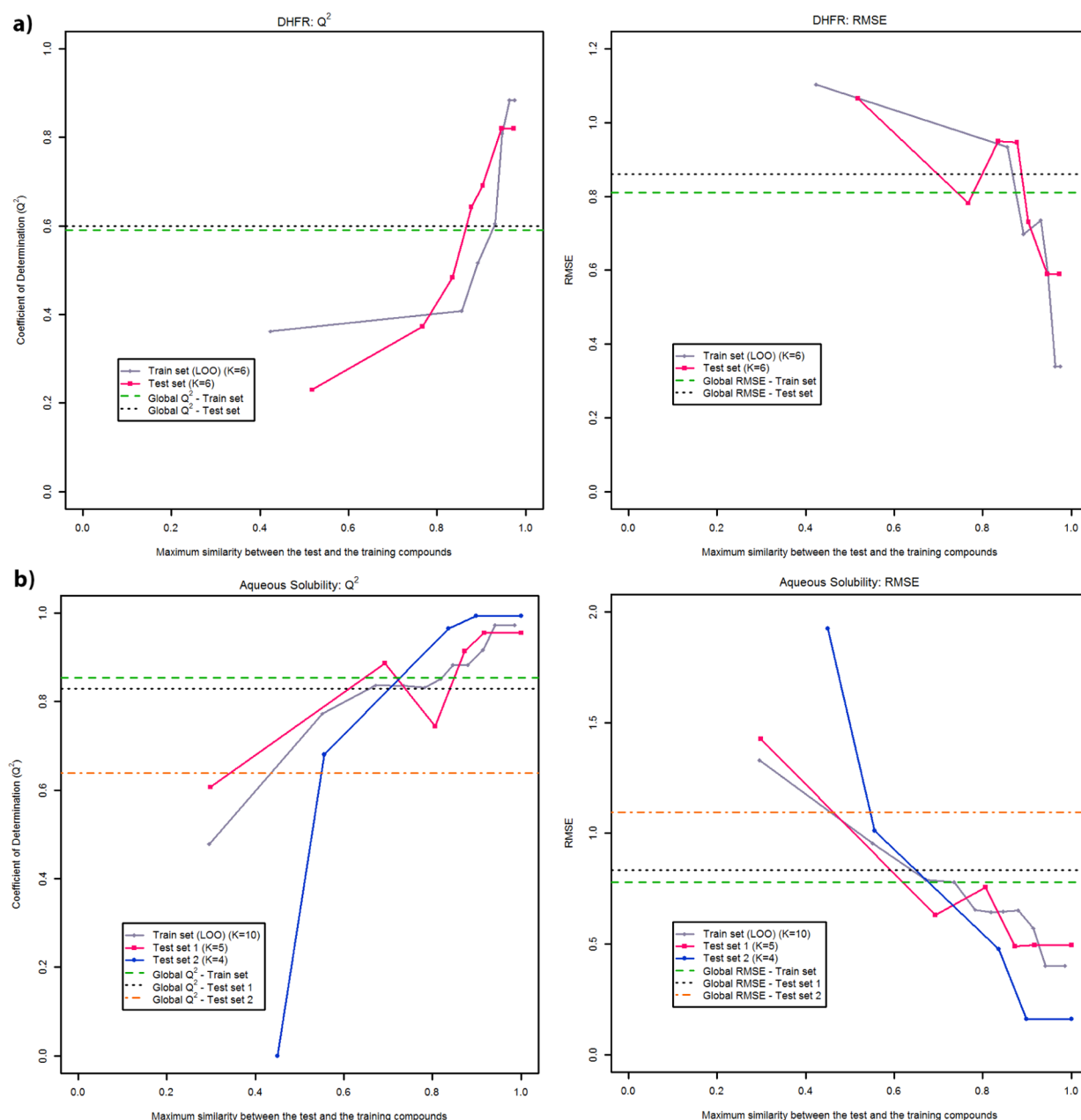
**5.2. Relationship between Prediction Error and Molecular Similarity.** Figure 3 presents the relationship between the maximum similarity among the test compounds and the compounds of the training set that were selected for predicting the property averaged by intervals and respective predictive performance using  $Q^2$  or RMSE for data set A and B. In general it is possible to verify that  $Q^2$  increases as the maximum similarity between the compounds increases and that RMSE decreases as the maximum similarity between the compounds increases. Figure 3 - a) shows that if a threshold of 94% for the most similar compound to the test compound selected from the training set in data set A is defined, which covers approximately 50% of all molecules, there is a high confidence in the predicted values with an expected  $Q^2 > 80\%$  and  $RMSE < 0.59$  which is significantly better than global  $Q^2$  of 59% and 60% and RMSE of 0.81 and 0.86 for the training and testing sets, respectively. It is important to highlight that for the last intervals of similarity the confidence in the results is very high, for example for the training set similarity scores range between 96.4% and 97.5% with an average  $Q^2$  of 88.4% and RMSE of 0.34.

For data set B (Figure 3 - b)) if a threshold of 84% for the most similar compound to the test compound selected from the training set is defined, which covers approximately 60% of the compounds, there is a high confidence in the predicted values with an expected  $Q^2 > 85\%$  for the training set and test set 1 and 96% for test set 2, and  $RMSE < 0.64$  for the training set and 0.47 for the testing sets 1 and 2. It is important to highlight that for the last intervals of similarity the confidence in the results is very high, for example for the training set similarity scores range between 96.4% and 97.5% with an average  $Q^2$  of

88.4% and RMSE of 0.34. Again it is important to emphasize the results for the last interval of similarity, for example for the test set 2 which *DistKrig* coupled with NAMS obtained worse results ( $Q^2 = 64\%$  and  $RMSE = 1.19$ ) than *CoordKrig* coupled with Fingerprints ( $Q^2 = 81\%$  and  $RMSE = 0.79$ ); however, for the last interval of similarity scores which ranges between 89.9% and 99.0%, the average  $Q^2$  is 99.0% and the RMSE is 0.16. These observations lead us to the conclusion that the existence of at least one compound in the training set that has a high similarity with the test compound allows making predictions with high confidence and minimal error. A complementary observation is that this method is able to identify regions of the molecular space that are lacking compounds with experimentally determined properties. These regions are ideal targets for experimental determination of properties of new molecules, which in turn will provide a broader coverage of the molecular space, resulting in a better model performance.

**5.3. Kriging Estimated Variance and Its Relationship with Prediction Errors.** The kriging estimated variance of each prediction depends on the arrangements of the observed values with respect to each other and with the location of the test compound in relation to the training compounds, and it is completely independent of the true property of the test compound. Therefore, kriging provides the estimated variance at every estimated point, which is a great indicator of the accuracy of the estimated value and signs areas for which more experimental measures are needed.

In general there is a strong correlation between the kriging estimated error and the absolute true predicted error of each compound of data sets A and B. There are some cases for which the kriging estimated error is higher than the true prediction

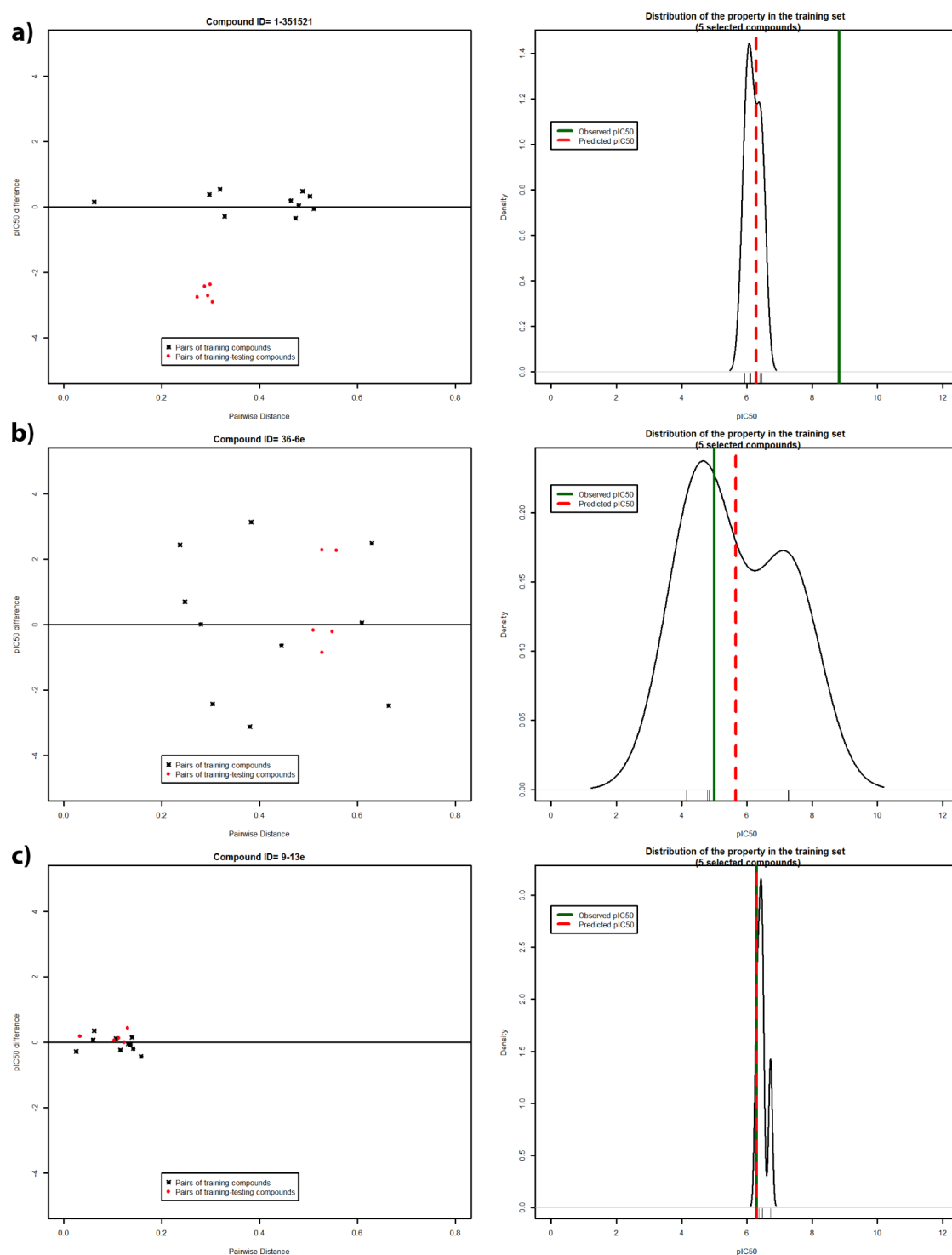


**Figure 3.** Plots showing the relationship between the maximum similarity between test compounds and compounds of the training set that were selected for predicting the property and predictive performance using  $Q^2$  or RMSE for data sets A and B. The points marked in the lines represent the boundaries of the interval (K) for which the predictive performance metric is being averaged. The horizontal lines highlight the global RMSE or  $Q^2$  obtained using all compounds. a) Data set A: The similarity between the compounds and respective  $Q^2$  or RMSE was averaged by 6 intervals containing 39 or 40 compounds each for training LOO and 20 or 21 compounds for testing. b) Data set B: The similarity between the compounds and respective  $Q^2$  or RMSE was averaged by 10 intervals containing 103 or 104 compounds each for training LOO, 5 intervals containing 51 or 52 compounds each for testing set 1 and 4 intervals containing 5 or 6 compounds each for testing set 2.

error which do not represent a problem, since it is in agreement with the notion of confidence interval; however, there are also some cases that deserve further investigation since the kriging estimated error is smaller than the true prediction error. For that purpose, three compounds were identified from test set A representing three different situations: a) a compound (ID: 1-351521) which has a high true prediction error of 2.546 and a low kriging estimated error of 0.170; b) a compound (ID: 36-6e) which has a low true prediction error of 0.649 and a high kriging estimated error of 1.834; and finally c) a compound (ID: 9-13e) which has a low true prediction error of 0.007 and a low kriging estimated error of 0.093. Figure 4 shows the

distance between the training and test compounds versus the  $pIC_{50}$  difference for each presented situation as well as the distribution of the properties in the compounds selected for training, the observed and predicted  $pIC_{50}$  of the test compound. Situation a) is represented in Figure 4 - a) which shows that the distance between the set of selected training compounds and the test compound 1-351521 is identical for all of them and relatively low (approximately 0.3). Also between the training compounds the property value is almost invariant. Therefore, this situation is translated in a low kriging estimated error, although due to small differences in key groups (similar to the case presented in Figure 2), the property of the test

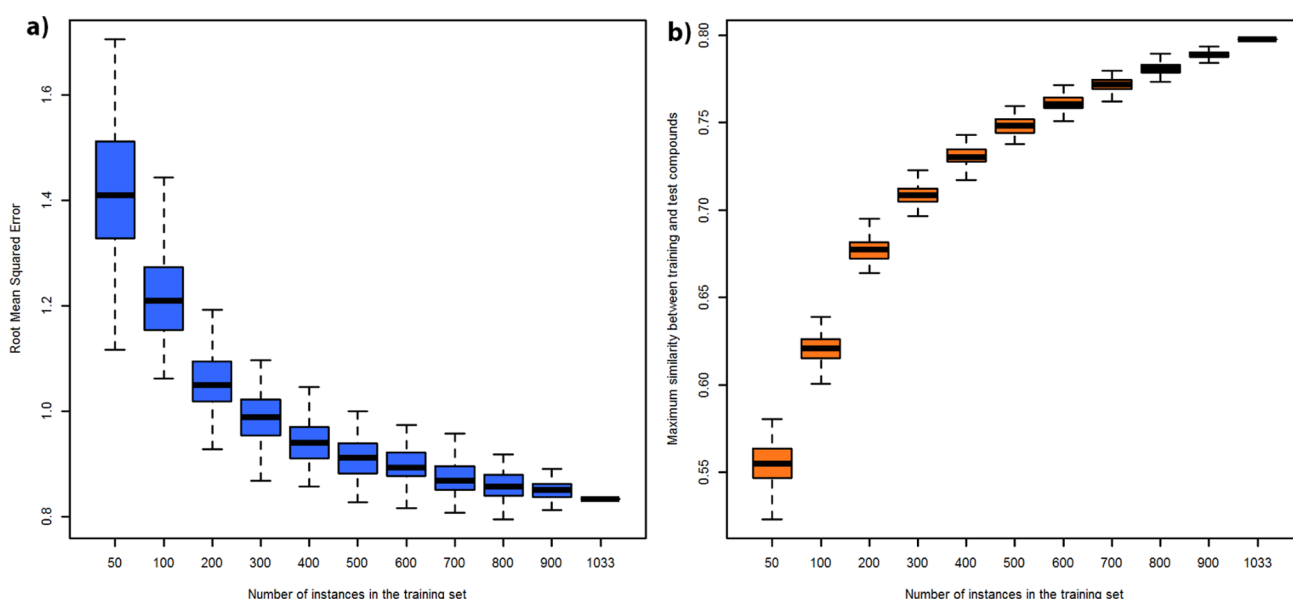




**Figure 4.** On the right side are represented the plots of the differences between the observed  $pIC_{50}$  values and the pairwise distance for all unique pairs between the 5 compounds selected for training (black points) and between the test compound (a) 1-351521, b) 36-6e, and c) 9-13e and the training compounds (red points) in data set A. On the left side are represented the density plots of the distribution of the property values in the selected training set (5 compounds) and the observed (green line) and predicted value (red dashed line) for the property of the test compound a) 1-351521, b) 36-6e, and c) 9-13e in data set A.

compound is significantly different from the properties observed for the training compounds. The situation b) is represented in Figure 4 - b), which is not as fallacious as situation a), since the true predicted error is within the estimated kriging error of  $[-1.834, 1.834]$ . The set of training compounds selected to predict the property of 36-6e are relatively distant in relation to each other and to the test

compound and the distribution of the property values in the compounds selected for train show a high range, thus a high estimated kriging error was expected. The situation c) is represented in Figure 4 - c), and it is an example of an ideal situation, since the true prediction error is within a narrow estimated kriging error interval of  $[-0.093, 0.093]$ . The set of training compounds used to predict the property of 9-13e is



**Figure 5.** Effect of the training set (data set B) size on **a)** test set predictive results using the Root Mean Squared Error (RMSE) and **b)** maximum similarity between training neighborhood and test compounds. Each of these training subsets were created by random sampling without replacement  $n$  compounds 100 times.

similar to the test compound and has a low extent of distance scores as well as comparable differences in the property value.

**5.4. Effect of the Training Set Size on the Predictive Results.** The obtained models are limited in applicability by the data from which they are constructed; however, this issue is seldomly addressed by reporting the kriging estimated variance of each prediction as a measure of extrapolation - high estimated kriging variances are obtained for compounds that are out of the applicability of the model by the lack of similar “neighbors”, while low estimated kriging variances indicate that the model is able to predict that property value with high confidence. As already shown, the methodology provides improved predictive results with the increase of similarity between the training and test compounds. Based on these results, it is likely that large data sets will improve the predictive performance of the method, as the metric space has more instances the probability of finding training compounds that are more similar to the test compound increases. To test this hypothesis, the training set of data set B ( $n = 1033$ ) was used to create smaller data sets with  $n = \{50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1033\}$  compounds. Each of these subsets were created by randomly sampling  $n$  compounds 100 times. The predictive results (using RMSE) using each of these subsets to predict the aqueous solubility of the test set compounds ( $n = 258$ ) are summarized in Figure 5 a). It is possible to observe that the predictive results increase as the number of compounds used as a training set increases approximately following a power law distribution; however, the gain in the predictive results becomes asymptotically reduced with the increase of the number of compounds in the training set. As expected, in Figure 5 b) it is possible to observe that the maximum similarity between training neighborhood and test compounds increases as the number of training compounds increases accompanying the improvement of the predictive results. It is important to note that this method always outputs the same predictive results as long as the same training set is used, as it can be observed when 1033 compounds are used for training.

**5.5. Assumptions and Limitations.** Kriging is based on a statistical method which creates an interpolated map and output error map with the standard errors of the estimates, as such, the assumptions of the method should be considered carefully. The main assumption is stationarity (spatial homogeneity). If the change of data points from one neighborhood to the next is too abrupt there may be discontinuities even though the actual phenomenon is continuous. If there is a spatial dependence between points that are closer together, the semivariogram will have small semivariance, and this semivariance is expected to increase with distance. If this assumption is held, just a few kriging model parameters have to be estimated from the data to make optimal predictions and valid statistical inferences. Therefore, the similarity metric used to map the compounds in the metric space should be able to discriminate the compounds according to the similarity principle. Furthermore, for the data sets in this study the assumption of being quasi-stationary does not apply to the entire data set but only to the search neighborhood under which the estimation model is fitted. Most of these subareas meet the local quasi-stationary assumption (homogeneity and density compromise) when analyzing the pattern of the semivariogram cloud. No properties are guaranteed, when the wrong variogram is used; however, typically still a good interpolation is achieved even in cases of no spatial dependence in which the kriging interpolation is only as good as the arithmetic mean. The error map reflects data locations, and it depends entirely on data configuration and semivariance function; therefore, discontinuities will also be reflected in the kriging estimated variance. Furthermore, the use of small neighborhoods is also advantageous in terms of computation since the basic math of this methodology involves the inverse of an  $n \times n$  matrix, where  $n$  is the number of data points used to predict the properties of a new compound.

As stated above, this method, as most QSPR/QSAR methods, relies on the similarity property principle, which states that molecules that are structurally similar are likely to have similar properties. However, there are some exceptions to

S(A, B)	1259	1261	1257	1256	1277	514	Log(S)	Neighborhood
1259	1						-4.40	
1261	0.937	1					-3.59	
1257	0.902	0.861	1				-6.70	
1256	0.885	0.886	0.865	1			-4.12	
1277	0.808	0.811	0.795	0.886	1		-4.46	
514	0.796	0.794	0.800	0.835	0.790	1	-3.99	

**Figure 6.** Example of a situation (data set B - aqueous solubility) in which the most similar compounds to the test compound (ID: 1259) have considerable differences in the property value. The relationship between training compounds leads to a correct property interpolation of the test compound with a low absolute prediction error of 0.27.

this similarity principle, most obviously in the case of activity cliffs where even a small structural change can be associated with a dramatic property shift. One of the advantages of OK in relation to other techniques based on distance (e.g., K-Nearest Neighbors) is that it considers not only the distance between the test and training compounds but also the distance between all training compounds. If the data set is broad, then this problem is amended because the relationship between training compounds might indicate this discontinuity. The example depicted in Figure 6 shows a situation in which the most similar compounds have considerable differences in the property value; however, the relationship between all training compounds leads to a correct property interpolation of the test compound with a low absolute prediction error of 0.27, which is in agreement with the expected error of approximately 0.25 for the maximum similarity level between training and test compounds.

Nevertheless, there are some rare situations in which the relationships between training compounds and the test compound are not enough to correctly interpolate properties of compounds that have small structural differences associated with a dramatic property shift. An example of such a situation is depicted in Figure 7, where the most similar compounds are all highly active and the test compound is highly similar, yet inactive. The property of this compound is predicted with a high prediction error of 4.40 when compared with the expected prediction error of approximately 0.81 for the maximum similarity level between training and test compounds. In such situations a specific weighting scheme for small structural modifications that are able to produce high property differences could be advantageous.

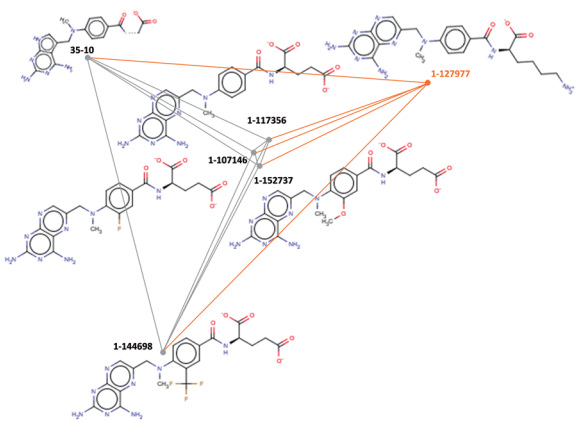
Even so, the principle is a very useful one for which there is substantial supporting evidence in the predictive results that were obtained, and in large part these capabilities of the model are enhanced by the degree of structural similarity between the training-set and test-set molecules. Therefore, the quality and coverage of the training set is a key element for the predictive capabilities of the method.

## CONCLUSIONS

In this work, we have proposed a new method for predicting chemical, physical, or biological properties of chemical compounds solely based on structural similarity functions to define the chemical space that is explored by kriging models, in order to predict unmeasured properties of compounds.

The overall predictive results of the proposed methodology applied to two different data sets were good and within the confidence margins of other studies found in the literature that used the same data sets. However, the presented approach showed several advantages relatively to other QSPR/QSAR approaches, namely (1) it makes use of the dissimilarity/similarity between the compounds and it is not necessary to use any feature selection and no prior knowledge of the problem or property is necessary, thus it may be applicable to most QSAR/QSPR studies directly and to any possible compound (even to compounds that were never synthesized) as long as its structure is known, (2) the similarity map that positions each molecule of the data set in the chemical space can be used to predict any chemical or biological property of the compounds as long as experimental data is available, (3) it is possible to identify the situations for which prediction errors are deemed to be higher (measure of extrapolation) by estimating the kriging variance for each prediction - high estimated kriging variances are obtained for compounds that are out of the applicability of the model by the lack of similar neighbor compounds, while low estimated kriging variances indicate that the model is able to predict that property value with high confidence, (4) the model is readily understandable rather than a black box model, as it is possible to verify which of the compounds more directly impact the current estimation, (5) new compounds can be easily included as well as removed from the pool of training compounds since in this approach the target function is approximated locally for each test compound instead of an overall model that requires retraining each time the training set is changed, (6) the method can be applied to data sets of any size, however the predictive results are more likely to improve with the increase of the number of training instances as the



S(A, B)	1-127977	1-117356	1-107146	1-152737	1-144698	35-10	pIC <sub>50</sub>	Neighborhood
1-127977	1						3.30	
1-117356	0.815	1					8.21	
1-107146	0.780	0.951	1				9.44	
1-152737	0.773	0.899	0.898	1			8.96	
1-144698	0.735	0.872	0.880	0.832	1		8.81	
35-10	0.700	0.847	0.806	0.766	0.743	1	7.22	

**Figure 7.** Example of a situation (data set A - DHFR inhibitors activity) in which the most similar compounds to the test compound (ID: 1-127977) are all highly active, yet the test compound is inactive. The relationship between training compounds does not lead to a correct property interpolation of the test compound (7.70) with a high prediction error of 4.40.

probability of finding neighbor compounds with higher similarity increases, and (7) searches for the relationship among measured properties (richer hypothesis space) rather than approximate the modeled system by fitting the parameters of the selected basis functions (single hypothesis space). This approach can simultaneously solve multiple problems and deal successfully with changes in the problem domain.

The results of the application of this methodology also showed that the existence of at least one compound in the training set that has a high similarity with the test compound allows making predictions with higher confidence and reduced error. Another important conclusion is that the current approach can be used to guide the extension of the training set and exploration of new promising regions within the molecular space by suggesting new molecules that can be used as seed compounds for experimental property determination, which in turn will improve the model quality by providing a broader coverage of the molecular space, as well as being used for data set curation proposes by analyzing the prediction error and the structure/property of the selected neighbor compounds.

The estimated variance resulting from kriging for each prediction showed a strong correlation with the true prediction error, which proves that it can be used as a quality measure of each kriging prediction since it provides a confidence interval in the predicted value. It is our conviction that kriging estimated variance can also be used to interactively determine the number of compounds that should be used to make a prediction based on the minimization of the kriging estimated variance.

For both data sets that were tested the similarity function NAMS to map the compounds in the chemical space using random or temporal data selection yields better validated predictions with a smaller number of compounds (as nearest neighbors) for each prediction than using molecular descriptors or fingerprints. NAMS yielding better results was expected since a preliminary analysis comparing the pairwise distance between the compounds and their property difference showed that NAMS was able to discriminate the compounds better in

accordance with the similarity principle: similar compounds tend to have similar property values and *vice versa*. The application of a feature selection technique prior to the calculation of molecular similarity using molecular descriptors could be advantageous; however, the objective of the study was to build a universal map of structural relationships in the chemical space that was not dependent on the property in study. The analysis of the prediction errors showed that a similarity metric that would weight differently substructures that have a higher (positive or negative) impact in the property value could improve the predictive performance of the method. However, any other similarity/dissimilarity approach can be applied in this methodology. In general, the predictive results are affected by redundancy between the compounds and by predictive maps that present patterns with several pairs of compounds at the same distance but with considerably different values of property and *vice versa*.

Ongoing research is being performed aiming in automatically search the neighborhood of a compound and determine the optimal number of neighbors which can be used to predict its property with a minimized prediction error. This study was limited to predict properties based solely on the structure of the compound as it may be used for any possible compound (even compounds that were never synthesized) and do not require any knowledge on their bioactivity or chemical/physical properties. Future work includes the development of a weighting schema in the similarity function to include both structural similarity and property profiles (using methods such as High Throughput Screening Fingerprints<sup>66</sup> or Similarity Ensemble Approaches<sup>67</sup>) in order to accurately predict similar properties of compounds even when molecules are not structurally similar.

## ■ ASSOCIATED CONTENT

### ⑤ Supporting Information

*Supporting Information 1 - Data set A - Dihydrofolate reductase (DHFR) inhibitors activity:* A complete table with SMILES, experimental values for the DHFR inhibition activity

in the rat liver, references, and publication year. *Supporting Information 2 - Data set B - Aqueous Solubility*: A complete table with compound name, SMILES, and aqueous solubility experimental values for training and testing sets. *Supporting Information 3 - Detailed Table of Results*: A complete table with the predictive results obtained using *DistKrig* and *CoordKrig* coupled with Molecular Descriptors, Fingerprints, and NAMS and with a different number of compounds selected for training. *Supporting Information 4 - Discriminate molecules with different property/activity value based on structural dissimilarity*: The capacity of each of the structural similarity methods in study (based on molecular descriptors, fingerprints, and NAMS) to discriminate molecules with different activity/property value solely based on their structural distance are analyzed. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [ateixeira@lasige.di.fc.ul.pt](mailto:ateixeira@lasige.di.fc.ul.pt).

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge the Portuguese Fundação para a Ciência e Tecnologia for the Multiannual Funding Programme of the LaSIGE laboratory and doctoral grant SFRH/BD/64487/2009.

## REFERENCES

- (1) Bachrach, S. Chemistry publication - making the revolution. *J. Cheminf.* [Online] **2009**, 1, Article 2. <http://www.jcheminf.com/content/1/1/2> (accessed September, 2013).
- (2) Katritzky, A. R.; Maran, U.; Lobanov, V.; Karelson, M. Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 1–18.
- (3) Katritzky, A. R.; Fara, D. C.; Petrukhin, R. O.; Tatham, D. B.; Maran, U.; Lomaka, A.; Karelson, M. The Present Utility and Future Potential for Medicinal Chemistry of QSAR/QSPR with Whole Molecule Descriptors. *Curr. Top. Med. Chem.* **2002**, 24, 1333–1356.
- (4) Doucet, J. P.; Panaye, A. QSARs in Data Mining. In *Three dimensional QSAR - Applications in Pharmacology and Toxicology; QSAR in Environmental and Health Sciences*; CRC Press: Boca Raton, 2011; pp 253–266.
- (5) Liu, Y. A Comparative Study on Feature Selection Methods for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 1823–1828.
- (6) Gonzalez, M. P.; Teran, C.; Saiz-Urra, L.; Teijeira, M. Variable Selection Methods in QSAR: An Overview. *Curr. Top. Med. Chem.* **2008**, 8, 1606–1627.
- (7) Teixeira, A.; Leal, J.; Falcao, A. Random forests for feature selection in QSPR Models - an application for predicting standard enthalpy of formation of hydrocarbons. *J. Cheminf.* [Online] **2013**, 5, Article 9. <http://www.jcheminf.com/content/5/1/9> (accessed September, 2013).
- (8) Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **2007**, 13, 3494–504.
- (9) Puzyn, T.; Leszczynski, J.; Cronin, M. T. D. *Recent Advances in QSAR Studies: Methods and Applications*; Springer: London, 2009.
- (10) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, 20, 241–266.
- (11) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, 29, 476–488.
- (12) Tetko, I.; Poda, G.; Ostermann, C.; Mannhold, R. Accurate In Silico log P Predictions: One Can't Embrace the Unembraceable. *QSAR Comb. Sci.* **2009**, 28, 845–849.
- (13) Oprea, T. L.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, 3, 157–166.
- (14) Walker, J.; Carlsen, L.; Jaworska, J. Improving Opportunities for Regulatory Acceptance of QSARs: The Importance of Model Domain, Uncertainty, Validity and Predictability. *QSAR Comb. Sci.* **2003**, 22, 346–350.
- (15) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
- (16) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *J. Med. Chem.* **2004**, 47, 4891–4896.
- (17) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity - a Review. *QSAR Comb. Sci.* **2003**, 22, 1006–1026.
- (18) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B.-T. Molecular similarity and diversity in chemoinformatics: From theory to applications. *Mol. Diversity* **2006**, 10, 39–79.
- (19) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, 39, 3049–3059, DOI: 10.1021/jm960290n.
- (20) Gute, B. D.; Basak, S. C. Molecular similarity-based estimation of properties: a comparison of three structure spaces. *J. Mol. Graphics Modell.* **2001**, 20, 95–109.
- (21) Basak, S. C.; Gute, B. D.; Mills, D.; Hawkins, D. M. Quantitative molecular similarity methods in the property/toxicity estimation of chemicals: a comparison of arbitrary versus tailored similarity spaces. *J. Mol. Struct.* **2003**, 622, 127–145.
- (22) Li, C.; Colosi, L. M. Molecular similarity analysis as tool to prioritize research among emerging contaminants in the environment. *Sep. Purif. Technol.* **2012**, 84, 22–28.
- (23) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 379–386, DOI: 10.1021/ci970437z.
- (24) Teixeira, A. L.; Falcao, A. O. Noncontiguous atom matching structural similarity function. *J. Chem. Inf. Model.* **2013**, 53, 2511–2524.
- (25) Tobler, W. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **1970**, 46, 234–240.
- (26) Matheron, G. *Les Variables Regionalisees et Leur Estimation*; Masson et Cie: Paris, 1965.
- (27) Isaaks, E.; Srivastava, R. *An Introduction to Applied Geostatistics*; Oxford University Press: New York, 1989.
- (28) Davis, J. *Statistics and Data Analysis in Geology*; John Wiley and Sons: New York, 2002.
- (29) Burden, F. R. Quantitative Structure-Activity Relationship Studies Using Gaussian Processes. *J. Chem. Inf. Model.* **2001**, 41, 830–835.
- (30) Fang, K.-T.; Yin, H.; Liang, Y.-Z. New Approach by Kriging Models to Problems in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 2106–2113.
- (31) Obrezanova, O.; Csanyi, G.; Gola, J. M. R.; Segall, M. D. Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties. *J. Chem. Inf. Model.* **2007**, 47, 1847–1857.
- (32) Hawe, G. I.; Alkorta, I.; Popelier, P. L. A. Prediction of the Basicities of Pyridines in the Gas Phase and in Aqueous Solution. *J. Chem. Inf. Model.* **2010**, 50, 87–96.
- (33) Sun, Y.; Brown, M.; Prapopoulou, M.; Davey, N.; Adams, R.; Moss, G. The application of stochastic machine learning methods in the prediction of skin penetration. *Appl. Soft Comput.* **2011**, 11, 2367–2375.
- (34) Negreiros, J.; Painho, M.; Aguilar, F.; Aguilar, M. Geographical Information Systems Principles of Ordinary Kriging Interpolator. *J. Appl. Sci.* **2010**, 10, 852–867.
- (35) Bohling, G. Introduction to geostatistics and variogram analysis. 2005 [Online]. <http://people.ku.edu/~gbohling/cpe940/Variograms.pdf> (accessed November, 2013).

- (36) R Core Team, *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
- (37) Ribeiro, P. J., Jr.; Diggle, P. J. *geoR: a package for geostatistical analysis*. *R-NEWS* **2001**, *1*, 14–18.
- (38) Diggle, P.; Ribeiro, P. J. *Model Based Geostatistics*; Springer: New York, 2007.
- (39) Venables, W.; Ripley, B. *Modern Applied Statistics with S*; Springer: New York, 2002.
- (40) Chambers, J. M. In *Statistical Models in S*; Hastie, T. J., Ed.; Wadsworth and Brooks/Cole: Pacific Grove, California, 1992.
- (41) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.
- (42) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. Virtual computational chemistry laboratory - design and description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453–63.
- (43) VCCLAB, Virtual Computational Chemistry Laboratory. 2005. <http://www.vcclab.org> (accessed in September 2013).
- (44) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.
- (45) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An open chemical toolbox. *J. Cheminf. [Online]* **2011**, *3*, Article 33. <http://www.jcheminf.com/content/3/1/33> (accessed March, 2013).
- (46) Ehrlich, H.-C.; Rarey, M. Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 68–79.
- (47) Teixeira, A. L.; Leal, J. P.; Falcao, A. O. *Automated Identification and Classification of Stereochemistry: Chirality and Double Bond Stereoisomerism*. Technical Report; Department of Informatics, Faculty of Sciences, University of Lisbon: 2013; arXiv:1303.1724.
- (48) Chen, M. J.; Shimada, T.; Moulton, A. D.; Cline, A.; Humphries, R. K.; Maizel, J.; Nienhuis, A. W. The functional human dihydrofolate reductase gene. *J. Biol. Chem.* **1984**, *259*, 3933–43.
- (49) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A comparison of methods for modeling quantitative structure-activity relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.
- (50) Yalkowsky, S. H. *Solubility and Solubilization in Aqueous Media*; Oxford University Press: New York, 1999.
- (51) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235–249.
- (52) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Representation/Prediction of Solubilities of Pure Compounds in Water Using Artificial Neural Network - Group Contribution Method. *J. Chem. Eng. Data* **2011**, *56*, 720–726.
- (53) Palmer, D. S.; Llinas, A.; Morao, I.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O. Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle. *Mol. Pharmaceutics* **2008**, *5*, 266–279.
- (54) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777.
- (55) Salahinejad, M.; Le, T. C.; Winkler, D. A. Aqueous Solubility Prediction: Do Crystal Lattice Interactions Help? *Mol. Pharmacol.* **2013**, *10*, 2757–2766.
- (56) Cheminformatics.org: QSAR datasets - Huuskonen Data Set. <http://cheminformatics.org/datasets/huuskonen/index.html> (accessed July, 2013).
- (57) Liu, R.; So, S.-S. Development of Quantitative Structure-Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633–1639.
- (58) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488–1493.
- (59) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429–434.
- (60) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2003**, *44*, 266–275.
- (61) Yalkowsky, S. H.; Banerjee, S. *Aqueous solubility: Methods of estimation for organic compounds*; Marcel Dekker: New York, 1992.
- (62) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure–Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906–1915.
- (63) Mittal, R. R.; Harris, L.; McKinnon, R. A.; Sorich, M. J. Partial Charge Calculation Method Affects CoMFA QSAR Prediction Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 704–709.
- (64) Totrov, M. Atomic Property Fields: Generalized 3D Pharmacophoric Potential for Automated Ligand Superposition, Pharmacophore Elucidation and 3D QSAR. *Chem. Biol. Drug. Des.* **2008**, *71*, 15–27.
- (65) Broughton, M.; Queener, S. Pneumocystis carinii dihydrofolate reductase used to screen potential antipneumocystis drugs. *Antimicrob. Agents Chemother.* **1991**, *35*, 1348–55.
- (66) Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.* **2012**, *7*, 1399–1409.
- (67) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.