molecules

MDPI

Article

Analysis and Comparison of Vector Space and Metric Space Representations in QSAR Modeling

Samina Kausar [1,2] **and Andre O. Falcao** [1,2,*]

[1] LASIGE, Faculdade de Ciencias, Universidade de Lisboa, 1749-016 Lisboa, Portugal; saminakausar.bioinfo@gmail.com

[2] BioISI—Biosystems & Integrative Sciences Institute, Faculdade de Ciencias, Universidade de Lisboa, 1749-016 Lisboa, Portugal

* Correspondence: aofalcao@ciencias.ulisboa.pt

Received: 25 February 2019; Accepted: 26 April 2019; Published: 30 April 2019

Abstract: The performance of quantitative structure–activity relationship (QSAR) models largely depends on the relevance of the selected molecular representation used as input data matrices. This work presents a thorough comparative analysis of two main categories of molecular representations (vector space and metric space) for fitting robust machine learning models in QSAR problems. For the assessment of these methods, seven different molecular representations that included RDKit descriptors, five different fingerprints types (MACCS, PubChem, FP2-based, Atom Pair, and ECFP4), and a graph matching approach (non-contiguous atom matching structure similarity; NAMS) in both vector space and metric space, were subjected to state-of-art machine learning methods that included different dimensionality reduction methods (feature selection and linear dimensionality reduction). Five distinct QSAR data sets were used for direct assessment and analysis. Results show that, in general, metric-space and vector-space representations are able to produce equivalent models, but there are significant differences between individual approaches. The NAMS-based similarity approach consistently outperformed most fingerprint representations in model quality, closely followed by Atom Pair fingerprints. To further verify these findings, the metric space-based models were fitted to the same data sets with the closest neighbors removed. These latter results further strengthened the above conclusions. The metric space graph-based approach appeared significantly superior to the other representations, albeit at a significant computational cost.

Keywords: QSAR modeling; non-contiguous atom matching structure similarity—NAMS; metric space; vector space; PCA; feature selection; random forest; support vector machines

1. Introduction

In the past 50 years, quantitative structure–activity relationship (QSAR) has become a powerful tool for drug design and discovery. The underlying principle in QSAR modeling is the assumption that molecular structure information is sufficient to model and predict biological or pharmacological activity. Hence, in QSAR studies, different molecular representations have been used to describe the information encoded in molecular structures so as to predict the quantitative relationships between biological activity (response-variable) and structural information (predictors) [1–5].

The performance of QSAR models for the accurate characterization of biological molecular properties largely depends on the relevance of the selected molecular representation. Such representations can be divided into two broad categories of methods, namely, vector space and metric space representations [6]. A vector space or linear space representation occurs when the set of modeling instances is represented as a vector, with its characteristics measured relative to some reference frame and thus having a notion of magnitude and direction from the origin. In most

QSAR modeling studies, vector space is the most common representation used, where each chemical structure is translated using a set of molecular descriptors. This is generally referred to as the "chemical feature space", which represents different structural characteristics/properties [5,7,8]. Nevertheless, vector space-based QSAR modeling has two major modeling issues. The first is the determination of the set of features capable of structural representation and, the second is the identification of the subset of features that, more significantly, are able to predict the desired property [9–13]. Metric space representation, on the other hand, is built on the principle of measured distances between a set of instances that we want to model. As sometimes it is difficult to identify specific features of a real world entity such as a molecule, oftentimes it is easier to quantify its distance or similarity to other instances. A typical case for using metric space representations is in protein functional annotation; while it is quite hard to define a set of features that characterize a protein, the similarity between proteins (whether structural or sequence-based) is commonly used to assign its function, as it is known that above a given similarity threshold, proteins maintain their function [14,15]. In in silico screening, the similarity principle leads to the simplest database screening methods. If a seed molecule has been experimentally determined as active, the first approach to find other actives is to identify similar molecules, as the probability of finding other actives increases with proximity to the base molecule [16,17]. QSAR metric space modeling is also hampered by two different issues. In the first place, we need to determine how to measure similarity between molecules—for which there are currently several and conflicting approaches—and secondly, it is necessary to compute the distance of each molecule to all the molecules in the training sets, which may entail difficult computational problems. Distance matrices, as they are quadratic to the number of instances of the data set, add difficulties to the modeling effort and do not scale well, even with the increased computational power available today. Any vector space is a metric space, as it is possible to compute the distances between instances using any common distance metric such as the Euclidean distance. There are some data sets for which no vector representation is known (e.g., proteins); however, it is possible to compute their distance. Thus, all vector spaces are metric spaces, but the reverse is not true (Figure 1).

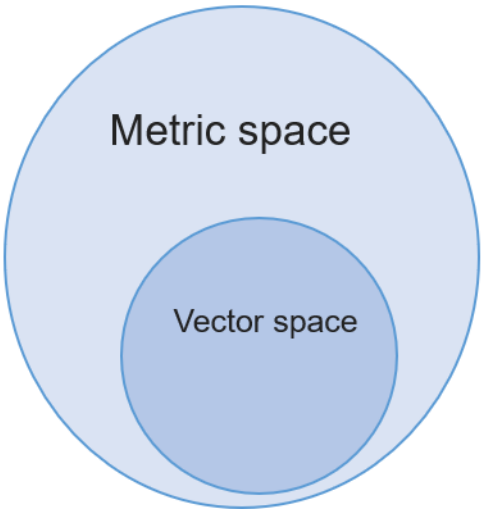

Figure 1. Vector space vs. metric space.

1.1. Molecular Similarity and Metric Space Representation

Molecular similarity largely depends upon an appropriate combination of two basic components including (a) a molecular structural representation to find the overlapping or similar features and (b) a similarity function/coefficient to quantify the similarity [18–26]. By far, the most commonly used structural representation for comparing molecules is the use of two-dimensional (2D) molecular fingerprints. Fingerprints are a sort of binary fragment descriptor, where each bit represents the hashing product of the possible chemical fragments of a molecule. There are currently several widely

used fingerprints that differ in the form that a molecule is decomposed, the size of the representation, and the hashing algorithm [27]. Some other descriptor-independent methods are also available for molecular similarity comparisons, including molecular graph matching approaches [28–31]. To quantify molecular similarity, the most common method used is the Tanimoto (Jaccard) similarity coefficient [32,33]; however, there are many other similarity/distance methods [20,25,26,33,34]. The one-complement D of the Tanimoto/Jaccard coefficient, where $D = 1 - J$, has been proven to be a real metric, satisfying all the known properties of distance measures [35]. In comparison to vector space-based methods, there is limited research reported in the literature exploring the quantitative relationship between computed molecular similarity and activity in QSAR/QSPR modeling [7,16,19,36–45].

1.2. Metric Spaces vs. Vector Spaces

With all of the aforementioned concerns, the main question that we want to address in this study is whether a metric space or a vector space modeling approach outperforms the other in QSAR regression problems. Therefore, in this work, we have carried out a comparative analysis of molecular structural representation using some of the most commonly used vector and metric space-based methodologies and compare the results. Overall, we seek to answer the following four questions:

- Is metric space representation as good as the most common vector space-based approaches?
- Which similarity representation carries the maximum chemical/structural information content to establish the best relationship between structural similarity and activity?
- How effective is the reduction of dimensionality of the feature space with principal components by the replacement of explicit descriptors/fingerprints in QSAR modeling?
- Is there any one molecular structure representation method that is generally better than the others?

To accomplish these goals, the following work was performed: Five distinct data sets with distinct modelability characteristics were selected and curated from ChEMBL23. Several modeling efforts were then systematically applied to all selected data sets, namely (i) a typical vector space representation of molecules was performed by using an extensive set of chemical descriptors then used for model fitting in a QSAR optimization framework that includes automated data processing, descriptors/fingerprints computation, and feature selection; (ii) similarity matrices were computed for all data sets using a variety of methods (five fingerprint-based and one graph-based), and these similarity matrices were then used for modeling by using their principal components as model components; (iii) the fingerprint-based representations, as they actually also represent molecular features, were further used in a vector-based model, using the same linear dimensionality reduction method. For all three different modeling choices, the number of features (or principal components) used in each model was selected by using five-fold cross-validation, and each final model was assessed against an independent validation set randomly selected from the initial data set and which was never used in any step of the model-fitting phase.

2. Methodology

2.1. Overview of the Methodology

We collected and curated the molecular data for each biological target from ChEMBL23 [46], then all molecules of each data set were represented using different fingerprint models and molecular descriptors and separated into different modeling problems. To perform all of the analyses, each data set was initially randomly split into training and independent validation sets (IVSs), the former used for training and model selection, and the latter for the final evaluation of the model. A state-of-the-art QSAR modeling approach [47] was used to build a predictive model using an optimized feature selection procedure. The other models investigated with the same data sets required first the computation of five different fingerprint sets. These were used for additional vector space modeling

and for the computation of similarity matrices between all molecules of each data set. Additionally, one graph-based structural similarity (NAMS) approach was used to make one further similarity matrix for metric-space modeling. Principal component analysis (PCA). was applied to both the similarity matrices and the bare fingerprints so as to create and evaluate models by iteratively increasing the number of principal components. The predictive performance of all data representations was assessed using the IVSs, which were never used during feature/PC selection (Figure 2). The details of each step of the followed methodology are covered in the following sections.

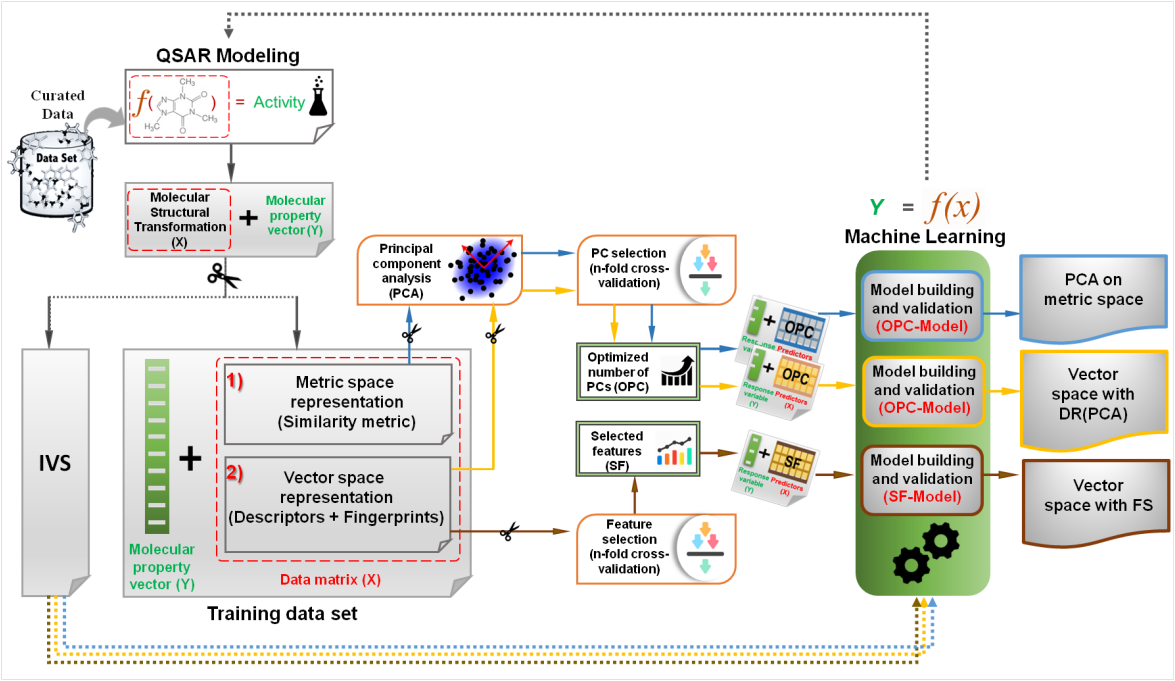

Figure 2. Quantitative structure–activity relationship (QSAR) modeling methods.

2.2. Vector Space Representation

In a vector space, each molecule is represented by using a feature vector that contains several molecular properties (descriptors) or structural features represented using a binary array of fixed size (fingerprints) [27,48].

2.2.1. Descriptor-Based Representations

Molecular descriptors aim to selectively describe the information encoded in the structure [48]. Some molecular descriptors are derived with mathematical formulae obtained from chemical graph theory, information theory, and quantum mechanics, among other methods, that directly illustrate some relevant features of the molecules [48,49]. Molecular descriptors can be divided into four broader categories: constitutional (1D), topological (2D), geometrical (3D), and physico-chemical properties-based (4D) descriptors [48,50]. 2D descriptors are the most commonly used types of descriptors.

2.2.2. Fingerprint-Based Representations

Another well-known molecular representation is molecular fingerprints, which are fixed-length bit-strings where each bit encodes a fragment or characteristic of a given molecule [27]. Molecular fingerprints are often very different in length and complexity, ranging from 2D/simple representations of relevant structural features to 3D/complicated pharmacophore arrangements. Thus, many types of fingerprints have been generated with different settings (generation method, length,

size of patterns, and number of bits activated by each pattern, etc.) and are further deployed as descriptors for predictive modeling to estimate biological activities [12,27,51–54].

In principle, 3D representation should have higher information content than 2D, but surprisingly, higher complexity is often more error-prone and less robust in performance [26,55–58]. 2D fingerprints can encode different structural information, for example, molecular fragments and structural patterns, topological pathways through compounds, or topological atom environments either as bit strings or feature sets. Numerous software packages have been developed to generate several types of fingerprint for drug discovery applications [54]. Moreover, the basic principle of fingerprints generating algorithms and their comparative performance in a variety of QSAR problems has been extensively studied [8,26,54,59]. The preferred molecular fingerprints can be grouped into the following three classes:

- Topological/path-based fingerprints (e.g., Daylight-like RDkit [27,60] and Atom Pair [61]) capture the paths between atom types by describing their different combinations and always assign the same bit's position to the same substructures within the compared molecules, which sometimes results in bit collisions but is also useful for clustering compounds.
- Circular fingerprints (e.g., ECFP [62]) record circular atom environments that grow radially from the central atom connections. In topological and circular fingerprints, an individual bit has no definite meaning.
- Structural keys fingerprints (e.g., MACCS [63], PubChem [64]), where each specific bit position represents the presence (1) or absence (0) of predefined functional groups, substructure motifs, or fragments.

2D fingerprints can easily be calculated by specialized, open-source, and readily available software packages (e.g., OpenBabel [65] or RDkit [60]). 2D fingerprint-based similarity analysis is the most widely used methodology in ligand-based virtual screening, clustering, and diversity analysis [24,26,59,66,67].

2.3. Metric Space Representation

A molecule in metric space is defined only as its relation (distance or similarity) to all other molecules in the data set. Technically, a metric space is computed using distances between all the elements of a data set, creating a distance matrix which can then be used in a variety of modeling techniques such as hierarchical agglomerative clustering or k-nearest neighbours models [68,69]. There is a variety of ways to transform similarities into distances [16,54]; however, as all the methodologies for comparing molecules produce similarity matrices, it was deemed unnecessary to transform the similarities into distances and we instead use similarity matrices directly for modeling, as this extra transformation would introduce one further step in the data preparation procedure with no clear advantage.

In descriptor-independent methods, graph matching approaches have been used. In these methods, graph theory is used to represent molecules as labeled graphs whose vertices correspond to the atoms and whose edges correspond to the covalent bonds. Several techniques, each with some advantages and limitations, are available to compare labeled graphs [29]. In the descriptor-independent methods, many advancements have been introduced to improve the sensitivity of graph matching methodology and obtain consistent and reliable molecular similarity results. One of these methods is non-contiguous atom matching structural similarity (NAMS), which has shown modeling advantages over other structural methods [28], although the computational cost of its application can be high.

2.3.1. Fingerprint-Based Similarity

Many types of 2D and 3D molecular fingerprints have been generated to code chemical structures/properties into bit-string representations [20,70,71]. Molecular fingerprint representation allows for an easy comparison of molecules by identifying and quantifying the amount of overlapping

elements between them. The applications of molecular fingerprints has been broadly reviewed and used in the literature [22,54,70,72]. There is a large variety of similarity and distance functions that have been introduced and return a molecular similarity score [54,59]. In cheminformatics, the prevalent approach is the use of the Tanimoto coefficient (T_c) over molecular fingerprints [26,33]. In the case of 2D fingerprint comparison, for binary vectors of fingerprints representing two molecules A and B, T_c is defined as

$$Tc(A, B) = \frac{A \cap B}{A \cup B} = \frac{c}{a + b - c}. \quad (1)$$

In Equation (1), a corresponds to the number of bits set to 1 in molecule A, b is the number of bits set to 1 in molecule B, while c is the number of common set bits in both molecules. $1 - T_c$ is an actual distance measure, encompassing all four property distance measures referred to above.

2.3.2. NAMS-Based Similarity

NAMS is a graph matching algorithm that uses a new atom alignment method to quantify the structural similarity between compared molecules [28]. NAMS breaks complex molecular structures into simpler parts to reduce molecules to atom–bond–atom structures and calculates a global structural similarity score from the best optimal alignment between the atoms of compared molecules. This algorithm has shown a higher discriminant power for biological activity than other structural or graph matching approaches. One of the reasons is that the applied atom matching methodology is able to consider important characteristics of atoms and bonds such as chirality and double bond stereo-isomerism that are oftentimes ignored in other approaches.

Given the structural representation of any two molecules, NAMS is able to compute its similarity score. NAMS can be fine-tuned with several parameters that allow users fo increase the importance of any specific molecular characteristics (atom or bond similarities and atomic characteristics like atom stereo-isomerism or double bond cis-trans isomerisms). Changing the parameters will change the resulting molecular similarities, but the overall results of comparing large and diverse data sets are not very much changed. For the current work, only the parameters were used.

2.4. Model Building

In QSAR modeling, the most well-known machine learning approaches include neural networks (ANN), support vector machines (SVM), decision trees, random forests (RF), and k-nearest neighbours [73,74]. In the last few years, RF [75] and SVM [76], two non-linear supervised learning methods, have become the most prevalent algorithms in QSAR studies [75,77–82]. One of the biggest advantages of SVM is its ability to deal with high dimensional and duplicated data with a lower risk of model overfitting [79–82], while, on the other hand, RF are considered specially robust in complex situations of high dimensional QSAR/QSPR data sets [75,77,78]. Hence, RF and SVM are the basic algorithms used in the learning phase of the current work.

As stated, some of the most prevalent issues in QSAR modeling approaches are variable redundancy or collinearity, with complex correlation patterns between descriptors or the presence of irrelevant features in the data set, which may reduce the quality of the produced models. These are consequences of the high dimensionality of such problems. Such issues are aggravated by the fact that in QSAR studies, there are oftentimes many more predictors than the number of actual instances to fit [9–13], which will make it more difficult to find adequately fitting models. Several approaches have been followed in the literature to solve the descriptor selection problem in QSAR modeling [73,77,83–85]. These approaches can be roughly divided into two different categories: feature reduction and feature selection. In feature reduction, the main purpose is to algebraically combine sets of features into statistically independent new components. There are several methods that purport to accomplish these goals, among which is principal component analysis (PCA) and singular value decomposition or kernel PCA [86]. PCA is by far the most commonly used method in feature reduction, while kernel

based PCA is beginning to get some traction in the literature [87]. Feature selection, on the other hand, is a more complex problem, and in essence can be summarized as finding and selecting the smallest set of features that are capable of producing the best model. Methods to address this problem include the identification of linear correlations between all variables, bootstrapping methods capable of deciding which variables have the highest impact on model quality, or the use of optimization meta-heuristics like genetic algorithms [73,77,83–85].

In this work, we used two of the most common methods for feature reduction. PCA was used with the metric space data produced from the similarity matrices and fingerprint data, while random forests were used to identify the most relevant features capable of producing the highest-scoring models.

2.4.1. Feature Reduction with PCA

Principal component analysis (PCA) is a linear reduction method used to calculate the most meaningful basis on which to re-express high dimensional data into a reduced space. However, PCA is a useful tool in QSAR modeling to deal with the problem of high data dimensionality and collinearity [4,68]. In typical QSAR studies, PCA is used to analyze the original data matrix in which molecules are represented by several types of predictor variables (molecular descriptors/fingerprints). PCA performs dimensionality reduction by transforming original descriptors' space into linear orthogonal combinations of original variables named principal components (PCs). The generated PCs are uncorrelated and always ranked according to the decreasing data variance of the original variables [68]. As the first components contain the highest amount of data variance, models can be fit to data by gradually incrementing the components in the model. A first model will use only the first component, a second model will use the first two components, and so on, and which of these models with reduced dimensions is capable of producing the least amount of error in k-fold cross-validation is evaluated. Since each PC is an independent source of the original data variance, PCs have been used as a model input mainly when high data dimensionality is a big issue, and most models are sensitive to the number of variables used [68]. Several studies in the literature apply PCA for dimensionality reduction in QSPR/QSAR problems [4,88–90].

In this study, we performed PCA in both vector space representations (descriptor and fingerprint data matrices) and metric space representations (fingerprint-based similarity data matrices and NAMS-based similarity data metric). The generated PCs were used to build QSAR models with dimensionality reduction (DR). We compared the predictive performance of QSAR models generated by the reduced dimensionality of metric space with typical PCA-based QSAR models where vector space is reduced by PCA.

2.4.2. Feature Selection with Random Forests

A random forest (RF) is an ensemble supervised nonlinear machine learning algorithm for classification or regression [75]. This algorithm generates a set of weakly independent decision trees that are built using randomly selected subsets of the data. Each generated tree is produced by randomly selecting a set of predictors from the full set and by sampling with replacement instances from the same data pool. This will create a set of randomly generated trees (a forest), each one created from different data and variable partitions. The RF algorithm then uses a consensus voting procedure to combine the predictions from all randomly generated weak models and make more robust predictions. One of the consequences of this bootstrap procedure is that it is possible to assess the power that each variable has in the final predictions. The trees that include such variables will typically have higher prediction power, and as such, it is possible to rank each variable in terms of its overall importance to the model quality. Many studies showed that RFs' voting procedure can be used for feature selection by ranking and selecting each variable according to its importance in RF models [77,85,91]. In this ensemble method, each variable's importance score is calculated using several variable importance (VI) measures. In regression problems, an increase in the mean squared error of a tree is one of the widely

used VI measures, which explains how much prediction error increases with the random permutation of any given variable while keeping all others unchanged in a node of a tree [75,85,91,92].

In this work, we followed the random forest (RF)-based feature selection method [77] to rank features in a high dimensional vector space according to their importance score. These are then later used in the feedforward feature selection procedure (Figure 2).

2.4.3. Support Vector Machine

An SVM [76] is a supervised machine learning algorithm that has been widely used for classification and regression-based data analysis in many fields, including QSAR studies [77,79–82]. For a given set of data instances, a discriminative SVM algorithm focuses on the identification of support vectors (data instances) to draw a decision hyperplane in a high dimensional space that best separates data instances with maximum margins. SVM uses different kernel functions for data transformation in a new hyperplane; these can be linear, radial basis functions, sigmoid, or polynomial, which are generally considered good choices for a majority of problems. The discovery of support vectors greatly depends on the selected kernel function. In contrast to other methodologies where there is a learning phase that heuristically searches thorough the multidimensional feature space, in SVM learning this search procedure is a mathematical optimization procedure, and it is guaranteed that an optimal solution can be found in polynomial time. This also implies that, as no random component is involved, the same solution model will be produced for each model. In this work, we used SVM in the process of feedforward feature selection where PCs from vector/metric reduced dimensionality space and RF importance score-based ranked variables from features/vector space were stepwise subjected to the SVM, and final QSAR models were developed with an optimized set of selected dimensions (Figure 2). In the current work, the radial basis function was selected for all problems.

2.4.4. Model Evaluation and External Validation

N-fold cross-validation or model internal validation is the simplest approach, where the training data set is randomly divided into a number (N) of folds (parts), and each part is used as an external set for the validation of the predictive model, which was fitted by using the remaining compounds in the other $N - 1$ partitions. Cross-validation is essential to optimize modeling parameters and variable selection, and to verify the internal predictive power and robustness of the QSAR model [89]. In our analysis, we performed N-fold cross-validation to find an optimized number of most relevant variables (variable/PCs selection). For this purpose, a feedforward approach was used to generate estimation models by sequentially adding the RF importance score-based ranked variables (more relevant to least significant) and PCs extracted from vector and metric spaces as an input in the SVM algorithm. The internal predictive performance of each model was assessed by computing the percentage of variance explained (PVE) and root mean squared error (RMSE) of each predictive model in cross-validation [93]. As the cross-validation may result in a different number of best-performing variables for different folds, an average of the PVE score was recorded across all folds each time. Finally, the set of dimensions that led to the smallest average predictive error score in all folds was considered as the selected number of descriptors/fingerprints/PCs. After performing all of this feature optimization, the whole training data set was reused to develop a model with the selected features to perform a blind external prediction using the independent validation set.

3. Data

We tested the proposed QSAR modeling methodology on five data sets for common human biological targets, retrieved from ChEMBL23 [46]. These were selected independently of any previous hypothesis (Table 1). We used an automated QSAR modeling workflow [47] to collect and curate data for each selected target. The bioactivity data of the selected targets was retrieved using the UniProt accession number (Table 1).

Table 1. Data set description.

Uniprot ID.	Gene Name	Target Protein Name	Associated Bioactivities (Y)	Total Number of Observations (N-Processed)
P35367	HRH1	Histamine H1 receptor	Ki	1222
Q99720	SIGMAR1	Sigma non-opioid intracellular receptor 1	Ki	226
Q12809	HERG	Potassium voltage-gated channel subfamily H member 2	Ki	1481
P35462	DRD3	D(3) dopamine receptor	Ki	2902
P28223	HTR2A	5-hydroxytryptamine receptor 2A	Ki	2088

Moreover, missing data, salt groups, and mixtures (e.g., in unconnected molecules, smaller fragments were excluded) were removed. In duplicated data, if more than one record was present for the same compound, the one kept would be its most recent measurement, according to the publication year. All data sets feature *Ki* as the bioactivity measure. However, the logarithm of *Ki* is more typically used for modeling and makes more biological sense. Also, to encompass several problems of the more extreme values, it was decided to clamp the values between an interval so that very weak or possibly inactive molecules receive the same low score, while it is oftentimes unnecessary to discriminate results with $Ki \leq 1$ nM, as these are very active molecules. Thus, the following expression (Equation (2)) was used for all data sets to transform *Ki* into *spKi* (scaled and clamped *pKi*):

$$spKi = \begin{cases} 0, & \text{if } Ki \geq 10{,}000 \text{ nM}, \\ \frac{4 - \log_{10}(Ki)}{4}, & \text{if } 1 \text{ nM} < Ki < 10{,}000 \text{ nM}, \\ 1, & \text{if } Ki \leq 1 \text{ nM} \end{cases} \quad (2)$$

spKi values are thus clamped between 0 and 1, the most active compounds having values closer or equal to 1, and the lesser active or inactives will have values of zero. This clamping assumes that *Ki* values below 1 nM are considered extremely active compounds, while molecules with *Ki* values above 10,000 nM are considered very weak or inactive.

Data Preparation for Vector and Metric Space Representations

For each data set, molecules were represented in metric and in vector spaces by using three different approaches: (a) common vector space methods using molecular descriptors or fingerprints, named vector space with FS (feature selection); (b) principal components over the similarity matrices, categorized as PCA on metric space; and (c) principal components over molecular descriptors and fingerprints placed in vector space, or DR (PCA) (Figure 2).

For vector space representation, we used 1348 descriptors (2D and 3D) calculated for each selected data set with the RDKit [60] toolkit (Table S1). Separate modeling efforts were performed by testing five different types of fingerprints separately, including ECFP6 (circular), PubChem (substructure keys) computed using the CDK [94] toolkit, MACCS (substructure keys), RDkit (path-based), and Atom Pair (path-based) generated using RDKit [60]. The data preparation for principal component over metric space representation involved the computation of the similarity matrices between all elements of the training set and computing the distances of the IVS to those of the training set. Using the Tanimoto index, similarity matrices were obtained for each of the five fingerprints by adding the NAMS graph-based molecular matching algorithm. Models generated using dimensionality reduction of metric and vector spaces were named "optimized number of PC models" (OPC-models), as the procedure emphasizes selecting the best number of PCs, capable of producing more reliable models. Predictive models built using vector space with FS were named SF-models (model having the selected number of features) (Figure 2).

Thus, a total of eighteen different molecular representations were used in this study and served as input data to a machine learning algorithm for the generation of ninety regression models for five selected QSAR problems.

4. Results

4.1. Implementation of Analysis

All molecular descriptors and fingerprints used in this study were calculated using CDK [94] and RDKit [60] built-in nodes of the open-source data-mining framework KNIME (version 3.2) [95]. All analyses were performed using R (version 3.4.4) [96] on a desktop workstation powered by a 6th-generation Core i7 Processor (3.41 GHz) with 16 GB RAM. Package e1071 [97] was used for the SVM algorithm and an R library, randomForest [98], for RF. Both SVM and RF algorithms were implemented with the default parameters. The R package factoextra was used for dimension reduction using PCA [99]. It is noteworthy that in the PCA-based QSAR modeling, orthogonal projections/PCs for test sets in N-fold and IVS were calculated by using R's PCA `predict()` function .

4.2. Results of Generated Models

OPC-models and SF-models were fitted with the training data sets of all selected targets. For all data sets, the training data was used to evaluate and select the model that was able to produce the smallest RMSE or PVE (ratio of the variance explained). Typically, this involved selecting models with a reduced number of features or PCs (Table S2 and Figure S1). The final models after feature selection were validated using the same IVS for each problem set (Table S3).

The first aspect that stands out from these results is that the most relevant factor for explaining model quality is the nature of the data itself. The predictive performance of QSAR models highly depends upon different characteristics of the data set (e.g., size, chemical diversity, and presence of activity cliffs) [100–106]. As an example, the *HERG* data set can easily be seen as a difficult problem, independently of the approach followed to model it (Figure 3). On the other hand, the human Histidine Receptor 1 (*HRH1*) generally appears as more easily modelable, while the remaining three problems (*SIGMAR1*, *DRD3*, and *HTR2A*) show intermediate modelability characteristics. Secondly, with some relevant cases noted below, no single method uniformly performs better than the others, and each method's performance seems to be heavily dependent upon the data set characteristics.

To have a more encompassing view of the produced results, we performed a Friedman ranked test [107]; this is a non-parametric test used to assess different treatments applied to different test situations, as is the current case. In the present situation, a modeling approach is considered a treatment, which is evaluated by its results for the different data sets. Each model is then ranked according to its performance, where the best models have a lower rank and vice versa. The Friedman test is then able to evaluate each performance according to its rank in all data sets, thus effectively providing a performance value for each modeling approach. Another advantage of the Friedman test is that it allows for a post hoc analysis that is able to better qualify the differences verified between treatments, for instance, by grouping similar models with similar performance values. For each modeling data set, the rank in PVE of each modeling approach was calculated in R's agricolae package (Figure 4) [108]. The test results showed that there were significant differences between treatments, with a Chi-squared test of 38.44 with 17 degrees of freedom giving a p-value of 2.2×10^{-3}, which strongly suggests that there are statistically significant differences between the different modeling approaches.

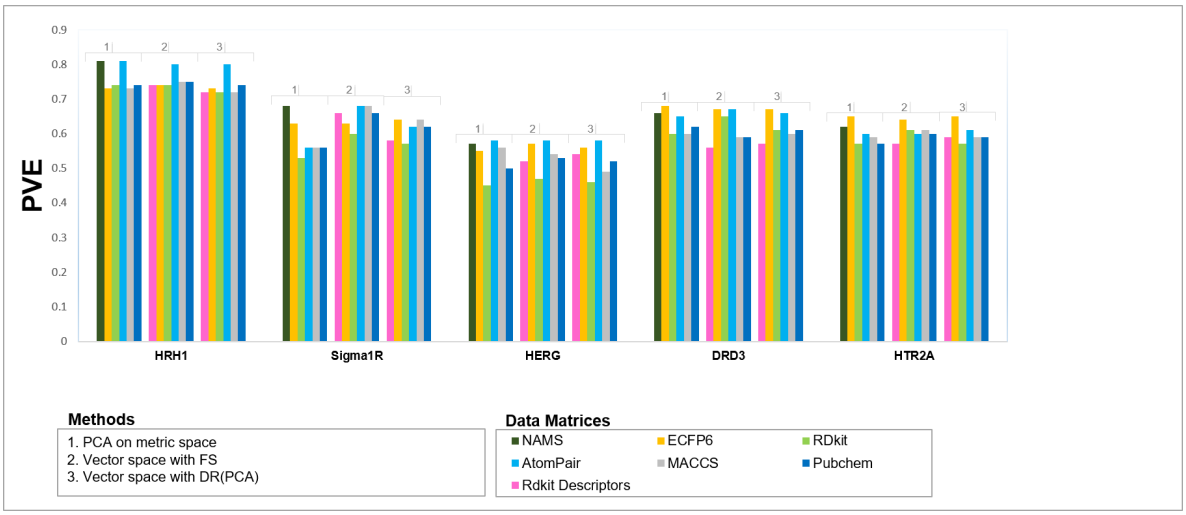

Figure 3. Comparisons of QSAR models' predictive performance using independent validation sets (IVSs). PVE: percentage of variance explained by the model.

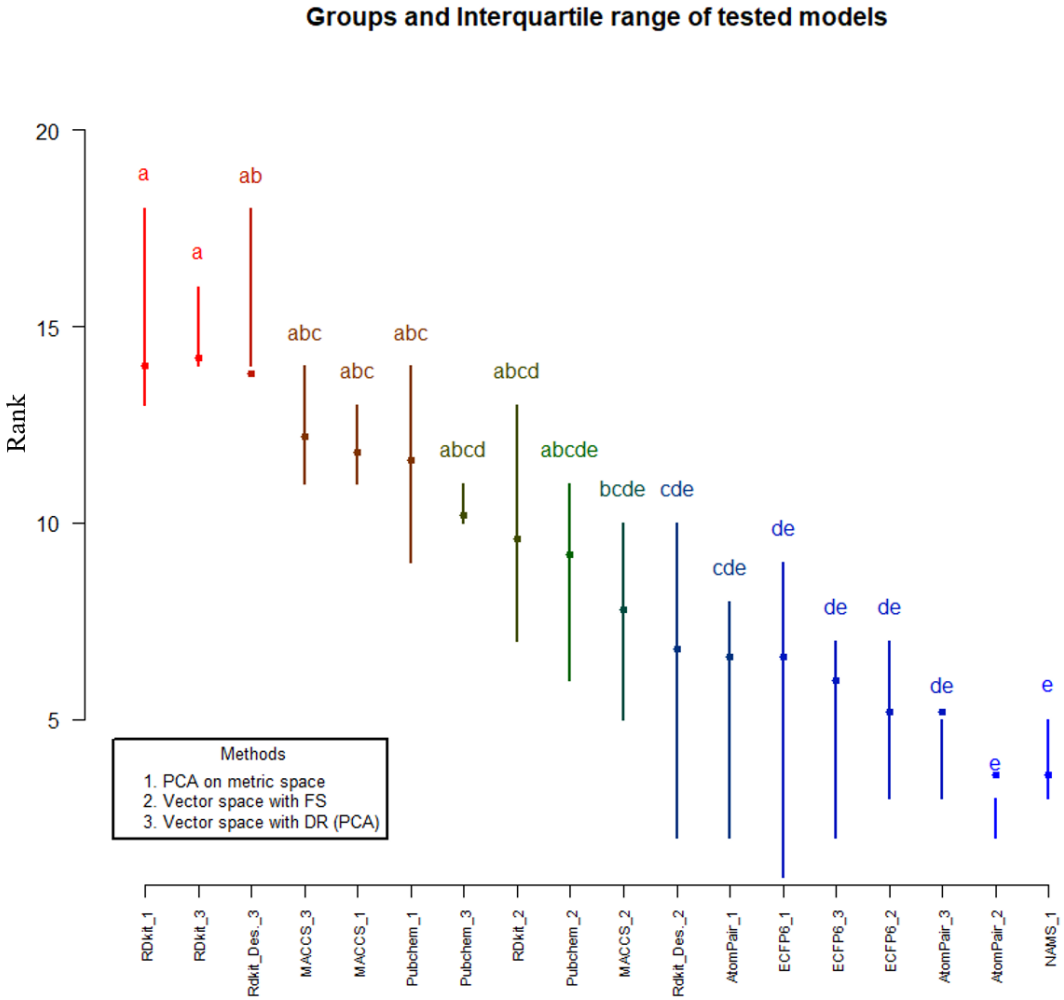

Figure 4. Friedman test results and interquartile ranges of tested models.

The post hoc analysis of the Friedman test allows groupings of statistically indistinct treatments under the same grouping [107]. A treatment can belong to several groups. Figure 4 indicates to

which groups each model belongs. The significance level used was 0.05, meaning that the model groupings are correct with at least 95% confidence. It can clearly be seen only the models that belong to grouping *e*—the one with model rankings consistently lower (thus indicating higher quality modeling approaches)—are NAMS metric space PCA and Atom Pair fingerprints with classical feature selection. Moreover, it can be observed that the use of RDkit fingerprints and molecular descriptors, both with metric space representation and PCA dimensionality reduction, consistently appear in the highest positions (worst models).

We further dissected these individualized results according to the four major questions that were the main objectives of our analysis. These questions are addressed one by one in the following sections.

4.2.1. Is Metric Space Representation as Good as the Most Common Vector Space-Based Approaches?

To answer this question, the results of all three different approaches (simple feature selection and PCA dimensionality reduction in both vector spaces and metric spaces) were analyzed. A comparison of OPC-models generated using PCA on metric and vector spaces and SF-models built using vector spaces with FS showed that the predictive performance of each QSAR model was influenced by the selected type of molecular structural representation (Figure 5), which was expected and consistent with the literature [50,103,109]. We performed a similar analysis using the Friedman test over the ranks of the median values of each data modeling approach from the explained variance (PVE) of the fitted models using each respective IVS (Figure 5).

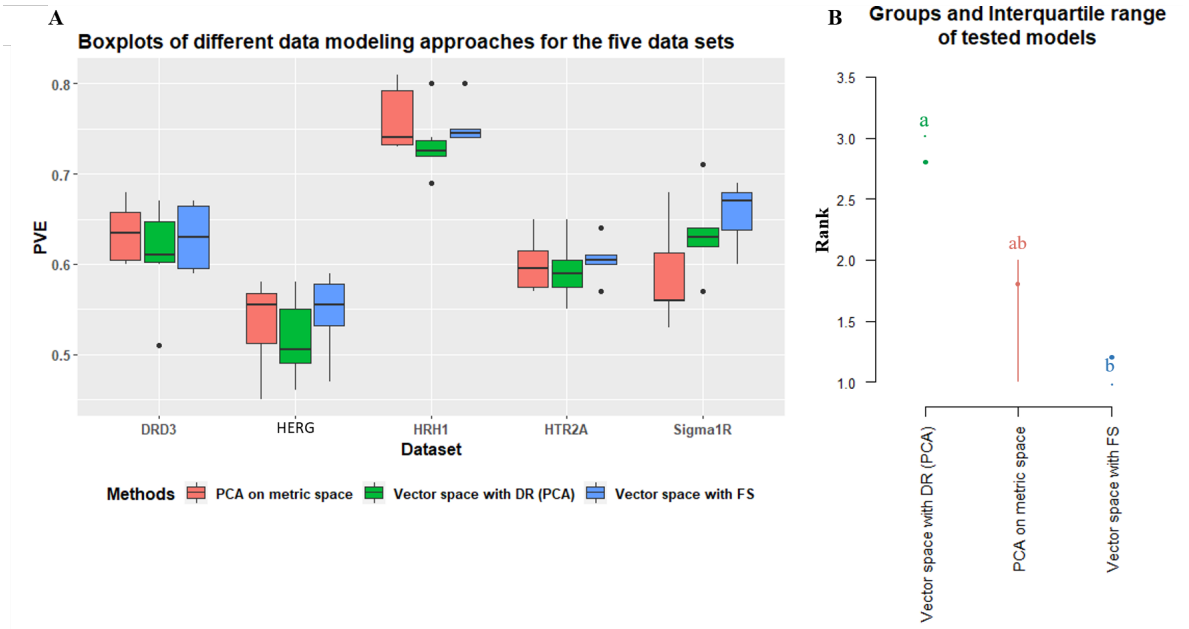

Figure 5. (**A**) Boxplots of the three modeling approaches grouped by the different data sets; (**B**) groups and interquartile ranges of the medians of tested models from the Friedman test post hoc analysis.

Feature selection over vector spaces has proven to be the most globally reliably modeling approach and appears to be significantly better relative to the use of PCA on the same data. Metric space PCA appears as somewhere in between, closer to the feature selection approach. The Friedman test for this data yielded a Chi-squared value of 6.0, which corresponded to a *p*-value of 0.049, just below the 0.05 threshold. With such results, it is fair to conclude that the usage of metric space data may compromise the quality of the models produced when comparing results to traditional vector space feature selection models, yet it clearly outperforms vector space PCA-based approaches. It is nonetheless striking that the highest-ranking method in the overall assessment is NAMS, a metric space-based approach, which may allow us to suggest that the other methods of calculating molecular

similarities may be responsible for this decreased performance and may not be as adequate to compute molecular distances.

4.2.2. Which Similarity Representation Carries the Maximum Chemical/Structural Information Content to Establish the Best Relationship between Local Similarities and Activity?

To analyze which similarity representation contributed more significantly to reliable predictive modeling, the overall performance of OPC-models generated using six similarity data matrices (NAMS, ECFP6, RDkit, Atom Pair, MACCS, and PubChem-based similarities) was evaluated again using the Friedman test (Figure 6). The ranking of each metric space-based approach was assessed for each data set and the overall quality of each model quantified through the use of the Friedman test and respective post hoc analyses. For the present case, NAMS clearly emerges as the best approach, followed closely by Atom Pair and ECFP6 fingerprints, the former appearing in the same group as NAMS. The Chi-squared test for the metric space-based approaches ranked comparison was 15.2 with 5 degrees of freedom, which corresponds to a p-value of 9.5×10^{-3}. Thus, test results again suggest that NAMS molecular similarity is able to more reliably capture important structural information, which eventually generates a better quantitative relationship between local similarities and compound activity.

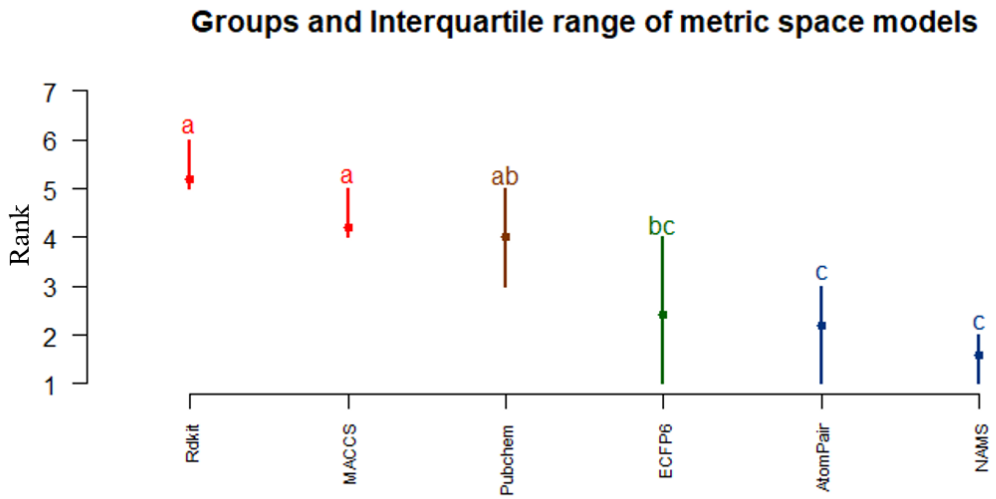

Figure 6. Overall performance of similarity representation using PCA on metric space-based QSAR modeling approach.

4.2.3. How Effective Is Using a Reduced Dimensionality of the Metric/Vector Space with Principal Components, Replacing Explicit Descriptors/Fingerprints, in QSAR Modeling?

This question can actually be answered by observing the previous results. It seems clear that when directly comparing PCA to direct feature selection (Figure 5), the latter produces markedly better results, which strongly suggests that the dimensionality reduction achieved with PCA is a poor proxy for a better structured search for the most relevant descriptors in a modeling problem. Nonetheless, using PCs from the similarity matrix allows us to capture the same information available from vector space modeling. These results also highlight the capability of fingerprints to produce high quality models without the need for other chemical descriptors. Furthermore, the fingerprint-generating method appears critical for producing the most reliable models. As is clear from the above results, Atom Pair and ECFP6 fingerprints appear as the best fingerprint-based similarity approaches, while the RDkit and PubChem fingerprints consistently lag behind all other models.

4.2.4. Is There Any Solution That Is Globally Better on a Variety of Difficult Problems?

From the above results, it is clear that there is no one single best approach for dealing with complex QSAR problems. Although metric space-based NAMS and Atom Pair come out in first place most of the time, they are not consistent for all data sets. For instance, Atom Pair fingerprint representation performs poorly for the HTR2A model, while NAMS does not appear on top for the DRD3 data set. Similarly, as mentioned above, there does not appear to be any intrinsic advantage in changing from a fingerprint vector space-based approach to similarity-based metric space modeling. The most consistent result was that the use of PCA with descriptor data was generally a poor modeling approach. PCA can nonetheless be used with distance matrices to capture reliable information for modeling.

5. Discussion

Many studies have demonstrated that the selection of molecular structural representation has a larger impact on the predictability of QSAR models than the choice of model optimization methods [8,26,54,59,67,109,110]. Our results confirm these findings further, suggesting that reduced metric space representation using NAMS-based similarity and Atom Pair fingerprints with feature selection are the methods that more consistently address a variety of modeling problems.

Nonetheless, one further concern over such studies is how much novel information is actually being discovered from the models, as it is a known fact that similar molecules tend to have similar biological properties. Therefore, a distinct possibility is that the use of similarity matrices for inference may be result in reliable predictions only when molecules very similar to the training data set are present. Thus, one further test for these modeling approaches is to understand how reliable these methods are for making models where all very similar molecules have been removed and no molecule, either in the training set or the IVS, has a high similarity to any other. This would allow the evaluation of the capability of each approach to make inferences when very diverse compounds are fed into the model. Therefore, to check the robustness of the tested methodologies, the five data sets were manipulated by converting them into harder problems with only structurally diverse molecules, making certain that no molecules within a given similarity threshold are present in each data set. Accordingly, five new data sets were created based on the initial ones but where no molecule was present if it was similar, within a given threshold, to others already present. As different similarity methods produce different scores for the same molecules, the thresholds were adjusted for each similarity method to make sure that the model would be trained with a similar number of instances (Table 2). This complementary analysis obviously relates only to metric space modeling, thus the following results will only focus on this modeling approach.

Table 2. Data size before and after removing nearest neighbors. Thr—similarity threshold; N—new data set size.

Target Protein Name	Data Size without Removing Nearest Neighbors	NAMS		ECFP6		RDkit		Atom Pair		MACCS		Pubchem	
		Thr	N	Thr	N	Thr	N	Thr	N	Thr	N	Thr	N
Histamine H1 receptor (HRH1)	1222	0.80	379	0.55	378	0.80	371	0.67	376	0.84	379	0.87	391
Sigma non-opioid intracellular receptor 1 (Sigma1R)	226	0.87	312	0.61	310	0.89	305	0.75	309	0.92	311	0.94	321
Potassium voltage-gated channel subfamily H member 2 (HERG)	1481	0.80	397	0.54	394	0.82	392	0.69	395	0.83	395	0.86	403
D(3) dopamine receptor (DRD3)	2902	0.80	478	0.52	481	0.77	470	0.67	480	0.87	484	0.86	484
5-hydroxytryptamine receptor 2A (HTR2A)	2088	0.80	432	0.47	432	0.78	424	0.63	426	0.83	429	0.85	437

After removing the nearest neighbors, all data sets were again randomly split into training and independent validation sets, and the same data processing procedures were repeated for these new, more challenging data sets. Moreover, the same modeling principles were repeated by training the models with the training set while simultaneously selecting the best feature set and finally validating

the best model with the corresponding IVS. The overall performance of the same models using these new data sets was assessed. Because the number of instances present in all new problems is different, both RMSE and PVE were used to adequately assess each model's performance (Figure 7).

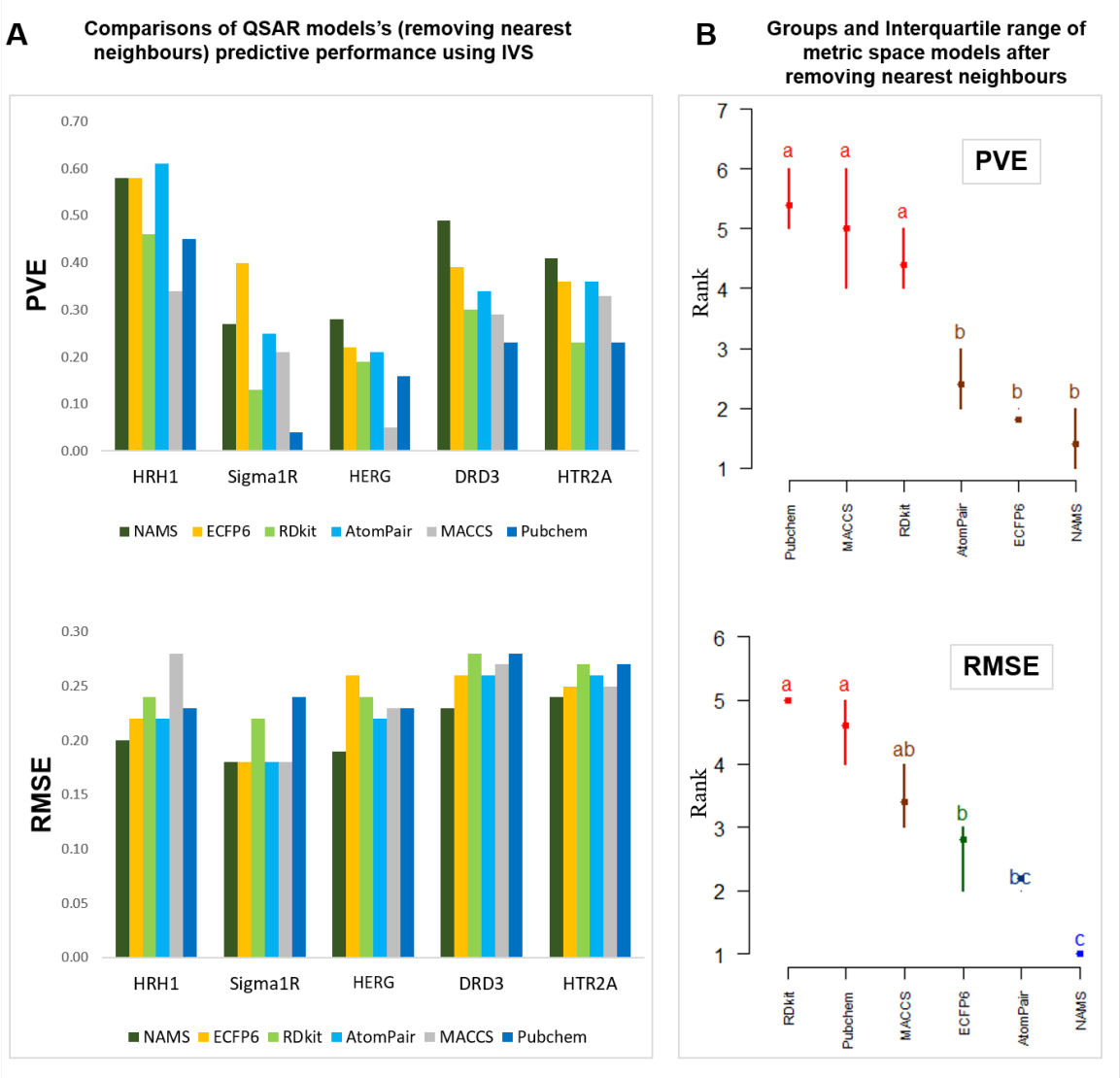

Figure 7. Overall performance of metric space representation after removing nearest neighbors in a PCA on metric space-based QSAR modeling approach.

As can be seen, with such hampered data the performance of QSAR models has naturally dropped, leading to a decrease in PVE ranging from 0.15 to 0.52 (Figure 7A). This finding is consistent with the literature [8] in that similar molecules present in models tend to inflate result statistics. It can also now promptly be seen that the differences between the different models are now amplified, and it is clearly easier to visually identify which approaches distinguish themselves from all others. Nonetheless, the overall model ranking was not significantly changed. Thus, NAMS similarity representation was, for these data sets, clearly the highest-performing model, achieving the lowest RMSE scores in all cases. Using the Friedman interquartile range graph (Figure 7B) performance scores for both Atom Pair and ECPF6 are dependent on the use of PVE or RMSE. All other fingerprint approaches were not up to the referred methods used in these more difficult challenges. The Friedman test for the PVE had a Chi-squared value of 21.1, with a p-value of 7.8×10^{-10}.

Computation Time

The execution time of QSAR models built from reduced dimensionality of metric space ranged between 60.61 and 48.88 min and for vector space, 52.53 to 15.34 min, whereas vector space with FS computational time ranged between 860 min (DRD3) and 17 min (Sigma1R). A comparative analysis of computational time showed that reduced dimensionality significantly reduced the complexity of the problem at hand and that computational time cost also decreased.

Computation time is an important issue when comparing different modeling approaches, especially when the use of metric space methods is being evaluated, as the use of a full similarity matrix is required for each data set. Furthermore, metric space modeling requires that one of the steps for inference is that the distance of each new molecule to all of the molecules in the training set is assessed. This is not typically a problem for academic studies but may put a large computational burden for actual screening efforts when several millions of molecules are being evaluated. This problem is aggravated in the case of the specific non-fingerprint approach we tested (NAMS). Although apparently able to produce a more accurate distance, which translates into better prediction models overall, it does so at a much higher computational cost. With current common hardware, the average computational cost to compute the similarity of two molecules is 12 ms, which for many problems may be too high for many problems. As an example, computing the similarity of one new molecule to a training set of 1000 molecules will require 12 s. Such computational costs (although the problem is trivially parallelizable) may involve unacceptable computational costs for very large data sets.

6. Conclusions

In this study, we compared different molecular representation approaches for input into QSAR machine learning methods. These approaches were divided into vector space- and metric space-based, with each molecule being represented as a vector of different characteristics in the former, and with a molecule being represented by its distance or similarity to others of known activity in the latter. We have tested five different fingerprint types (RDKit-FP2-based, MACCS, PubChem, Atom Pair, and Morgan's ECFP6) both as vectors of descriptors and, in metric space approaches, with Tanimoto scores computed for similarity. One exclusively vector space approach was also tested, where common chemical descriptors were computed and used in vector space modeling, as well as a pure metric space method with a molecular graph-based similarity (NAMS). We also tested whether it was more adequate to use dimensionality reduction methods (as with PCA) or a more computer-intensive feature selection procedure. These representation and dimensionality reduction methods were tested over five different data sets of different modelabilities and analyzed by the Friedman test for ranking models. Results showed that the choice of molecular representation to compute molecular similarity is more important than the modeling approach followed, thus certain methods produced consistently better results. ECFP6 and Atom Pair fingerprints were clearly the best approaches for modeling in vector spaces, surpassing all other methods. Classic molecular descriptors did not show any advantage for any of the data sets in this study. Regarding dimensionality reduction methods, the use of principal components appeared to be inferior to the use of random forest-based feature selection. The former method, albeit faster to process, generally produced results not on par with the latter.

In this study, metric space modeling by itself did not appear clearly superior to a vector space-based approach and, for the same representation, using fingerprints as descriptors tended to produce better results than using molecular distances from those same fingerprints. However, when using metric space representations, the differences between similarity methods become even more clear, with NAMS and Atom Pair fingerprints appearing objectively better than all other representations. When verifying whether metric space-based representations can be used for more remote inferences where the chemical space is evaluated in regions distant from the training data, the above conclusions regarding metric space modeling appear to have been amplified, and a larger distance between similarity methods was observed, with NAMS and Atom Pair fingerprints appearing clearly separated from the others.

Finally, metric space-based methods are more computationally expensive, requiring the computation of molecular similarity to every instance of the training set for each new molecule. This is a particularly severe cost for the graph-based similarity algorithm used (NAMS), and computation cost is a serious factor that may hamper its applicability in a real world virtual screening approach, despite overall being the method that is more consistently capable of producing high-quality QSAR models.

Supplementary Materials: The following are available online, Supplementary data (Additional file 1) contains three supplementary tables including: Table S1: List of RDkit 2D and 3D descriptors, Table S2: 5-fold cross-validation results, Table S3: External validation results, Figure S1: Selection of optimized number of PCs: PVE vs. number of PCs plot from PCA on metric space.

Author Contributions: A.O.F. designed and supervised the study. S.K., under the guidance and support of A.O.F., performed all analyses. All authors contributed to manuscript writing and approved its final version.

Funding: This research was funded by the Foundation for Science and Technology (Fundação para a Ciência e Tecnologia, FCT) for MIMED project funding (PTDC/EEI-ESS/4923/2014).

Acknowledgments: The authors gratefully acknowledge the Foundation for Science and Technology (Fundação para a Ciência e Tecnologia, FCT) for a doctoral grant (SFRH/BD/111654/2015) and the LASIGE Research Unit, ref. UID/CEC/00408/2019 for providing the infrastructure.

Conflicts of Interest: The authors declare that they have no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

NAMS	Non-contiguous atom matching structure similarity
DR	Dimensionality reduction
PC	Principal component
OPC-models	Optimized number of PC-based models
SF-models	Selected number of features-based model
RF	Random forest
SVM	Support vector machines
IVS	Independent validation sets

References

1. Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y.C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010. [CrossRef]
2. Dudek, A.Z.; Arodz, T.; Galvez, J. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. *Comb. Chem. High Throughput Screen.* **2006**, *9*, 213–228. [CrossRef]
3. Hansch, C.; Maloney, P.P.; Fujita, T.; Muir, R.M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178–180. [CrossRef]
4. Yoo, C.; Shahlaei, M. The applications of PCA in QSAR studies: A case study on CCR5 antagonists. *Chem. Biol. Drug Des.* **2017**. [CrossRef] [PubMed]
5. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Volume 11; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2008; p. 688. [CrossRef]
6. Chávez, E.; Navarro, G.; Baeza-Yates, R.; Marroquín, J.L. Searching in Metric Spaces. *ACM Comput. Surv.* **2001**, *33*, 273–321. [CrossRef]
7. Gasteiger, J. *Handbook of Chemoinformatics: From Data to Knowledge*, Volumes 1–4; Wiley-VCH: Weinheim, Germany, 2008; pp. 1–1870. [CrossRef]
8. O'Boyle, N.M.; Sayle, R.A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminform.* **2016**, *8*, 36. [CrossRef] [PubMed]

9. Yasri, A.; Hartsough, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218–1227. [CrossRef] [PubMed]
10. Puzyn, T.; Leszczynski, J.; Cronin, M.T. *Recent Advances in QSAR Studies: Methods and Applications (Challenges and Advances in Computational Chemistry and Physics)*; Springer: Berlin, Germany, 2009.
11. 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. [CrossRef] [PubMed]
12. Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504. [CrossRef] [PubMed]
13. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* **2010**, *29*, 476–488. [CrossRef] [PubMed]
14. Lesk, A.M. *Introduction to Bioinformatics*, 4th ed.; Oxford University Press: Oxford, UK, 2014; p. 400.
15. Orengo, C.A.; Bateman, A. *Protein Families: Relating Protein Sequence, Structure, and Function*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2013; p. 552. [CrossRef]
16. Teixeira, A.L.; Falcao, A.O. Structural similarity based kriging for quantitative structure activity and property relationship modeling. *J. Chem. Inf. Model.* **2014**, *54*, 1833–1849. [CrossRef]
17. Martin, Y.C.; Kofron, J.L.; Traphagen, L.M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350–4358. [CrossRef]
18. Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity—A Review. *QSAR Comb. Sci.* **2003**, *22*, 1006–1026. [CrossRef]
19. Johnson, M.A.; Maggiora, G.M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, NY, USA, 1990.
20. Willett, P.; Barnard, J.M.; Downs, G.M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996. [CrossRef]
21. Bender, A.; Glen, R.C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218. [CrossRef]
22. Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 3186–3204. [CrossRef]
23. Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: Foundations, limitations and novel approaches. *Drug Discov. Today* **2007**, *12*, 225–233. [CrossRef] [PubMed]
24. Stumpfe, D.; Bajorath, J. Similarity searching. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 260–282. [CrossRef]
25. Maggiora, G.M.; Shanmugasundaram, V. Molecular Similarity Measures. In *Methods in Molecular Biology*; Springer: Clifton, NJ, USA, 2004; pp. 1–50. [CrossRef]
26. Bajorath, J. Molecular Similarity Concepts for Informatics Applications. In *Bioinformatics: Volume II: Structure, Function, and Applications*; Keith, J.M., Ed.; Springer: New York, NY, USA, 2017; pp. 231–245. [CrossRef]
27. James, C.; Weininger, D.; Delaney, J. *Daylight Theory Manual Version 4.9*; Daylight Chemical Information Systems, Inc.: Laguna Niguel, CA, USA, 2011.
28. Teixeira, A.L.; Falcao, A.O. Noncontiguous atom matching structural similarity function. *J. Chem. Inf. Model.* **2013**, *53*, 2511–2524. [CrossRef] [PubMed]
29. 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. [CrossRef]
30. Raymond, J.W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521–533. [CrossRef] [PubMed]
31. Barnard, J.M. Substructure searching methods: Old and new. *J. Chem. Inf. Model.* **1993**, *33*, 532–538. [CrossRef]
32. Flower, D. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Model.* **1998**, *38*, 379–386. [CrossRef]
33. Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **2015**, *7*, 1–13. [CrossRef]

34. Tversky, A. Features of similarity. *Psychol. Rev.* **1977**, *84*, 327–352. [CrossRef]
35. Leskovec, J.; Rajaraman, A.; Ullman, J.D. *Mining of Massive Datasets*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2014.
36. Benigni, R.; Cotta-Ramusino, M.; Giorgi, F.; Gallo, G. Molecular similarity matrices and quantitative structure-activity relationships: A case study with methodological implications. *J. Med. Chem.* **1995**, *38*, 629–635. [CrossRef] [PubMed]
37. So, S.S.; Karplus, M. Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications. *J. Med. Chem.* **1997**, *40*, 4360–4371. [CrossRef]
38. Robert, D.; Amat, L.; Carbó-Dorca, R. Quantum similarity QSAR: Study of inhibitors binding to thrombin, trypsin, and factor Xa, including a comparison with CoMFA and CoMSIA methods. *Int. J. Quantum Chem.* **2000**, *80*, 265–282. [CrossRef]
39. Gironés, X.; Carbó-Dorca, R. Molecular quantum similarity-based QSARs for binding affinities of several steroid sets. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1185–1193. [CrossRef]
40. Besalú, E.; Gironés, X.; Amat, L.; Carbó-Dorca, R. Molecular quantum similarity and the fundamentals of QSAR. *Acc. Chem. Res.* **2002**, *35*, 289–295. [CrossRef]
41. Carbó-Dorca, R. About the prediction of molecular properties using the fundamental Quantum QSPR (QQSPR) equation †. *SAR QSAR Environ. Res.* **2007**, *18*, 265–284. [CrossRef] [PubMed]
42. Carbó-Dorca, R.; Mezey, P.G. *Advances in Molecular Similarity*; Number v. 2 in Advances in Molecular Similarity; Elsevier Science: Amsterdam, The Netherlands, 1999; p. 296.
43. Urbano Cuadrado, M.; Luque Ruiz, I.; Gómez-Nieto, M.Á. A Steroids QSAR Approach Based on Approximate Similarity Measurements. *J. Chem. Inf. Model.* **2006**, *46*, 1678–1686. [CrossRef]
44. Girschick, T.; Almeida, P.R.; Kramer, S.; Staišlring, J. Similarity boosted quantitative structure-activity relationship—A systematic study of enhancing structural descriptors by molecular similarity. *J. Chem. Inf. Model.* **2013**. [CrossRef]
45. Luque Ruiz, I.; Gómez-Nieto, M.Á. QSAR classification and regression models for β-secretase inhibitors using relative distance matrices. *SAR QSAR Environ. Res.* **2018**, *29*, 355–383. [CrossRef] [PubMed]
46. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954. [CrossRef]
47. Kausar, S.; Falcao, A.O. An automated framework for QSAR model building. *J. Cheminform.* **2018**, *10*, 1. [CrossRef]
48. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2009. [CrossRef]
49. Katritzky, A.R.; Lobanov, V.S.; Karelson, M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24*, 279–287. [CrossRef]
50. Gasteiger, J. *Handbook of Chemoinformatics*; Volumes 1–4; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2003; pp. 1–1870. [CrossRef]
51. Bajorath, J. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*, Volume 275; Humana Press: Totowa, NJ, USA, 2004. [CrossRef]
52. Roy, K.; Kar, S.; Das, R.N. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*; Elsevier: Amsterdam, The Netherlands, 2015. [CrossRef]
53. Varnek, A.; Baskin, I.I. Chemoinformatics as a theoretical chemistry discipline. *Mol. Inform.* **2011**, *30*, 20–32. [CrossRef]
54. Cereto-Massagué, A.; Ojeda, M.J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63. [CrossRef]
55. McGaughey, G.B.; Sheridan, R.P.; Bayly, C.I.; Culberson, J.C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.F.; Cornell, W.D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519. [CrossRef] [PubMed]
56. Muegge, I. Synergies of Virtual Screening Approaches. *Mini-Rev. Med. Chem.* **2008**, *8*, 927–933. [CrossRef] [PubMed]

57. Sheridan, R.P.; Kearsley, S.K. Why do we need so many chemical similarity search methods? *Drug Discov. Today* **2002**, *7*, 903–911. [CrossRef]
58. Zhang, Q.; Muegge, I. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. *J. Med. Chem.* **2006**, *49*, 1536–1548. [CrossRef] [PubMed]
59. Muegge, I.; Mukherjee, P. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin. Drug Discov.* **2016**, *11*, 137–148. [CrossRef] [PubMed]
60. Landrum, G. RDKit Documentation. *Release* **2018**, *1*, 1–79.
61. Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Model.* **1985**, *25*, 64–73. [CrossRef]
62. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef]
63. Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280. [CrossRef]
64. U.S. National Library of Medicine. *PubChem Substructure Fingerprint*; U.S. National Library of Medicine: Bethesda, MD, USA, 2009.
65. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33. [CrossRef]
66. Willett, P. The Calculation of Molecular Structural Similarity: Principles and Practice. *Mol. Inform.* **2014**, *33*, 403–413. [CrossRef]
67. Jasial, S.; Hu, Y.; Vogt, M.; Bajorath, J. Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Research* **2016**, *5*, 591. [CrossRef]
68. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*; Elsevier: Amsterdam, The Netherlands, 2012. [CrossRef]
69. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2009. [CrossRef]
70. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006**, *11*, 1046–1053. [CrossRef]
71. Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold Hopping Using Two-Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010**, *53*, 5707–5715. [CrossRef]
72. Willett, P. Similarity-based approaches to virtual screening. *Biochem. Soc. Trans.* **2003**, *31*, 603–606. [CrossRef]
73. Liu, P.; Long, W. Current mathematical methods used in QSAR/QSPR studies. *Int. J. Mol. Sci.* **2009**, *10*, 1978–1998. [CrossRef]
74. Lima, A.N.; Philot, E.A.; Goulart Trossini, G.H.; Barbour Scott, L.P.; Maltarollo, V.G.; Honorio, K.M. Use of machine learning approaches for novel drug discovery. *Expert Opin. Drug Discov.* **2016**, *11*, 225–239. [CrossRef]
75. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
76. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
77. Teixeira, A.L.; Leal, J.P.; Falcao, A.O. Random forests for feature selection in QSPR models—An application for predicting standard enthalpy of formation of hydrocarbons. *J. Cheminform.* **2013**, *5*, 1. [CrossRef]
78. Statnikov, A.; Wang, L.; Aliferis, C. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinform.* **2008**, *9*, 319. [CrossRef]
79. Yee, L.C.; Wei, Y.C. Current Modeling Methods Used in QSAR/QSPR. In *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2012; pp. 1–31. [CrossRef]
80. Varnek, A.; Baskin, I. Machine Learning Methods for Property Prediction in Chemoinformatics. *J. Chem. Inf. Model.* **2012**, *52*, 1413–1437. [CrossRef]
81. Gertrudes, J.C.; Maltarollo, V.G.; Silva, R.A.; Oliveira, P.R.; Honório, K.M.; da Silva, A.B.F. Machine learning techniques and drug design. *Curr. Med. Chem.* **2012**, *19*, 4289–4297. [CrossRef]
82. Dobchev, D.; Pillai, G.; Karelson, M. In silico machine learning methods in drug development. *Curr. Top. Med. Chem.* **2014**, *14*, 1913–1922. [CrossRef]
83. González, M.P.; Terán, C.; Saíz-Urra, L.; Teijeira, M. Variable selection methods in QSAR: An overview. *Curr. Top. Med. Chem.* **2008**, *8*, 1606–1627. [CrossRef]

84. Dehmer, M.; Varmuza, K.; Bonchev, D.; Emmert-Streib, F. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2012; p. 32434.
85. Genuer, R.; Poggi, J.M.; Tuleau-Malot, C. Variable selection using Random Forests. *Pattern Recognit. Lett.* **2012**, *31*, 2225–2236. [CrossRef]
86. Zaki, J.M.; Meira, W. *Data Mining and Analysis: Fundamental Concepts and Algorithms*; Cambridge University Press: New York, NY, USA, 2014.
87. Lee, J.A.; Verleysen, M. *Nonlinear Dimensionality Reduction*; Information Science and Statistics; Springer: New York, NY, USA, 2007. [CrossRef]
88. Eriksson, L.; Andersson, P.L.; Johansson, E.; Tysklind, M. Megavariate analysis of environmental QSAR data. Part I—A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). *Mol. Divers.* **2006**, *10*, 169–186. [CrossRef]
89. Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–701. [CrossRef]
90. Katritzky, A.R.; Petrukhin, R.; Tatham, D.; Basak, S.; Benfenati, E.; Karelson, M.; Maran, U. Interpretation of Quantitative Structure-Property and -Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 679–685. [CrossRef]
91. Genuer, R.; Poggi, J.M.; Tuleau, C. Random Forests: Some methodological insights. *Inria* **2008**, *6729*, 32.
92. Biau, G. Analysis of a Random Forests Model. *J. Mach. Learn. Res.* **2012**, *13*, 1063–1095.
93. Spiess, A.N.; Neumeyer, N. An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: A Monte Carlo approach. *BMC Pharmacol.* **2010**, *10*, 6. [CrossRef]
94. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500. [CrossRef]
95. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME—The Konstanz Information Miner. *SIGKDD Explor.* **2009**, *11*, 26–31. [CrossRef]
96. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Development Core Team: Vienna, Austria, 2011. [CrossRef]
97. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F. *Misc Functions of the Department of Statistics (e1071), TU Wien*; R Development Core Team: Vienna, Austria, 2014.
98. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
99. Kassambara, A.; Mundt, F. *Package 'Factoextra' for R: Extract and Visualize the Results of Multivariate Data Analyses*; R Development Core Team: Vienna, Austria, 2017.
100. Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Modeling robust QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 2310–2318. [CrossRef] [PubMed]
101. Fourches, D.; Muratov, E.; Tropsha, A. Trust but verify: On the importance of chemical structure curation in chemoinformatics and QSAR modeling research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204. [CrossRef]
102. Fourches, D.; Tropsha, A. Using graph indices for the analysis and comparison of chemical datasets. *Mol. Inform.* **2013**, *32*, 827–842. [CrossRef]
103. Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. Are the chemical structures in your QSAR correct? *QSAR Comb. Sci.* **2008**, *27*, 1337–1345. [CrossRef]
104. Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data set modelability by QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 1–4. [CrossRef]
105. Golbraikh, A.; Fourches, D.; Sedykh, A.; Muratov, E.; Liepina, I.; Tropsha, A. *Modelability Criteria: Statistical Characteristics Estimating Feasibility to Build Predictive QSAR Models for a Dataset*; Springer: Boston, MA, USA, 2014; pp. 187–230. [CrossRef]
106. Marcou, G.; Horvath, D.; Varnek, A. Kernel Target Alignment Parameter: A New Modelability Measure for Regression Tasks. *J. Chem. Inf. Model.* **2016**, *56*, 6–11. [CrossRef]
107. Hollander, M.; Wolfe, D.; Chicken, E. *Nonparametric Statistical Methods*, 3rd ed.; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2015. [CrossRef]
108. Mendiburu, F.D. *Agricolae: Statistical Procedures for Agricultural Research*; R Package Version 1.2-8; R Package Team: Vienna, Austria, 2017.

109. Tetko, I.V.; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746. [CrossRef]
110. Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatical, P.; Öberg, T.; Dao, P.; Cherkasov, A.; Tetko, I.V. Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. *J. Chem. Inf. Model.* **2008**, *48*, 766–784. [CrossRef]