JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

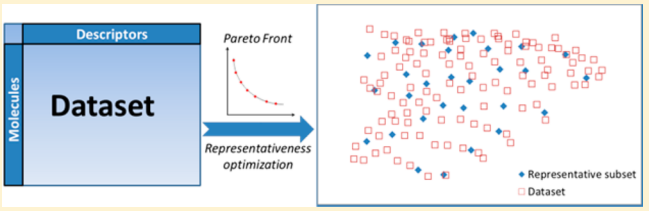
Optimization of Molecular Representativeness

Abraham Yosipof and Hanoch Senderowitz*

Department of Chemistry, Bar Ilan University, Ramat-Gan 52900, Israel

Supporting Information

ABSTRACT: Representative subsets selected from within larger data sets are useful in many chemoinformatics applications including the design of information-rich compound libraries, the selection of compounds for biological evaluation, and the development of reliable quantitative structure−activity relationship (QSAR) models. Such subsets can overcome many of the problems typical of diverse subsets, most notably the tendency of the latter to focus on outliers. Yet only a few algorithms for the selection of representative subsets have been reported in the literature. Here we report on the development of two algorithms for the selection of representative subsets from within parent data sets based on the optimization of a newly devised representativeness function either alone or simultaneously with the MaxMin function. The performances of the new algorithms were evaluated using several measures representing their ability to produce (1) subsets which are, on average, close to data set compounds; (2) subsets which, on average, span the same space as spanned by the entire data set; (3) subsets mirroring the distribution of biological indications in a parent data set; and (4) test sets which are well predicted by qualitative QSAR models built on data set compounds. We demonstrate that for three data sets (containing biological indication data, logBBB permeation data, and *Plasmodium falciparum* inhibition data), subsets obtained using the new algorithms are more representative than subsets obtained by hierarchical clustering, *k*-means clustering, or the MaxMin optimization at least in three of these measures.



1. INTRODUCTION

The introduction of combinatorial chemistry and high throughput screening (HTS) techniques in the 1990s has presented the drug development community with new opportunities for the synthesis and biological testing of large arrays of drug like compounds. However, it soon became apparent that even under these favorable conditions, only a small fraction of the accessible chemistry space could be synthesized and tested.[1] This in turn has pioneered the development of rational approaches for the design of screening libraries. One of the most common criteria typically employed in such design efforts is diversity, namely, the attempt to include within the screening library compounds which are different from one another as much as possible. The rationale behind selecting diverse sets is rooted in the similar property principle[2] which states that similar compounds have similar properties. A natural outcome of this principle is that structurally similar compounds tend to be redundant in terms of the structure−activity relationship (SAR) information they convey. Thus, diverse subsets are expected to be more information-rich. Multiple diversity algorithms[3] have been developed including dissimilarity-based compound selection algorithms,[4,5] clustering algorithms,[6,7] cell-based algorithms,[8,9] and optimization methods.[10−12] Treating the selection of a diverse subset as an optimization problem has the advantage that this objective could be combined with other objectives (e.g., cost, ADME/t properties, etc.) into a multiobjective optimization problem (MOOP).[13,14] MOOP problems could be solved either by combining the individual objectives into a single fitness function using a set of user-defined weights[15] or through Pareto-based optimization.[13,16,17]

Maximally diverse subsets tend to be biased toward the inclusion of outliers and consequently misrepresent the bulk of the compound collection from which they were selected. In drug discovery, focusing on outliers may translate into selecting and testing "extreme" compounds (e.g., high molecule weight compounds, complex compounds, or compounds with unrealistically large number of chemical groups) which are often not ideal starting points for subsequent lead optimization campaigns.[4] In cases where extreme compounds could be identified based on a single descriptor, they could be easily removed from the compound collection using simple filters. However, when compound "extremeness" results from a combination of descriptors, their identification and subsequent removal are more challenging. Such compounds are therefore often retained in the data set.

Some of the problems encountered with maximally diverse subsets could be overcome by selecting representative subsets, namely, subsets that mirror the distribution of the parent data set in some predefined descriptors space. Under this paradigm, members of a representative subset are still readily distinguishable from one another (i.e., diverse) yet are not bizarre (unless many bizarre compounds exist in the parent data set) thereby embodying the potential of the entire parent data set. However, despite their potential usefulness, representative subsets have

Received: December 4, 2013

   dx.doi.org/10.1021/ci400715n | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

gained much less attention than diverse subsets and accordingly far fewer algorithms for their selection have been described in the literature. Clark[4,18] has developed the Optimizable K-Dissimilarity Selection algorithm (OptiSim) as a generalization of the maximum and minimum dissimilarity selection algorithms and demonstrated its ability to select subsets which are both diverse and representative as a function of the subsample size. OptiSim was compared with hierarchical agglomerative clustering methods and in many cases was found to produce better behaved clusters.[19] Yet, to the best of our knowledge, an algorithm for the direct selection of representative subsets has not been reported to date nor has the selection of such subsets been treated as an optimization problem.

Representative subsets could be used under two general scenarios: (1) A representative subset is selected and analyzed and the results are used to infer the properties of the parent data set. A typical example for this scenario is the selection of a representative subset from within a large data set composed of compounds with different biological activities. In this case, the distribution of activities within a representative subset is expected to mirror that of the data set. Furthermore, if this is indeed the case then testing only the subset will convey a similar amount of information as would have been gained by testing the entire data set. (2) A representative subset is selected, set aside, and used to validate models built on the rest of the data set. This scenario is often used for evaluating the performances of quantitative structure−activity relationship (QSAR) models (built from the unselected portion of the data set) in terms of their ability to accurately predict the activities of compounds residing within their applicability domain (represented by the subset).[20]

Here we present a new algorithm for the selection of a representative subset from a parent data set. This algorithm applies a Monte Carlo/simulated annealing[21] (MC/SA) procedure to the optimization of a representativeness function based on pairwise distances between subset and data set compounds. Like diversity, representativeness could be combined with other objectives into a MOOP problem. Here we demonstrate this concept by the Pareto-based optimization of the representativeness and the MaxMin functions. In order to test the performances of the new representativeness function either alone or in combination with MaxMin under the two scenarios discussed above, we devised several test cases and representativeness measures. First we demonstrate that for three different data sets (containing biological indication data from the Comprehensive Medicinal Chemistry (CMC) database,[22] logBBB permeation data, and *Plasmodium falciparum* inhibition data), the new algorithms are able to produce subsets which are closer to data set compounds and which better span the chemical space of the parent data sets in comparison with other subset selection algorithms (hierarchical clustering, *k*-means clustering, or the MaxMin optimization). Then, for the first scenario, we demonstrate that a representative subset selected by the two new algorithms adequately mirrors the distribution of biological indications in the CMC database. This is in marked contrast with subsets selected by the other methods. This test case is also used to assess the ability of the new algorithms to select a relatively large representative subset (hundreds of compounds) from within a larger data set (thousands of compounds). Finally, for the second scenario, we take advantage of machine learning principles[23] in general and more specifically of recent developments in the QSAR field. Following the work of Tropsha[24] and others,[20] it is today widely recognized that the true predictive power of any QSAR model could only be evaluated from its performances on an external test set, namely, a set of compounds which was not used for model construction. Such test sets could include new compounds with relevant activity data if these are available or could be drawn from the modeling set. In the latter case, a test set could be extracted from the modeling set using either random or rational approaches. Randomly selected test sets are running the risk that certain regions in chemistry space populated by the modeling set compounds will not be well represented by the test set thereby providing biased performance measures. In contrast, test sets that span the chemistry space populated by the modeling set well could provide better performance estimates for QSAR models albeit only for compounds which reside within their applicability domain. Indeed Martin et al.[25] have recently demonstrated that rational selection methods yield better prediction statistics for test set compounds. Thus, we argue that such statistics provide a good estimate of how well a set of compounds represents the parent data set from which it was selected. Here we demonstrate that for two modeling sets (logBBB and *Plasmodium falciparum* inhibition) and five classification techniques (decision trees, random forests, ANN, SVM, and *k*NN) subsets obtained using the new representativeness function or found on the Pareto front generated by the simultaneous optimization of representativeness and MaxMin are better predicted than subsets selected by hierarchical clustering, *k*-means clustering, or the MaxMin function suggesting that they are indeed more representative.

2. MATERIAL AND METHODS

2.1. Experimental Data Sets. The first data set corresponds to the Comprehensive Medicinal Chemistry (CMC) database.[22] This database currently contains 9522 pharmaceutical compounds classified into different indications. Indications including one or two compounds were removed leading to a filtered data set consisting of 4855 compounds which cover 105 different biological indications (Supporting Information Table S1).

The second data set consists of 152 compounds with known logBBB permeation values which were compiled from the literature (Supporting Information Table S2).[26−28] Structures of all compounds were validated by manual inspection. Compounds were classified into two groups, namely, BBB permeable (BBB^+, logBBB $\geq$ 0, 81 compounds) and BBB nonpermeable (BBB^-, logBBB < 0, 71 compounds).

The third data set is based on inhibition data obtained in a high throughput cell based assay measuring the proliferation of *Plasmodium falciparum* (*Pf*) in erythrocytes. These data were retrieved from the work of Huwyler et al.[29] The original data set consisted of 550 compounds (201 active, 349 inactive). In the present study, 9 compounds for which we were unable to calculate all descriptors were omitted leaving a total of 541 compounds with 195 active and 346 inactive (Supporting Information Table S3).

2.2. Descriptors Calculation. Molecular descriptors were calculated using Discovery Studio version 3.5.[30] A total of 50 2D descriptors were calculated for the CMC data set, a total of 160 2D and 3D descriptors were calculated for the logBBB data set, and a total of 150 1D and 2D descriptors were calculated for the *Pf* data set. In all cases, the initial set of descriptors was filtered by removing constant descriptors and correlated

dx.doi.org/10.1021/ci400715n | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

descriptors. Finally, for the logBBB and *Pf* data sets for which qualitative activity data (in the form active/inactive) are available, descriptors were ranked according to the information gain filter.[23] Briefly, the information gain of a descriptor reflects the ability of this descriptor to split the data set into two "homogeneous" groups. Homogeneity is determined according to Shannon's entropy measure. These filtration procedures led to final sets of 8, 15, and 10 descriptors for the CMC, logBBB, and *Pf* data sets, respectively (Tables 1, 2, 3, S2, and S3). The diversities of the data sets were estimated by pairwise Euclidean distance distributions calculated in the space defined by their respective (normalized) descriptors. The resulting histograms are presented in Supporting Information Figures S1 (CMC), S2 (logBBB), and S3 (*Pf*).

Table 1. Descriptors Selected for the CMC Data Set

Structural and Thermodynamic	
AlogP98	Molecular Solubility
# Hydrogen bond acceptor	# Aromatic Bonds
# Hydrogen bond donor	# Rings
# Rotatable bonds	# Double Bonds

Table 2. Descriptors Selected for the logBBB Data Set

Jurs descriptors	structural and thermodynamic	E-state keys	spatial descriptors	topological descriptor
TPSA	AlogP98	S_sssN	Radius of gyration	Balaban index JY
FNSA3	# Hydrogen bond acceptor	S_ssCH$_2$		
DPSA1	# Hydrogen bond donor	S_dO		
RPCG	# Rotatable bonds	S_aaCH		
RNCG				

Table 3. Descriptors Selected for the *Plasmodium falciparum* Inhibition Assay Data Set

atom counts	structural and thermodynamic	E-state keys	topological descriptor
O_Count	AlogP98	ES_Count_aaCH	JX
	Molecular_Solubility	ES_Count_aasC	
	Molecular_SurfaceArea	ES_Count_sCH$_3$	
		ES_Count_ssCH$_2$	
		ES_Count_ssO	

2.3. Selection of Representative Subsets. Representative subsets of different sizes were selected from the CMC, logBBB, and the *Plasmodium falciparum* data sets. Selections were performed using five different algorithms including two newly developed ones, namely, optimization of a new representativeness function and Pareto-based optimization of the representativeness function and the MaxMin function and three common algorithms taken from the literature (diversity selection using MaxMin optimization, hierarchical clustering using the Ward's method, and *k*-means clustering). For the CMC data set, a single subset of 200 compounds was selected. For the logBBB and *Pf* data sets three subsets of different sizes were selected. This effectively led to splitting of these data sets into three pairs of modeling and test sets as indicated in Table 4. For all splits, similar proportions of active and inactive compounds were selected for the test sets by applying independent selection procedures to the two activity categories.

2.3.1. Representativeness Optimization. Here we present a new algorithm for the direct selection of a representative subset from within a data set. Representativeness is treated as an optimization problem, and accordingly, we define a representativeness function, based on pairwise distances between subset and data set compounds which is optimized by means of Monte Carlo/simulated annealing (MC/SA) procedure. Our algorithm consists of the following steps (Figure 1):

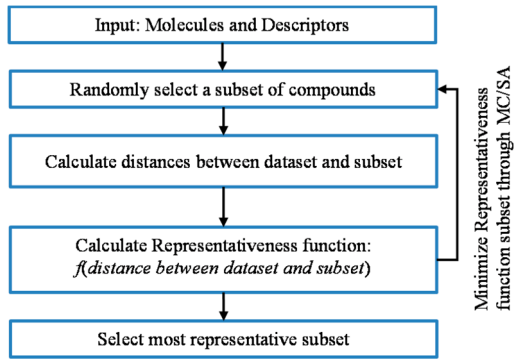


Figure 1. Schematic representation of the new representativeness optimization algorithm.

1. Characterize each compound in the data set by a set of molecular descriptors.
2. Normalize descriptors by converting them into Z-scores according to eq 1 where μ is the mean and σ is the standard deviation:

$$Z = \frac{x_i - \mu}{\sigma} \quad (1)$$

The resulting scores are then converted to [0, 1] range by calculating the cumulative probability assuming a normal distribution ($\mu = 0$; $\sigma^2 = 1$).
3. Select a random subset s of size k from within the l compounds comprising the data set.
4. Calculate the Euclidean distance between compound i from the data set and all k compounds comprising subset s.
5. Take the minimum Euclidean distance from step 4 as the score for compound i: $\text{score}_i = \min(\text{dist}_{i,\{s\}})$.

Table 4. Modeling and Test Sets for the logBBB and *Plasmodium falciparum* Inhibition Assay Data Sets

	logBBB				*Plasmodium falciparum*			
	modeling set		test set		modeling set		test set	
split	active/inactive	total	active/inactive	total	active/inactive	total	active/inactive	total
1	71/61	132	10/10	20	165/296	461	30/50	80
2	61/51	112	20/20	40	155/276	431	40/70	110
3	51/41	92	30/30	60	130/231	361	65/115	180

dx.doi.org/10.1021/ci400715n | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

6. Repeat steps 4 and 5 for all $l-k$ compounds remaining in the data set.
7. Calculate the average score over all $l-k$ compounds. This score characterizes subset s: $\text{score}_s = (1/l - k)\sum_{i=1}^{l-k} \text{score}_i$.
8 Minimize score_s through a MC/SA procedure.[21] At each step replace, at random, a single compound from s with a candidate compound from the unselected portion of the data set, calculate a new score, $\text{score}_{s'}$, and accept it with the Metropolis probability: $P_{acc} = \min(1, e^{-(\Delta\text{score}/kT)})$ where Δscore is $\text{score}_s - \text{score}_{s'}$ and kT is the effective temperature. In the current implementation, the MC procedure was typically run for 10^6 steps replacing a single compound at each step and the effective temperature was set to produce an initial acceptance rate of ∼10% and an average acceptance rate of ∼0.5%.

This algorithm ensures the closeness of the data set compounds to the subset but not the opposite, that is, the closeness of the subset compounds to the data set (but, see the Results section below). The complexity of the representativeness function is $O(\delta M(N - M))$ where δ is the number of dimensions (descriptors), N is the number of compounds in the data set, and M is the number of compounds in the subset. The algorithm was coded in C++ by using the visual studio professional 2012. The code runs on a Linux cluster.

2.3.2. MaxMin Optimization. MaxMin optimization was performed using an in-house implementation of the algorithm originally described by Hassan et al.[10] Briefly, for a given subset, MaxMin calculates the square of the minimal distance, d_{ij}^2, over all (i, j) pairs comprising the subset according to eq 2:

$$\text{MaxMin} = \text{Max}(\underset{i\neq j}{\text{Min}}(d_{ij}^2)) \quad (2)$$

Where d_{ij} is the distance between compounds i and j and the summation runs over all the descriptors. The MaxMin function is optimized (maximized) by means of a MC/SA algorithm to produce the subset with the largest value. The present implementation started from a randomly selected subset (see Table 4) which was typically optimized by 10^6 MC steps. At each step, a single compound was replaced and the effective temperature was set to produce an initial acceptance rate of ∼10% and an average acceptance rate of ∼0.5%.

2.3.3. Pareto-Based Optimization. A major advantage in treating the selection of a representative subset as an optimization problem is the ability to combine it with additional objectives into a multiobjective optimization problem (MOOP).[13,14] In contrast with single-objective optimization (SOOP), in MOOP there is no guarantee that a single optimal solution exists that outperforms all other solutions in all criteria. Instead, several equally good (nondominated) solutions exist representing various compromises among the objectives. A nondominated solution is one where an improvement in one objective results in deterioration in one or more of the other objectives when compared with the other solutions in the population. The set of nondominated solutions represents the Pareto front.

In this work we simultaneously optimized the above-described representativeness function together with the MaxMin function using Pareto-based optimization. The Pareto-based optimization algorithm evaluates the MaxMin function (eq 2) and the representativeness function (step 7 above) for a selected subset (termed a solution to the MOOP problem) and assigns to it a Pareto rank based on the number of solutions dominating it. A nondominant solution (i.e., a solution on the Pareto front) is assigned a rank of zero. In the present case, solution i dominates solution j if $\text{MaxMin}(i) < \text{MaxMin}(j)$ and $\text{score}(i) < \text{score}(j)$ (where score is calculated according to step 7 above). Note that under this selection of dominance criteria the value of MaxMin is minimized. This is in contrast with the Maximization of the MaxMin function which is typically used for diversity selection. MaxMin minimization helps to further bias the selected subset toward the more populated regions of the database allowing the function to work in concert with the representativeness function and not against it. The Pareto rank is then optimized using Metropolis Monte Carlo, the solutions with rank = 0 are kept and are used to construct the Pareto front. Finally, a solution on the Pareto front is randomly selected.

2.3.4. Hierarchical Clustering. In hierarchical clustering a dendogram of clusters is produced in which the root node represents a single cluster that contains all the compounds and each leaf node represents a single compound. Agglomerate hierarchical clustering starts with the leaf nodes and iteratively combines closest neighbors clusters until the root node is reached, whereas divisive clustering starts at the root node and iteratively divides clusters until the leaf nodes are reached. In the present case we used the Ward's hierarchical clustering method as implemented in the WEKA version 3.7.9 software.[31] Ward's clustering operates by combining clusters so as to minimize the total within-cluster variance. Following clustering, the resulting dendogram was cut at the level which produced the desired number of clusters (e.g., size of the subset), a single compound was selected from each cluster for the subset (test set), and all other compounds were selected for the modeling set.

2.3.5. k-Means Clustering. k-Means clustering[32] is a nonhierarchical clustering method. It proceeds iteratively with the first step involving the selection of a user-defined number of seed compounds, k, and an initial set of clusters is formed by assigning each compound to its closest seed. The centroid of each cluster is then calculated and the compounds are relocated to their closest centroid. This process is repeated for a user defined number of iterations or until the clusters are stable, that is, no compounds are relocated. Following clustering, a single compound was selected from each cluster for the subset (test set) and all other compounds were selected for the modeling set. k-Means clustering was performed using WEKA version 3.7.9.[31]

2.4. Classification-Based QSAR Methods. For each division of the data sets into modeling and test sets (Table 4), classification models for the logBBB and *Plasmodium falciparum* data sets were generated using five different algorithms, namely, decision tree, random forests (RF), support vector machine (SVM), artificial neuronal network (ANN), and k nearest neighbors (kNN). In each case a QSAR model was built and validated using the modeling set and subsequently used to predict the activities of the test set compounds. All models were generated with algorithms implemented in the WEKA version 3.7.9 software[31] using default parameters unless otherwise noted.

The decision tree algorithm[33] operates by iteratively splitting a data set characterized by activity data and descriptors into smaller subsets. At each step, all descriptors are considered in search for one that upon splitting a parent node would produce the most uniform (activity-wise) child nodes. This procedure is

dx.doi.org/10.1021/ci400715n | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

repeated until no more splits are warranted either since all compounds within all (terminal) nodes have identical activities or since the gain in uniformity upon additional splits is not statistically significant. In the present study we used the J4.8, a C4.5 variant algorithm.

In 2001, Breiman[34] introduced the principle of random forests (RF) as an extension to the decision tree algorithm. In RF multiple trees (rather than a single tree) are generated using randomly selected feature sets. Activity predictions are made by all trees and combined using a majority vote rule. In the present study the number of trees was set to the default value of 10.

Support vector machine (SVM)[35] is an algorithm which has proven useful for noisy data. Under this paradigm, models are built by identifying a rigid decision hyperplane which leads to the greatest possible margins between activity classes. Nonlinear data could be handled by transposing the original feature space to higher dimensionalities using kernels. In this study, we have chosen to use the polynomial kernel function.

Artificial neuronal network (ANN)[36] is a nonlinear classification method inspired by the behavior of biological networks of neurons. Within this approach, objects (i.e., compounds) are represented by vectors containing their features (i.e., descriptors). Each feature is passed to one of the input neurons to which a weight is assigned. On the basis of these weights, input is passed to the output layer over a number of (optional) hidden layers. The output layer combines these signals to produce a result (e.g., activity prediction). Initially, weights are set to random values. As the network is repeatedly presented with input data, these weights are adjusted so that the total output of the network approximates the observed end point values associated with the compounds. In the present study we used multilayer perceptrons (MLP) with a single hidden layer.

k-Nearest neighbor (*k*NN)[23] is a lazy learning classification method, which assigns new compounds to the most common class of known compounds in their immediate neighborhood. Closest neighbors are identified by calculating Euclidian distances in a predefined descriptor space.

2.5. Evaluation of Subset Representativeness.

2.5.1. $1NN_{data\ set}$ and $1NN_{subset}$ Indices. In order to provide a direct measure of subset representativeness we calculate for each subset two indices, namely, $1NN_{data\ set}$ and $1NN_{subset}$ which correspond, respectively, to the average distance of each compound from the unselected portion of the data set (e.g., modeling set) to its closest neighbor in the subset (e.g., test set) and the average distance of each compound from the subset set to its closest neighbor in the unselected portion of the data set.

2.5.2. χ^2 Goodness of Fit Test. Each of the 200 compounds subsets selected from the CMC database was evaluated for its ability to accurately mirror the distribution of indications in the parent database using the χ^2 goodness of fit test. Under this test, the null hypothesis (H_0) states that the distribution of biological indications within the subset and data set are similar, i.e. the subset represents the data set distribution, whereas the H_1 hypothesis states that they are significantly different from each other. The goal in the present case is therefore to stay on the null hypothesis. The χ^2 statistics is defined as

$$\chi^2 = \sum_{i=1}^{n=105} \frac{(O_i - E_i)}{E_i} \tag{3}$$

where O_i and E_i represent respectively, the observed and expected frequencies for a biological indication i in the 200 compounds subset. E_i is derived from the frequency of indication i in the parent data set.

2.5.3. Prediction Statistics. For the logBBB and *Plasmodium falciparum* data sets, models were internally validated by using 10-fold cross validation[37] and externally validated using the selected test sets.

In all cases activity predictions were evaluated using the corrected classification rate (CCR) and Matthews correlation coefficient (MCC) measures. CCR is given by

$$\mathrm{CCR} = \frac{1}{2}\left(\frac{T_N}{N_N} + \frac{T_P}{N_P}\right) \tag{4}$$

*W*here, T_N and T_P represent the number of true negative and true positive predictions, respectively, and N_N and N_P represent the total number of the two activity classes. MCC is given by

$$\mathrm{MCC} = \frac{T_N T_P - F_N F_P}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}} \tag{5}$$

Where F_N and F_P denote false negative and false positive predictions, respectively.

3. RESULTS

3.1. $1NN_{data\ set}$ and $1NN_{subset}$ Indices. $1NN_{data\ set}$ and $1NN_{subset}$ indices for the CMC, logBBB, and *Pf* data sets are given in Tables 5, 6, and 7, respectively. As noted in section 2.3, for the logBBB and *Pf* data sets individual subset selections were performed for active and inactive compounds. Thus, in these cases $1NN_{data\ set}$ and $1NN_{subset}$ were summed over the two subsets.

Table 5. Representative Indices for the CMC Data Set[a]

method	$1NN_{data\ set}$	$1NN_{subset}$
hierarchical clustering	0.23	0.16
k-means clustering	0.25	0.18
MaxMin optimization	0.24	0.16
representativeness optimization	0.19	0.10
Pareto-based optimization	0.19	0.10

[a] $1NN_{data\ set}$ denotes the average distance of each compound in the unselected part of the data set to its closest neighbor in the subset. $1NN_{subset}$ denotes the average distance of each compound in the subset to its closest neighbor in the unselected part of the data set. For the CMC data set, a 200 compounds subset was selected.

These results clearly demonstrate that in all cases the optimization of the representativeness function leads to the best (smallest) $1NN_{data\ set}$ whereas Pareto-based optimization yields the best (smallest) $1NN_{subset}$ values, independent of the size of the selected subset (a single exception is found for the logBBB data set where $1NN_{subset}$ for split 3 is lower for *k*-means than for the Pareto-based optimization). Furthermore, in all cases except two, representativeness and Pareto-based optimizations yield the two best values for both indices.

3.2. Results for the CMC Data Set. We have used the different algorithms to choose a subset of 200 compounds from within the 4855 compounds comprising the filtered CMC data set. To determine whether the resulting subsets indeed represent the distribution of biological indication within the parent data set, we used the χ^2 goodness of fit test. The results of this test are presented in Table 8 and demonstrate that the distribution of indications within the subsets selected by the representativeness optimization and the Pareto-based optimi-

dx.doi.org/10.1021/ci400715n | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 6. Representative Indices for the logBBB Data Set[a]

method	$1NN_{data\ set}$	$1NN_{subset}$
Split 1		
hierarchical clustering	1.59	1.04
k-means clustering	1.68	1.14
MaxMin optimization	1.72	1.12
representativeness optimization	1.36	0.82
Pareto-based optimization	1.41	0.75
Split 2		
hierarchical clustering	1.35	1.24
k-means clustering	1.41	1.15
MaxMin optimization	1.36	1.27
representativeness optimization	1.12	1.07
Pareto-based optimization	1.18	1.06
Split 3		
hierarchical clustering	1.13	1.28
k-means clustering	1.18	1.11
MaxMin optimization	1.12	1.29
representativeness optimization	0.98	1.23
Pareto-based optimization	1.00	1.19

[a] $1NN_{data\ set}$ denotes the average distance of each compound in the unselected part of the data set to its closest neighbor in the subset. $1NN_{subset}$ denotes the average distance of each compound in the subset to its closest neighbor in the unselected part of the data set. In this case the unselected part of the data set and the subset correspond to the modeling set and test set respectively. Subsets of three different sizes were selected (splits 1, 2, and 3, see Table 4).

Table 7. Representative Indices for the *Pf* Data Set[a]

method	$1NN_{data\ set}$	$1NN_{subset}$
Split 1		
hierarchical clustering	1.01	0.80
k-means clustering	1.00	0.78
MaxMin optimization	1.12	0.92
representativeness optimization	0.85	0.62
Pareto-based optimization	0.93	0.69
Split 2		
hierarchical clustering	0.93	0.84
k-means clustering	0.91	0.76
MaxMin optimization	1.00	0.91
representativeness optimization	0.79	0.69
Pareto-based optimization	0.85	0.66
Split 3		
hierarchical clustering	0.80	0.89
k-means clustering	0.82	0.76
MaxMin optimization	0.85	0.90
representativeness optimization	0.68	0.78
Pareto-based optimization	0.78	0.74

[a] $1NN_{data\ set}$ denotes the average distance of each compound in the unselected part of the data set to its closest neighbor in the subset. $1NN_{subset}$ denotes the average distance of each compound in the subset to its closest neighbor in the unselected part of the data set. In this case the unselected part of the data set and the subset correspond to the modeling set and test set respectively. Subsets of three different sizes were selected (splits 1, 2, and 3, see Table 4).

zation are statistically indistinguishable from that in the parent database. In contrast, subset selected by *k*-means clustering, hierarchical clustering, and MaxMin display distributions which are markedly different from the data set.

3.3. Results for the logBBB and *Plasmodium falciparum* Data Sets. In addition to the $1NN_{data\ set}$ and $1NN_{subset}$ indices described above, we provide an indirect measure of subset representativeness for the logBBB and *Pf* data sets. This measure is based on ideas discussed above, namely, that activity predictions obtained for a test set (but not for a modeling set) are good estimates of how well this test set represents the parent data set from which it was selected and that better results correspond to more representative selections. With this in mind, we evaluated the performances of the different selection algorithms considered in this work by calculating the corrected classification rate (CCR) and the Matthews correlation coefficient (MCC) over all subset (i.e., test set) compounds.

Table 8. Results of χ^2 Test of Goodness of Fit for the Subset of 200 Compounds Selected from the CMC Data Set

method	χ^2	df	*p*-value
hierarchical clustering	274.97	104	<0.001
k-means clustering	272.92	104	<0.001
MaxMin optimization	133.93	104	<0.05
representativeness optimization	91.16	104	>0.05
Pareto-based optimization	93.66	104	>0.05

3.3.1. Results for the logBBB Data Set. Subsets of three different sizes (splits 1, 2, and 3 corresponding to test set sizes of 20, 40, and 60 compounds; see Table 4) were selected using each of the five algorithms considered in this work. The Pareto fronts from which the Pareto-based subsets were selected are shown in Figure 2. In all cases, the set selected by optimizing the representativeness function alone has the minimal *X*-value and a certain *Y*-value and is consequently found on the Pareto front (marked by a square). For reasons discussed in section 2.3.3, the set with the maximal MaxMin value is not on the Pareto front.

The statistical results obtained from the different QSAR methods for the logBBB data set are presented in Tables S4 (modeling set) and 9 (test set). An almost perfect correlation ($R^2 > 0.98$) was found between the two measures (CCR and MCC) across all data points in both the training set and the test set, and consequently, only the CCR data will be further discussed.

The overall performances of the different test set selection algorithms considered in this work were evaluated by calculating average CCR values across all split sizes and classification methods. Based on this criterion, the best algorithm is the Pareto-based optimization method ($\tilde{CCR}$ = 0.80) followed by the new representativeness function, the *k*-means clustering, and the hierarchical clustering ($\tilde{CCR}$ = 0.76) while the MaxMin function ($\tilde{CCR}$ = 0.69) lags behind. Moreover, 8 out of the 13 best results (including the best result) for this data set were obtained when selecting the test set by means of either the representativeness or the Pareto method ($CCR^{Pareto}_{RF-Split1}$ = 0.95; $CCR^{hierarchical}_{SVM-Split1}$ = 0.90; $CCR^{Pareto}_{J48-Split1}$ = $CCR^{Pareto}_{ANN-Split1}$ = $CCR^{Pareto}_{SVM-Split1}$ = $CCR^{Pareto}_{kNN-Split1}$ = $CCR^{Pareto}_{RF-Split2}$ = $CCR^{representativeness}_{SVM-Split1}$ = $CCR^{representativeness}_{kNN-Split1}$ = $CCR^{hierarchical}_{RF-Split1}$ = $CCR^{hierarchical}_{ANN-Split1}$ = $CCR^{hierarchical}_{kNN-Split1}$ = $CCR^{k-means}_{ANN-Split2}$ 0.85; bolded in Table 9). However, these results are not evenly spread across all splits as 11 of the best values were obtained for the smallest (20 compounds) test set ($CCR^{Pareto}_{RF-Split1}$; $CCR^{Pareto}_{J48-Split1}$; $CCR^{Pareto}_{ANN-Split1}$; $CCR^{Pareto}_{SVM-Split1}$; $CCR^{representativeness}_{SVM-Split1}$; $CCR^{Pareto}_{kNN-Split1}$; $CCR^{representativeness}_{kNN-Split1}$; $CCR^{hierarchical}_{SVM-Split1}$; $CCR^{hierarchical}_{RF-Split1}$; $CCR^{hierarchical}_{ANN-Split1}$; $CCR^{hierarchical}_{kNN-Split1}$). For the smallest test sets, the performances of the hierarchical clustering method were similar to or even marginally better than the representativeness

dx.doi.org/10.1021/ci400715n | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

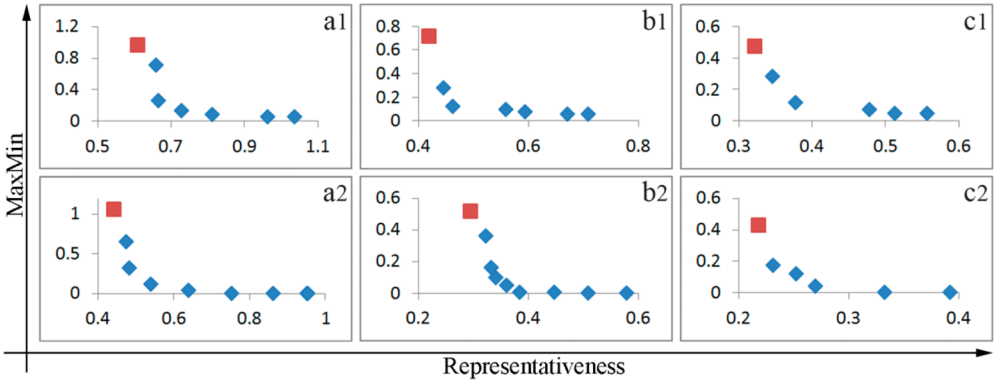


Figure 2. Pareto fronts for the logBBB data set obtained from the simultaneous optimization of the new representativeness function (step 7 above) and the MaxMin function (eq 2). The direction of the Pareto front is determined by the dominance criteria. See section 2.3.3 for more details. Parts a–c represent, respectively, the Pareto fronts for selecting subsets of 20 (split 1), 40 (split 2), and 60 (split 3) compounds. Since independent selection procedures were applied to each activity class (see text for more details), each split is represented by two curves where the upper one (denoted by 1) describes the selection from within the inactive pool (10, 20, and 30 for splits 1, 2, and 3, respectively) and the lower one (denoted by 2), the selection from within the active pool (10, 20, and 30 for splits 1, 2, and 3, respectively).

Table 9. CCR and MCC Values for the External Test Sets Selected from the logBBB Data Set[a]

		Split 1		split 2		split 3	
Model	Method	CCR	MCC	CCR	MCC	CCR	MCC
j48	Hierarchical clustering	0.80	0.65	0.68	0.36	0.67	0.34
	k-means clustering	0.75	0.50	0.75	0.50	0.83	0.67
	MaxMin	0.60	0.20	0.68	0.36	0.68	0.37
	Representativeness	0.80	0.65	0.75	0.50	0.72	0.44
	Pareto-based	**0.85**	0.73	0.80	0.60	0.72	0.48
RF	Hierarchical clustering	**0.85**	0.70	0.75	0.50	0.72	0.43
	k-means clustering	0.70	0.40	0.78	0.57	0.82	0.64
	MaxMin	0.80	0.60	0.73	0.46	0.68	0.37
	Representativeness	0.80	0.60	0.73	0.45	0.73	0.47
	Pareto-based	**0.95**	0.90	**0.85**	0.71	0.78	0.57
ANN	Hierarchical clustering	**0.85**	0.70	0.78	0.55	0.70	0.40
	k-means clustering	0.70	0.41	**0.85**	0.71	0.75	0.52
	MaxMin	0.60	0.20	0.65	0.31	0.72	0.43
	Representativeness	0.80	0.65	0.73	0.45	0.75	0.51
	Pareto-based	**0.85**	0.70	0.75	0.50	0.77	0.53
SVM	Hierarchical clustering	**0.90**	0.80	0.78	0.55	0.65	0.30
	k-means clustering	0.75	0.50	0.80	0.60	0.82	0.64
	MaxMin	0.60	0.20	0.73	0.45	0.73	0.47
	Representativeness	**0.85**	0.70	0.78	0.55	0.70	0.40
	Pareto-based	**0.85**	0.70	0.75	0.50	0.78	0.57
***k*NN**	Hierarchical clustering	**0.85**	0.70	0.78	0.56	0.70	0.40
	k-means clustering	0.65	0.30	0.73	0.45	0.75	0.51
	MaxMin	0.75	0.50	0.78	0.55	0.68	0.37
	Representativeness	**0.85**	0.70	0.78	0.55	0.65	0.30
	Pareto-based	**0.85**	0.70	0.75	0.50	0.70	0.40

[a]Splits 1, 2, and 3 represent, respectively, splitting the data set into 132 modeling and 20 test set compounds, 112 modeling and 40 test set compounds, and 92 modeling and 60 test set compounds.

method but lower than the Pareto-based optimization ($C\tilde{C}R_{Split1}$ = 0.85, 0.82, and 0.87 for hierarchical clustering, representativeness, and Pareto-based optimization, respectively). For the larger test sets, the performances of the *k*-means clustering method were similar to or even marginally better than those of the two new algorithms ($C\tilde{C}R_{Split2}$ = 0.78, 0.75, and 0.78 for *k*-means, representativeness, and Pareto-based optimization, respectively; $C\tilde{C}R_{Split3}$ = 0.79, 0.71, and 0.75 for *k*-means, representativeness, and Pareto-based optimization, respectively). Finally, the performances of the different QSAR classification methods across all split sizes and test set selection algorithms were almost identical ($C\tilde{C}R_{J48}$ = 0.74; $C\tilde{C}R_{RF}$ = 0.78; $C\tilde{C}R_{ANN}$ = 0.75; $C\tilde{C}R_{SVM}$ = 0.76; $C\tilde{C}R_{kNN}$ = 0.75). While evaluating different classification methods was not part of this work, these results demonstrate that the overall performances of the test set selection algorithms are not biased by the particular model building method.

3.3.2. Results for the Plasmodium falciparum (Pf) Data Set. Subsets of three different sizes (splits 1, 2, and 3 corresponding to test set size of 80, 110, and 180 compounds; see Table 4) were selected using each of the five algorithms considered in this work. The corresponding Pareto fronts are

dx.doi.org/10.1021/ci400715n | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

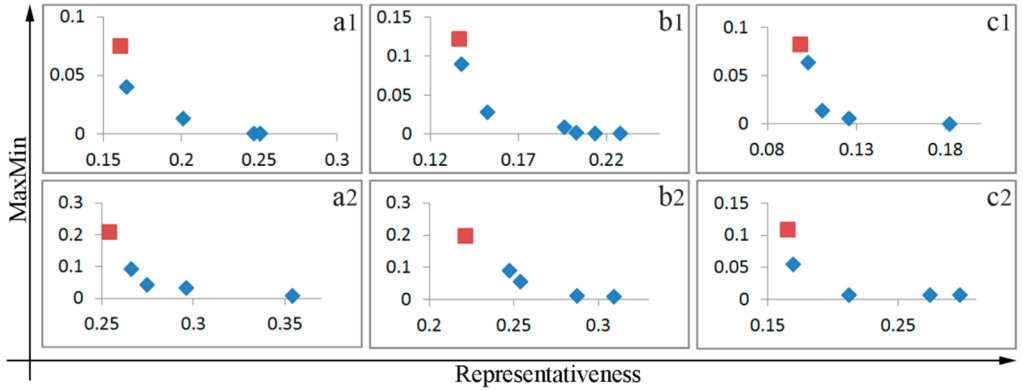


Figure 3. Pareto fronts for the *Plasmodium falciparum* inhibition assay data set obtained from the simultaneous optimization of the new representativeness function and the MaxMin function. The direction of the Pareto front is determined by the dominance criteria. See section 2.3.3 for more details. Parts a, b, and c represent, respectively, the Pareto fronts for selecting subsets of 80 (split 1), 110 (split 2), and 180 (split 3) compounds. Since independent selection procedures were applied to each activity class, each split is represented by two curves where the upper one (denoted by 1) describes the selection from within the inactive pool (50, 70, and 115 for splits 1, 2, and 3 respectively) and the lower one (denoted by 2), the selection from within the active pool (30, 40, and 65 for splits 1, 2, and 3, respectively).

Table 10. CCR and MCC Values for the External Test Sets Selected from the *Pf* Data Set[a]

		split 1		split 2		split 3	
model	method	CCR	MCC	CCR	MCC	CCR	MCC
*J*48	hierarchical clustering	0.77	0.54	0.76	0.50	0.78	0.57
	k-means clustering	**0.90**	0.81	**0.88**	0.76	0.78	0.57
	MaxMin	0.75	0.48	0.75	0.49	0.79	0.59
	representativeness	**0.86**	0.76	0.84	0.67	0.80	0.60
	Pareto-based	0.83	0.64	0.82	0.62	0.85	0.69
RF	hierarchical clustering	0.78	0.54	0.83	0.64	0.76	0.52
	k-means clustering	0.84	0.70	0.85	0.72	**0.86**	0.71
	MaxMin	0.73	0.47	0.74	0.51	0.83	0.66
	representativeness	0.83	0.71	0.85	0.69	0.85	0.70
	Pareto-based	**0.86**	0.70	0.85	0.69	**0.86**	0.72
ANN	hierarchical clustering	0.79	0.59	0.84	0.65	0.84	0.66
	k-means clustering	0.82	0.65	0.84	0.70	**0.88**	0.73
	MaxMin	0.78	0.54	0.72	0.45	0.80	0.59
	representativeness	**0.91**	0.84	**0.86**	0.76	0.85	0.71
	Pareto-based	**0.88**	0.76	**0.88**	0.78	**0.88**	0.76
SVM	hierarchical clustering	0.76	0.50	0.82	0.62	0.80	0.60
	k-means clustering	**0.86**	0.73	0.85	0.72	0.84	0.70
	MaxMin	0.81	0.62	0.79	0.59	0.82	0.66
	representativeness	0.85	0.73	0.84	0.70	0.83	0.68
	Pareto-based	0.85	0.70	0.85	0.72	**0.88**	0.76
*k*NN	hierarchical clustering	0.75	0.51	0.76	0.52	0.77	0.55
	k-means clustering	0.83	0.68	0.83	0.70	0.83	0.67
	MaxMin	0.80	0.60	0.83	0.66	0.78	0.57
	representativeness	0.84	0.74	0.84	0.70	0.83	0.66
	Pareto-based	0.84	0.68	0.85	0.74	0.84	0.69

[a]Splits 1, 2, and 3 represent, respectively, splitting the data set into 461 modeling and 80 test set compounds, 431 modeling and 110 test set compounds, and 361 modeling and 180 test set compounds.

shown in Figure 3. As before, the representativeness set was found to lie on the upper part of the Pareto front.

The statistical results for the *Pf* data set are presented in Tables S5 (modeling set) and 10 (test set). As for the logBBB data set, the correlation between CCR and MCC across all data points was found to be high (R^2 = 0.93 for the modeling set and R^2 = 0.96 for the test set), and consequently, only the CCR data will be further discussed.

The data obtained for the *Pf* data set largely mirror those obtained for the logBBB set. Based on the average CCR values across all split sizes and QSAR methods the best three algorithms are again the Pareto-based optimization method, the new representativeness function, and the *k*-means clustering ($\bar{CCR}$ = 0.85) while both MaxMin ($\bar{CCR}$ = 0.78) and Hierarchical clustering ($\bar{CCR}$ = 0.79) have poorer performances. Nine out of the 14 best results (including the best result) for this data set were obtained using the representativeness or the Pareto method while the other 5 were obtained with *k*-means clustering. These 14 best results are bolded in Table 10. Somewhat in contrast with the logBBB data, the best results are

dx.doi.org/10.1021/ci400715n | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

more evenly spread across all split sizes. This is also reflected in the almost identical averaged CCR values across the three splits ($\tilde{CCR}_{Split1}$ = 0.82; $\tilde{CCR}_{Split2}$ = 0.82; $\tilde{CCR}_{Split3}$ = 0.83). Interestingly, none of the best results for this data set were obtained with the *k*NN method although as before, the performances of the different QSAR methods across all split sizes and test set selection algorithms were almost identical ($\tilde{CCR}_{J48}$ = 0.81; $\tilde{CCR}_{RF}$ = 0.82; $\tilde{CCR}_{ANN}$ = 0.84; $\tilde{CCR}_{SVM}$ = 0.83; $\tilde{CCR}_{kNN}$ = 0.82).

4. DISCUSSION AND CONCLUSIONS

Representative subsets are useful in many chemoinformatic applications including the design of information-rich compound libraries, the selection of compounds for biological evaluation, and the development of reliable QSAR models. Such subsets can overcome many of the problems typical of diverse subsets, most notably the tendency of the latter to focus on outliers. Treating the selection of a representative subset from within a parent data set as an optimization problem, has the added advantage that it could be combined with additional objectives into a multiobjective optimization problem (MOOP). Yet despite its potential usefulness, representativeness analysis has not been as widely considered and studied as diversity analysis and only few algorithms for the selection of representative subsets have been published.

Here we present a new MC/SA-based algorithm for the selection of representative subsets. In the current implementation we only retain the subset which yields the best value for the representativeness function. However, it is trivial to extend this algorithm to provide additional solutions (for example by retaining a stack with the best *n* solutions). Such an extension may prove useful if the best solution is limited by other factors, for example, if it contains compounds which do not conform to some desired properties or are otherwise too expensive to purchase. Furthermore, the algorithm can ensure that such multiple solutions are either similar to or different from one another (e.g., through the usage of an exclusion radius around each solution). Alternatively, additional objectives could be directly optimized for by using multiobjective optimization approaches. Here we provide a proof of concept for this strategy by combining the representativeness function with the MaxMin function and simultaneously minimize both by means of Pareto ranking. By minimizing rather than maximizing the MaxMin function we further focus the selected subset on the more populated regions of the parent data sets. Indeed we find multiple cases where the Pareto-based optimization outperformed the representativeness function in terms of the subset representativeness matrices considered in this work.

The performances of the new representativeness function, either alone or in combination with MaxMin were evaluated by selecting representative subsets of different sizes from three parent data sets containing either compounds with known biological activities grouped into different indications (CMC, see Supporting Information Table S1) or qualitative logBBB and proliferation of *Plasmodium falciparum* in erythrocytes data (see Supporting Information Tables S2 and S3). The three data sets differ in size, number of descriptors, and in the case of logBBB and *Pf*, by the distribution of "active" and "inactive" compounds (logBBB 53% active and 47% inactive; *Pf*, 36% active and 64% inactive). The selected subsets were evaluated by different measures. First we calculated for each subset the average distance of each compound from the unselected portion of the data set to its closest neighbor in the subset which we term $1NN_{data\ set}$ and the average distance of each compound from the subset to its closest neighbor in the unselected portion of the data set which we term $1NN_{subset}$. In all cases the best and second best $1NN_{data\ set}$ values, independent of subset size, were obtained when using the representativeness function or the Pareto-based optimization. For $1NN_{subset}$, in almost all cases (except two) the best values were obtained when using again the Pareto-based optimization and the representativeness function. As for $1NN_{data\ set}$, this result is independent of split size. While the new representativeness algorithm could be considered as a $1NN_{data\ set}$ optimizer and hence its good performances in this respect are not surprising, $1NN_{subset}$ was not explicitly considered as a selection criteria either in the representativeness or the Pareto-based optimization.

For the CMC database we evaluated the ability of the 200 compounds subsets selected by the different procedures to accurately mirror the distribution of indications across the parent (filtered) data set. We found that the distribution of indications within the subsets selected by the new representativeness algorithm and the Pareto-based optimization are statistically indistinguishable from those in the parent database. This test case also serves to demonstrate the usefulness of the new algorithms when data obtained on a subset are used to infer the properties of the entire data set.

The logBBB and *Pf* data sets pertain to a different scenario where the selected subset is used to evaluate the performances of QSAR models. The favorable properties of the representativeness and Pareto algorithms, as reflected in the $1NN_{data\ set}$ and $1NN_{subset}$ values obtained for the two data sets are in line with criteria previously established by Golbraikh and Tropsha[5] for the selection of appropriate training (i.e., modeling) and test sets for reliable QSAR modeling, namely, (1) closeness of the representative points of the test set to representative points of the training set in the multidimensional descriptor space and (2) closeness of the representative points of the training set to representative points of the test set in the multidimensional descriptor space. With this in mind, we have indirectly evaluated the performances of the new algorithms by treating the selected subsets and the remaining (unselected) data sets as test and training sets, respectively, and subjected them to five classification QSAR modeling procedures (decision tree, random forests, ANN, SVM, and *k*NN). For comparison, we performed similar training and test sets selections using three additional methods, namely, MaxMin, hierarchical clustering, and *k*-means clustering. We assume that more representative test sets would be better predicted by QSAR models.

As the data in Supporting Information Tables S4 and S5 indicate, all classification methods generated internally validated models for the two data sets and for all split sizes (CCR_{logBBB} > 0.65, CCR_{Pf} > 0.81). More revealing however are the data for the test sets (Tables 9 and 10). Overall (i.e., across the two data sets, the five classification algorithms and the three split sizes), the best performances were obtained with the Pareto method ($\tilde{CCR}$ = 0.83) followed by the representativeness function and *k*-means clustering ($\tilde{CCR}$ = 0.80) while hierarchical clustering ($\tilde{CCR}$ = 0.78) and the MaxMin function ($\tilde{CCR}$ = 0.74) led to poorer performances. A more in-depth analysis of the data, focusing on the individual classification QSAR methods and split sizes (five methods × three split sizes × two data sets for a total of 30 results), reveals that Pareto, representativeness, *k*-means clustering, hierarchical clustering, and MaxMin yielded the best CCR values in 18, 5, 13, 4, and 1 cases, respectively

dx.doi.org/10.1021/ci400715n | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(for a total of 41 data points; the deviation from the total number of results (30) arises since in several cases more than a single method yielded the best result). Furthermore, both the representativeness and in particular the Pareto methods tend to outperform *k*-means clustering for the selection of small and medium subsets (splits 1 and 2) where Pareto, representativeness, and *k*-means clustering yielded the best CCR values in 12, 5, and 7 cases, respectively. This situation changes for larger subsets (split 3) where Pareto, representativeness, and *k*-means clustering yielded the best CCR values in 6, 0, and 6 cases, respectively. These observations could be interpreted as follows: When partitioning a nonuniform data set into a small number of clusters, the resulting cluster representatives are unlikely to mirror the distribution of the compounds in the data set. This is because each cluster is expected to contain dissimilar compounds. However, when the number of cluster increases, they become more uniform and consequently compounds selected from them better represent the data set. It follows that representing a data set by a small subset is more difficult than representing it by a larger subset. Thus, we argue that the new representativeness and Pareto algorithms are able to select subsets which are inherently more representative of their corresponding parent data sets than the other methods considered in this work.

In conclusion, we present new algorithms for the direct selection of representative subsets from within a parent data set. When considering the $1NN_{data\ set}$, $1NN_{subset}$, and χ^2 goodness of fit test, these algorithms outperformed other subset selection methods such as MaxMin, hierarchical clustering, and *k*-means clustering. When considering the CCR matrix, these algorithms performed better than MaxMin or hierarchical clustering and depending on the split size either better or slight worse than *k*-means clustering. We expect that these new algorithms will be useful in various chemoinformatic applications.

■ ASSOCIATED CONTENT

Supporting Information

Figure S1−S3 and Tables S1−S5. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: hsenderowitz@gmail.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Access to the CMC database was generously provided by Accelrys Inc. (www.accelrys.com).

■ REFERENCES

(1) Drew, K. L. M.; Baiman, H.; Khwaounjoo, P.; Yu, B.; Reynisson, J. Size estimation of chemical space: how big is it? *J. Pharm. Pharmacol.* **2012**, *64*, 490−495.

(2) Johnson, M. A.; Maggiora, G. M. *Concepts and applications of molecular similarity*; John Wiley & Sons, 1990.

(3) Gillet, V. J. Diversity selection algorithms. *Wiley Interdis. Rev.: Comput. Molec. Sci.* **2011**, *1*, 580−589.

(4) Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.

(5) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Molec. Des.* **2002**, *16*, 357−69.

(6) Jr, J. H. W. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.

(7) Forgy, E. W. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics* **1965**, *21*, 768−769.

(8) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity measures for rational set selection and analysis of combinatorial libraries: the Diverse Property-Derived (DPD) approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599−614.

(9) Pearlman, R.; Smith, K. M. Novel software tools for chemical diversity. *Persp. Drug Discov. Des.* **1998**, *9−11*, 339−353.

(10) Hassan, M.; Bielawski, J.; Hempel, J.; Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Divers.* **1996**, *2*, 64−74.

(11) Agrafiotis, D. K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.

(12) Waldman, M.; Li, H.; Hassan, M. Novel algorithms for the optimization of molecular diversity of combinatorial libraries. *J. Molec. Graph. Modell.* **2000**, *18*, 412−426.

(13) Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green, D. V. S. Combinatorial Library Design Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 375−385.

(14) Gillet, V. J.; Willett, P.; Fleming, P. J.; Green, D. V. S. Designing focused libraries using MoSELECT. *J. Molec. Graph. Modell.* **2002**, *20*, 491−498.

(15) Agrafiotis, D. K. Multiobjective optimization of combinatorial libraries. *J. Comput.-Aided Molec. Des.* **2002**, *16*, 335−56.

(16) Nicolaou, C. A.; Brown, N.; Pattichis, C. S. Molecular optimization using computational multi-objective methods. *Curr. Opin. Drug Discov. Dev.* **2007**, *10*, 316−324.

(17) Nicolaou, C. A.; Brown, N. Multi-objective optimization methods in drug design. *Drug Discov. Today: Technol.* **2013**, *10*, e427−e435.

(18) Clark, R. D.; Kar, J.; Akella, L.; Soltanshahi, F. OptDesign: extending optimizable k-dissimilarity selection to combinatorial library design. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 829−36.

(19) Clark, R. D.; Langton, W. J. Balancing Representativeness Against Diversity using Optimizable K-Dissimilarity and Hierarchical Clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1079−1086.

(20) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361−75.

(21) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.

(22) http://accelrys.com/products/databases/bioactivity/comprehensive-medicinal-chemistry.html (accessed Nov. 26, 2013).

(23) Mitchell, T. M. *Machine Learning*; McGraw-Hill: New York, 1997.

(24) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molec. Inf.* **2010**, *29*, 476−488.

(25) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* **2012**, *52*, 2570−2578.

(26) Zhang, L.; Zhu, H.; Oprea, T.; Golbraikh, A.; Tropsha, A. QSAR Modeling of the Blood−Brain Barrier Permeability for Diverse Organic Compounds. *Pharm. Res.* **2008**, *25*, 1902−1914.

(27) Platts, J. A.; Abraham, M. H.; Zhao, Y. H.; Hersey, A.; Ijaz, L.; Butina, D. Correlation and prediction of a large blood−brain distribution data set—an LFER study. *Eur. J. Med. Chem.* **2001**, *36*, 719−730.

(28) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Dobchev, D. A.; Fara, D. C.; Karelson, M.; Acree, W. E., Jr; Solov'ev, V. P.; Varnek, A. Correlation of blood−brain penetration using structural descriptors. *Bioorg. Med. Chem.* **2006**, *14*, 4888−4917.

(29) Hammann, F.; Suenderhauf, C.; Huwyler, J. A Binary Ant Colony Optimization Classifier for Molecular Activities. *J. Chem. Inf. Model.* **2011**, *51*, 2690−2696.
(30) *Discovery Studio Modeling Environment*, Release 3.5; Accelrys Software Inc.: San Diego, 2013.
(31) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **2009**, *11*, 10−18.
(32) MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, 1967; Vol. *1*: Statistics, pp 281−297.
(33) Quinlan, J. R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81−106.
(34) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(35) Vapnik, V. N. *The nature of statistical learning theory*; Springer-Verlag: New York, 1995; p 188.
(36) Hassoun, M. H. *Fundamentals of Artificial Neural Networks*; MIT Press, 1995; p 537.
(37) Kohavi, R. In *Proceedings of the 14th international joint conference on Artificial intelligence*; Morgan Kaufmann Publishers Inc.: Montreal, Quebec, Canada, 1995; Vol. *2*, pp 1137−1143.