Supervised Self-Organizing Maps in Drug Discovery. 1. Robust Behavior with Overdetermined Data Sets

Yun-De Xiao,[†] Aaron Clauset,[‡] Rebecca Harris,[†] Ersin Bayram,[§] Peter Santago, II,[§] and
Jeffrey D. Schmitt*[,†]

Molecular Design Group, Targacept Inc., 200 East First Street, Suite 300, Winston-Salem,
North Carolina 27101-4165, Department of Computer Science, University of New Mexico, Albuquerque,
New Mexico 87131, and Virginia Tech - Wake Forest University School of Biomedical Engineering and
Sciences, Medical Center Boulevard, Winston-Salem, North Carolina 27157-1022

Received March 11, 2005

The utility of the supervised Kohonen self-organizing map was assessed and compared to several statistical methods used in QSAR analysis. The self-organizing map (SOM) describes a family of nonlinear, topology preserving mapping methods with attributes of both vector quantization and clustering that provides visualization options unavailable with other nonlinear methods. In contrast to most chemometric methods, the supervised SOM (sSOM) is shown to be relatively insensitive to noise and feature redundancy. Additionally, sSOMs can make use of descriptors having only nominal linear correlation with the target property. Results herein are contrasted to partial least squares, stepwise multiple linear regression, the genetic functional algorithm, and genetic partial least squares, collectively referred to throughout as the “standard methods”. The *k*-nearest neighbor (*k*NN) classification method was also performed to provide a direct comparison with a different classification method. The widely studied dihydrofolate reductase (DHFR) inhibition data set of Hansch and Silipo is used to evaluate the ability of sSOMs to classify unknowns as a function of increasing class resolution. The contribution of the sSOM neighborhood kernel to its predictive ability is assessed in two experiments: (1) training with the *k*-means clustering limit, where the neighborhood radius is zero throughout the training regimen, and (2) training the sSOM until the neighborhood radius is reduced to zero. Results demonstrate that sSOMs provide more accurate predictions than standard linear QSAR methods.

INTRODUCTION

Quantitative Structure Activity Relationship (QSAR) models are frequently used in drug discovery to relate calculable molecular descriptors to biological activity[1] via a mathematical function. Typically, these models employ a linear technique such as multiple linear regression (MLR)[2] or partial least squares (PLS).[3,4] The descriptors that comprise QSAR data models can be based on numerous measured or calculated physicochemical properties and can easily number in the thousands. The predictive ability of models can actually decrease as a result of having too many descriptors,[5] as most QSAR methodologies are known to be dimensionality sensitive. To maximize their predictive power and generalizability, an objective selection strategy such as a genetic algorithm (GA)[6] is typically employed that selects only those descriptors considered to be predictive of the target activity. In this manuscript we extend previously reported work on sSOMs,[7] assessing their sensitivity to overdetermined descriptor spaces. Another weakness in many statistical methods used in QSAR, such as PLS, is an underlying assumption that descriptors are independent and that the relationship between descriptors and target property is linear. Several approaches to exploring nonlinear descriptor-target relationships have been developed. For instance, the genetic functional algorithm (GFA) uses a genetic algorithm to build populations of predictive equations, while mutations act on the population to introduce nonlinear basis functions.[8] Machine learning techniques such as Artificial Neural Networks (ANN) have also been widely applied to QSAR data to achieve nonlinear mapping.[9]

SOMs (also known as Kohonen networks) nonlinearly map a high-dimensional metric space into a low-dimensional representation while respecting the underlying data topology to the maximal extent under the Kohonen map learning rules; thus, they have the potential to address the limitations mentioned above and to offer important advantages over existing QSAR methodology. The use of SOMs has been reported in the chemical and QSAR literature.[10−15] For instance, Zupan and Gasteiger designed a scheme for the application of the SOM algorithm for the projection of the molecular surface properties into two-dimensional feature maps,[16] and Polanski and Bek used a similar scheme for the comparison of molecules represented by atomic coordinates and superimposed by moments of inertia.[17] Prior to our original study, sSOMs had, to our knowledge, not been used to predict biological endpoints via classification of unknowns. In this paper, we demonstrate the efficacy of sSOMs as a tool for QSAR, by developing classification models for ligand−receptor binding that are easy to use and interpret

* Corresponding author e-mail: jeff.schmitt@targacept.com.
[†] Targacept Inc.
[‡] University of New Mexico.
[§] Virginia Tech - Wake Forest University School of Biomedical Engineering and Sciences.

10.1021/ci0500839 CCC: $30.25
Published on Web 10/04/2005

and are well suited for predicting target properties. Because QSAR data are typically both redundant and noisy, we use stepwise variable reduction to explore the impact of descriptor redundancy and also examine the effect of noise on the predictive ability of sSOMs. Next, a series of experiments is conducted to elucidate the influence of the SOM neighborhood kernel on classification ability. By setting the initial training radius and then the final training radius to zero, we create two versions of sSOMs that are increasingly like *k*-means, allowing us to demonstrate the importance of the neighborhood size in predictive capability. We purposefully use a broad set of generic descriptors and have not sought to find optimal descriptors for the DHFR data set.

TECHNICAL BACKGROUND

Self-Organizing Maps. The SOM infers a nonlinear statistical relationship from the topology of the input samples and creates an easily visualized image of the relationship on a low-dimensional display. The relationship is constructed by combining vector prototyping with self-organized learning.[18,19] From a neural network perspective, an *n*-dimensional (usually a 2-dimensional) ordered array of neuronal nodes is created, with each node possessing a prototype vector of dimensionality equal to that of the given input space (the descriptor space in the case of QSAR). The self-organization of neuronal nodes is divided into rough learning and fine-tuning phases, which are distinguished by different training parameters. During each learning phase, the SOM algorithm first treats the map as a competitive network, searching for the node that is most similar, according to a given distance metric, to randomly chosen input vectors. The winning node and its nearest neighbors are then updated using the Kohonen rule (described below) so that they move closer to the input vector. This rearranging will bring similar prototype vectors closer while keeping dissimilar ones apart, thereby preserving distance and proximity relationships within the input data. Iterative application of the SOM algorithm over the two learning phases transforms the initial prototype vectors into an approximation of the input distribution. It is this topology-preserving characteristic that allows the SOM to visualize a high-dimensional data set in a way that artificial neural networks cannot.

Although SOMs are often described as neural networks, it is instructive to describe the self-organization process in terms of vector prototyping.[20] The task of the SOM is to model N observations described by D descriptors. The descriptors form a manifold, X, of input vectors $x = [\xi_1, ..., \xi_D]^T$, where $x \in X$ and $X \subset \mathscr{R}^D$. Let C be a set of i prototype vectors $m_i = [\mu_{i1},..., \mu_{iD}]^T$ defined such that $m_i \in \mathscr{R}^D$. SOM training uses these prototype vectors to divide pattern space such that the probability density of the original manifold is preserved (topology preservation or ordering property), while simultaneously clustering the data (clustering property). A geometric interpretation of this process is that the prototype manifold divides pattern space into cojoint Voronoi cells. The practical result is that a new input x' is mapped to the nearest prototype m_c, known as the best matching unit (BMU), according to a distance metric $||\cdot||$ such that

$$||x - m_c|| = \min_i ||x - m_i|| \qquad (1)$$

In this implementation, the input manifold, X, is transformed with the variance-covariance matrix, A, yielding the Mahalanobis distance metric (recall that if A is the identity matrix then one obtains the Euclidian distance):

$$||(X - C)^T A^{-1}(X - C)|| \qquad (2)$$

Kohonen's batch-training algorithm[18] is used in all SOM training described herein. The batch-training algorithm is defined as follows:

(1) Initialize (randomly in our case) all prototype vectors m_i;

(2) Using the distance metric described above, determine the best matching prototype vector m_c for each input vector x; that is, determine which input vectors are associated with each Voronoi cell;

(3) Readjust the Voronoi cell structure by updating the prototype vectors according to the Kohonen rule

$$m_i(t + 1) = m_i(t) + \alpha h(c,i)(x(t) - m_i(t)) \qquad (3)$$

where α is a scalar-valued "learning-rate" that determines how much influence the update rule has. The function *h(c,i)* is the neighborhood kernel, or smoothing function, that defines a set of points around node *i*

$$h(c,i) = \exp\left(-\frac{||r_c - r_i||^2}{2\sigma^2}\right) \qquad (4)$$

where r_c and r_i are the respective locations of prototype vectors for the best matching unit and the Voronoi cell being updated, and σ is the neighborhood radius. Equation 4 is a Gaussian function that allows neighboring Voronoi cells to have some influence on the outcome of the update and thus sets SOMs apart from *k*-means clustering. Taking other cell centroids (prototype vectors) into account allows elasticity in the surface defined by the prototype vectors. Both the learning rate, $\alpha(t)$, and the neighborhood radius, $\sigma(t)$, are monotonically decreasing functions of time;

(4) Repeat from step 2 until the SOM converges, i.e. until the expectation value of $m_i(t+1)$ equals the expectation value of $m_i(t)$ for all i

$$\forall i{:}E\{h(c,i)(x - m_i)\} = 0 \qquad (5)$$

(For a discussion of SOM convergence, see ref 18.)

Training starts with a large neighborhood function to ensure proper topological ordering of the SOM, and then $h(c,i)$ is decreased over each iteration until the condition in step 4 is met. As an alternative to this typical methodology, in one set of experiments we stop training only when $h(c,i) = 0$. This criterion drives convergence and ensures that the SOM represents as closely as possible the probability density of the input manifold X.

The Relationship of SOMs to *k*-Means Clustering. *k*-means clustering is a nonhierarchical clustering method that seeks a solution that minimizes variability within clusters.[21] It is an iterative method in which n points are classified into k clusters by assigning each data point to the closest (usually by a Euclidean distance measure) centroid.

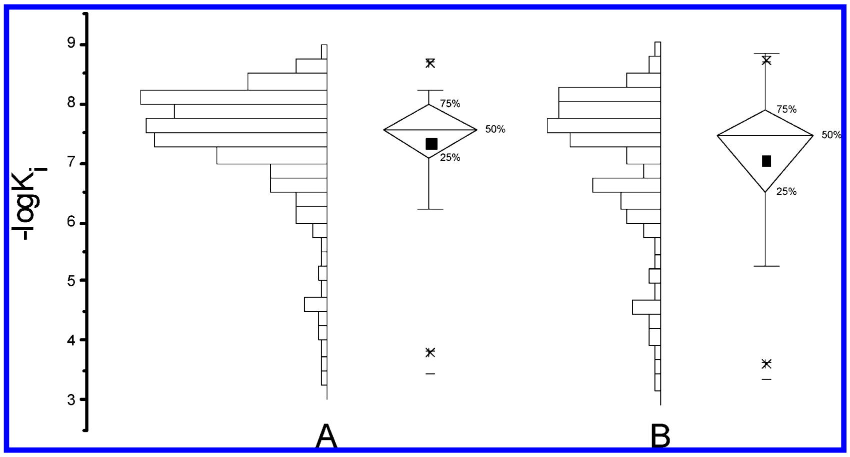

Figure 1. A: the original and B: the selected DHFR data set target value distribution. For the data set, a box chart and histogram show the distribution of target property, with 25th, 50th, and 75th percentile values labeling the box and whiskers are placed at ± SD; (■) data mean; (×) outliers; (−) minimum and maximum values.

The centroid is then reassigned as the average of the cluster members' coordinates. Iterations continue until recomputation of the centroids produces no significant change. Mathematically speaking, the *k*-means algorithm seeks to minimize the total distance of the input vectors, *X*, from *k* cluster centroids

$$E_K = \sum_k \sum_i ||x_i - m_{c(x,k)}||^2 \tag{6}$$

where $m_{c(x,k)}$ is the centroid closest to each $x_i \in X$. Typically, this cost function is minimized by iterative adjustment of the positions of the *k* cluster centroids until changes in E_K cease. Although no explicit cost function exists for SOMs, one can generate a function similar to eq 6 by fixing the SOM neighborhood radius and setting the neighborhood kernel to a Dirac delta function

$$E_K = \sum_k \sum_i h(c,i)||x_i - m_{c(x,k)}||^2 \quad \text{where } h(c,i) = \begin{cases} 1, & \text{if } m_{c(x,k)} \in N_c \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

with N_c being the set of cells within the neighborhood radius. Thus, the approximate SOM cost function closely resembles the function that the *k*-means algorithm seeks to minimize.[22] Kaski explains that the primary difference between these two techniques is that SOMs take the distance of each input vector from all of the prototype vectors (instead of only the distance from the closest one) into account, weighted by the neighborhood kernel. Thus, if the width of the neighborhood kernel radius is zero, the SOM functions as a conventional *k*-means clustering algorithm.

Supervised Self-Organizing Maps. In this manuscript we adopt the definition of the sSOM introduced by Kohonen.[23] Here, the manifold, *X*, of input vectors contains both descriptors, ξ, and class information, β: $x = [\xi_1,...,\xi_D]^T[\beta_1,...,\beta_K]^T$. Thus class information now explicitly influences topological ordering of the map during training. When the map is used for prediction, the β dimension is ignored. β represents a binary column vector containing bioactivity class information. The section on model development below provides information about how sSOMs are implemented in this manuscript.

Table 1. Results of Standard Statistical Methods[a]

	PLS	GFA	GPLS	swMLR
R^2	0.44	0.41	0.66	0.53
r^2	0.51	0.54	0.64	0.20
q^2	0.17	0.46	0.39	0.17
bs-r^2	0.28	0.53	0.61	0.21
PRESS	138.2	94.8	82.6	113.0

[a] R^2 = predictive correlation coefficient for test set; r^2 = correlation coefficient for training set; q^2 = leave-one-out cross validation coefficient for training set; bs-r^2 = bootstrap correlation coefficient for training set; PRESS = predicted sum of squares in the validation procedure for training set.

METHODS

Matlab[24] in combination with the SOM Toolbox[25] was used for all SOM, *k*-means, *k*NN,[26] and PLS programming. The standard methods, PLS, stepwise multiple linear regression (swMLR), GFA, and genetic partial least squares (GPLS) were implemented via Cerius2 Modeling Environment[27] using parameters listed in the Supporting Information. In the case of GFA and GPLS, where more than one model is created, statistics on the most predictive models are reported. All compute-intensive SOM calculations were carried out on a 20-CPU Intel cluster running Linux (version 2.4.17) and Scyld cluster management software (Scyld Beowulf release 2812, Scyld Computing Corp.).

The Data Set. As shown in Figure 1A, the 256 compound Hansch and Silipo dihydrofolate reductase (DHFR) data set[28] is highly skewed; to minimize bias in the modeling procedures, 121 compounds were eliminated so that the data set could be partitioned into evenly populated classification bins. The box-chart and histogram in Figure 1B show the distribution of the resulting 135-compound data set used in this manuscript. Next, to minimize the chances of sampling bias in our experiments, we built 10 working sets, each consisting of 80% training samples and 20% test samples. To construct the working sets, the 121-compound data set was divided into 8 groups according to bioactivity, and then a principal components analysis (3 components) was conducted on the descriptors from each group, leading to a small number of clusters in each group. Twenty percent of each cluster (from the 8 groups) was randomly selected and combined to form a given working set's test samples, with the remaining 80% forming the training samples. This

Table 2. Comparative Results of Standard Statistical Methods[a]

	3-way			4-way			5-way			6-way		
	GFA	GPLS	swMLR	GFA	GPLS	swMLR	GFA	GPLS	swMLR	GFA	GPLS	swMLR
%test	56	67	59	44	56	48	48	52	52	37	48	48
%train	63	69	42	48	54	29	42	46	29	38	46	21
test FP	7	7	11	19	15	30	22	19	26	26	19	33
test FN	0	0	0	0	0	0	0	0	0	11	4	4
train FP	2	1	6	8	6	17	14	12	23	16	15	31
train FN	1	1	3	3	3	7	5	4	9	9	7	13

[a] FP = false positives; FN = false negatives.

Table 3. sSOM Results with Varied Binning Resolution: No Variable Selection[a]

	3-way						4-way					
	best	mean	std	χ^2 (threshold at 0.01)	best rt	mean RT	best	mean	std	χ^2 (threshold at 0.01)	best rt	mean RT
SSOM												
%test	*80.8*	*66.5*	7.1	81.26 (9.21)	46.2	31.2	*55.6*	*42.2*	9.9	28.51 (11.34)	42.3	23.5
%train	90.8	89.6	1.6	1039.00 (9.21)	80.7	80.8	85.2	83.5	1.9	1499.02 (11.34)	78.0	74.9
z-score		3.7						3.2				
E_q	0.71	0.69	0.02		0.72	0.76	0.73	0.70	0.02		0.71	0.77
E_t	0.07	0.02	0.02		0.05	0.13	0.09	0.04	0.03		0.04	0.11
PLS												
%test	*65.4*	*53.0*	7.4				*40.7*	*35.4*	3.9			
%train	51.4	52.6	5.4				45.4	37.8	6.3			
PCs	3	3.2					5	3.2				
*K*NN												
test	40.7	30.2	5.3				33.3	22.0	5.3			
train	31.5	34.1	1.3				22.2	25.0	1.3			

	5-way						6-way					
	best	mean	std	χ^2 (threshold at 0.01)	best rt	mean RT	best	mean	std	χ^2 (threshold at 0.01)	best rt	mean RT
SSOM												
%test	*42.3*	*37.7*	4.4	58.25 (13.28)	40.7	18.4	*55.6*	*42.5*	8.2	108.95 (15.09)	29.6	14.2
%train	81.7	77.3	3.0	1811.16 (13.28)	65.7	72.3	80.6	77.7	2.9	2465.54 (15.09)	71.3	68.0
z-score		2.3						3.8				
E_q	0.73	0.70	0.02		0.71	0.73	0.71	0.69	0.01		0.71	0.76
E_t	0.09	0.04	0.03		0.03	0.08	0.05	0.02	0.01		0.02	0.06
PLS												
%test	*46.2*	*32.9*	9.8				*38.5*	*27.3*	6.0			
%train	29.4	33.5	5.3				28.4	28.9	4.3			
PCs	3	3.2					3	3.2				
*K*NN												
%test	*33.3*	*17.9*	6.2				*25.9*	*14.5*	5.3			
%train	16.7	20.5	1.6				13.9	16.7	1.3			

[a] best = highest performing working set; mean = average performance of the 10 working sets; best rt = best randomization trial; mean RT = average of 20 randomization trials; PCs = number of components. For χ^2 calculations, degrees of freedom = (#bins−1), values are calculated from the pooled test and training samples from the 10 working data sets.

process was repeated 9 times. The following biologically meaningful partitions were applied in our binning process: 3-way = {6.75,7.75}, 4-way = {6.50,7.46,7.90}, 5-way = {6.30,7.21,7.68,7.98}, 6-way = {6.17,6.80,7.47,7.76,8.03}. Our data set contains 177 commonly used 1D and 2D descriptors calculated using QSARIS[29] and Cerius2[27] software.

Model Development. sSOM models were developed for each working set, both with and without descriptor selection, using each of the 10 training sets; preliminary validation was then conducted using the respective test sets. In our implementation of sSOMs the class information [β_1,..., β_K] is a matrix with one vector for each of the K categorical bins. Each element has a value 1 or 0 depending upon whether the training compound is a member of that class or not. To predict the classification of a compound whose class is unknown, its descriptor data [ξ_1, ..., ξ_D] were presented to the trained map resulting in the 'firing' of one SOM node: the BMU. Since the training procedure labels every prototype vector (SOM node), class membership could be thus determined. The working set that yielded the best sSOM model (highest classification accuracy as a percentage of the test set: %test) was used for evaluation of the standard statistical methods, whose results are shown in Tables 1 and 2.

Descriptor Selection. We employ a variation of the forward stepwise regression method originally described by Draper and Smith,[2] using the SOM training set classification accuracy (%train) as the fitness criteria instead of the F-statistic. The descriptor ξ_i possessing the highest linear

Table 4. Confusion Matrix Containing Results from All 10 Working Set sSOM Models[a]

	predicted bin					
actual bin	1	2	3	4	5	6
			3-Way: Train			
1	93.4	2.7	4.0			
2	5.5	88.4	6.1			
3	4.5	8.5	86.9			
			3-Way: Test			
1	71.2	12.3	16.4			
2	21.3	61.8	16.9			
3	10.2	25.5	64.3			
			4-Way: Train			
1	89.6	5.7	2.5	2.2		
2	7.9	82.0	3.0	7.1		
3	2.1	5.2	78.3	14.5		
4	2.4	4.3	9.1	84.3		
			4-Way: Test			
1	51.0	33.3	7.8	7.8		
2	27.4	38.4	17.8	16.4		
3	8.3	20.0	38.3	33.3		
4	6.6	13.2	42.1	38.2		
			5-Way: Train			
1	77.4	14.8	3.5	2.6	1.7	
2	11.5	75.6	3.7	3.2	6.0	
3	1.9	6.5	86.9	3.7	0.9	
4	0.0	3.2	9.1	72.1	15.5	
5	4.3	2.4	8.6	9.5	75.2	
			5-Way: Test			
1	55.0	30.0	7.5	5.0	2.5	
2	32.1	24.5	17.0	7.5	18.9	
3	3.6	21.4	53.6	17.9	3.6	
4	2.0	7.8	25.5	23.5	41.2	
5	5.0	16.7	5.0	46.7	26.7	
			6-Way: Train			
1	76.2	14.8	1.1	4.2	3.7	0.0
2	10.7	81.6	1.0	0.5	2.0	4.1
3	4.7	13.6	69.2	4.1	3.6	4.7
4	2.7	1.1	7.6	75.5	7.1	6.0
5	1.1	3.2	7.4	1.6	83.1	3.7
6	4.3	2.5	0.6	1.8	9.8	81.0
			6-Way: Test			
1	45.2	32.3	3.2	12.9	3.2	3.2
2	11.4	59.1	6.8	0.0	2.3	20.5
3	9.8	25.5	29.4	11.8	17.6	5.9
4	2.8	11.1	19.4	50.0	11.1	5.6
5	0.0	12.2	17.1	7.3	36.6	26.8
6	5.3	7.0	5.3	5.3	45.6	31.6

[a] All figures represent the percentage of molecules in specified category that are correctly binned.

correlation coefficient (r) with the target variable β_k is chosen for the initial round of sSOM training. The remaining descriptors are then chosen at random and added one at a time to the input manifold. The descriptor is kept only if the %train SOM increases. Previously selected descriptors are then tested in consort with the newly added descriptor and are only kept if their removal results in a lower %train. In one set of experiments, the descriptor sets of each sSOM model were reused in subsequent PLS runs to assess the relative performance of the two methodologies with equivalent descriptor sets.

Measures of SOM Quality. We employ two simple measures of SOM quality, quantization error (E_q) and topographic error (E_t).[19] Quantization error, E_q, is defined as the average distance of each input data vector to its BMU:

$$E_q = \frac{1}{N}\sum ||x - m_c||^2 \tag{8}$$

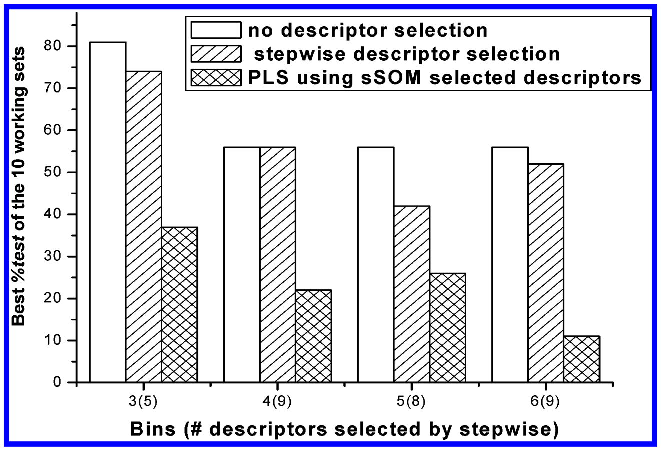

Figure 2. Comparison of sSOM and PLS performance with and without descriptor selection shows robustness of sSOM.

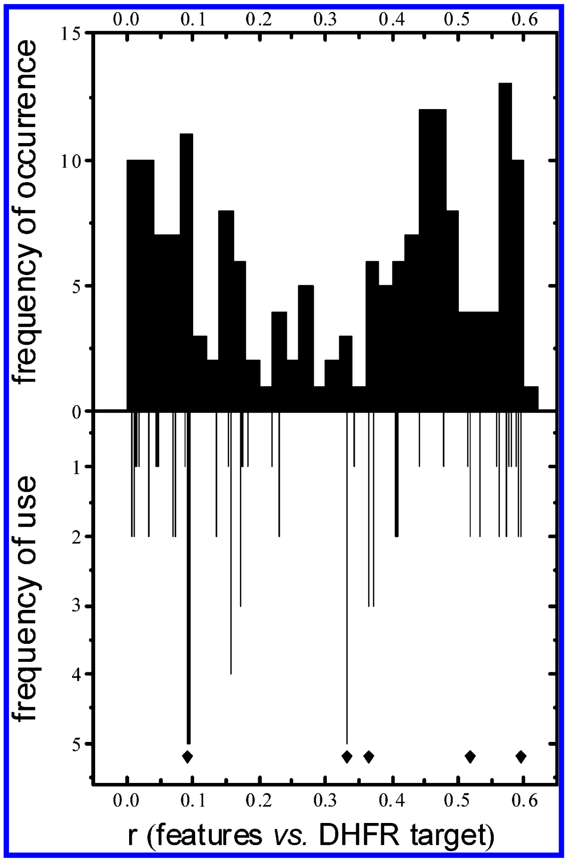

Figure 3. sSOMs make use of descriptors with low linear correlation (r) to target variable. Upper histogram shows distribution of DHFR data set descriptors according to linear correlation to target value. Lower panel shows descriptors used by the 20 DHFR models generated, as described; (◆) r-values of the 5 descriptors used by best DHFR model.

Topographic error, E_t, is defined as the proportion of all input vectors for which the first and second BMUs are not neighbors on the SOM network:

$$E_t = \frac{1}{N}\sum \Omega(x), \text{ where } \Omega(x) = \begin{cases} 1 & \text{if the first and second BMUs are adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

In this manuscript, E_t is included simply to demonstrate that the trained SOM models do not significantly break down

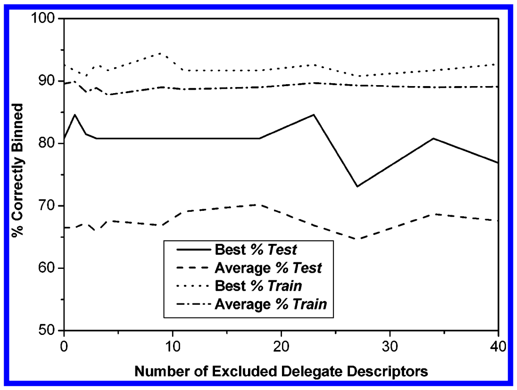

Figure 4. Small effect of removing delegate descriptors from feature manifold of sSOM demonstrates redundancy in descriptor set.

topographic ordering of the input manifold. The use of E_t may be used as a quantitative measure only when the two SOMs have the same number of prototype vectors; thus, absolute values are not comparable for maps with different numbers of prototype vectors. Model assessment by E_q and E_t are provided in Table 3.

Measures of sSOM Training Quality and Susceptibility to Chance Correlation. To assess the ability of sSOMs to classify unknowns relative to the battery of standard methods, the following statistics are provided for all models: percent correct classification (binning) of target property β_κ for training (%train) and test sets (%test); *z*-score, and χ^2 (chi-squared) test. The *z*-score is defined as $(\overline{\%\text{test}} - \mu)/\sigma$, where $\overline{\%\text{test}}$ is the average %test of the 10 models generated from the 10 working sets, μ = (number of bins)$^{-1}$, and σ is the standard deviation calculated from 20 randomization trials. The χ^2 statistic was calculated on the classification results of the 10 models, and the null hypothesis, that sSOM classification results come from a random decision, was tested at a 99% confidence level. We note that because the training and test data set were partitioned into evenly populated categorical bins, there was no probabilistic bias in random classification. As a final validation step, randomization tests were conducted using the methodology described in the Cerius2 3.0 QSAR+ user guide to ascertain the susceptibility of sSOMs to chance correlation. For randomization tests, the target property class (i.e., $[\beta_1,..., \beta_K]^T$) for the data set was scrambled prior to sSOM training. This procedure was conducted with two of the working sets: the set yielding the highest %test and the set whose results were closest to $\overline{\%\text{test}}$. The randomization procedure was repeated 19 times to create randomization statistics at the 95% confidence level.

RESULTS AND DISCUSSION

Predictive Ability of Supervised SOMs. To assess the relative classification ability of sSOMs, training and test results were compared to five standard methods, all widely used in QSAR modeling: PLS, GFA, GPLS, swMLR, and *k*NN (see Tables 1 and 2). sSOMs consistently outperformed these methods in terms of %train and %test, as shown in Tables 2 and 3. For example, in the DHFR 3-way binning experiment, the best sSOM model yields a %test of 81% compared to 65.4% for PLS and 40.7% for *k*NN (see the italicized parts in Table 3). The mean values of %train and %test (averaged over the 10 working sets: $\overline{\%\text{train}}$ and $\overline{\%\text{test}}$, respectively) yielded similarly encouraging results. It is important to note that these results with sSOM were achieved although the models resulting from the standard methods were trained on higher resolution information than the sSOM models, i.e., actual bioactivity values rather than categorical bin numbers.

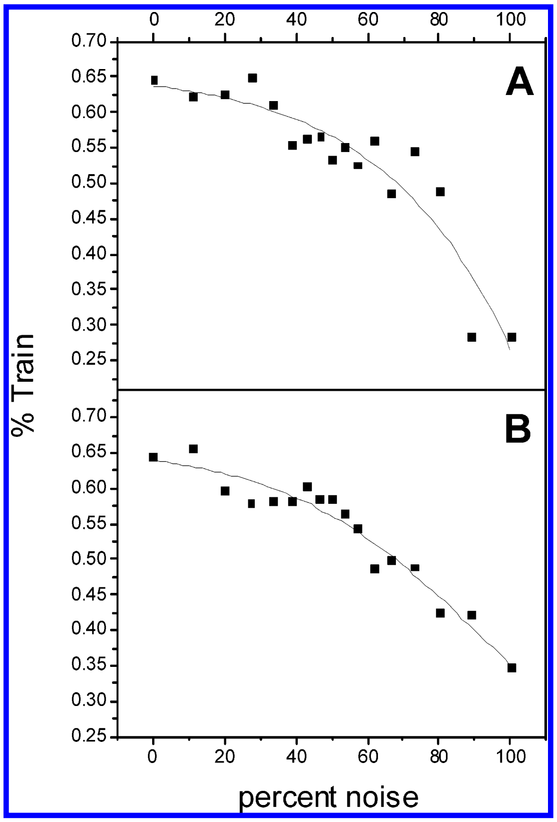

Figure 5. sSOMs provide robust classification under conditions of increasing noise. A: (■) random noise. B: (■) shuffling noise. Lines represent best-fit using a Boltzmann exponential.

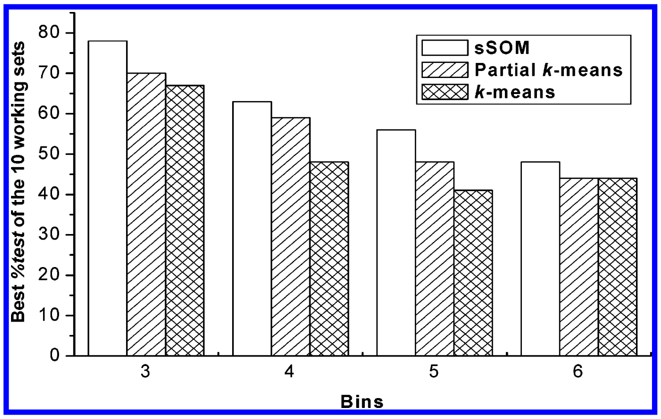

Figure 6. Assessment of the neighborhood kernel on sSOM performance shows that predictive power decreases as the *k*-means limit is approached.

We note that there are shortcomings in assessing QSAR model quality on the basis of %test without further external validation and address this in a forthcoming companion manuscript.

Examination of the data set at finer degrees of binning provided the encouraging result that sSOM performance degrades more slowly than the statistical expectation (see Table 3). The *z*-score (one-tail hypothesis testing) values for

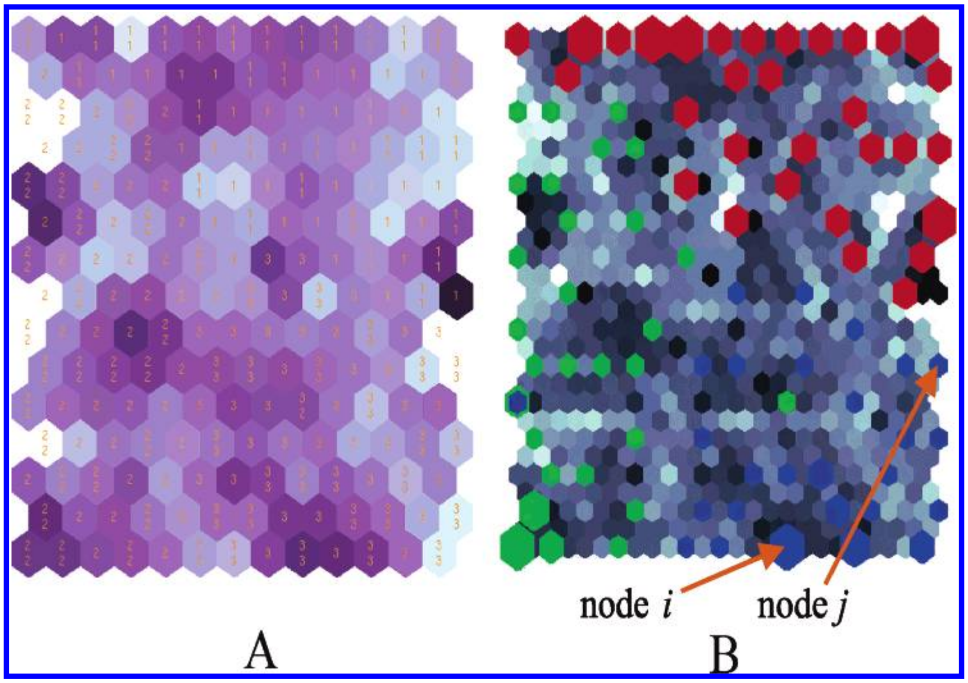

Figure 7. Matrix representations of best DHFR sSOM (3-way binning). A: Distance (D-) matrix representation describing the median distance from one node to its neighbors. B: Ultisch's U-matrix in which each hexagon represents the pairwise distance between nodes. Red, green, and blue hexagons represent BMUs, where red = bin 1, green = bin 2, and blue = bin 3, the size is proportional to the number of input vectors for which that hexagon is the BMU. In both matrix representations, brightness is proportional to distance.

the DHFR 3-way and 6-way models are nearly identical, but the value for 5-way is significantly different, signaling that the numbers of categorical bins as well as bin cutoffs may be important considerations in generation of optimal sSOM classification models. This behavior leads us to surmise that for a sSOM (or any other QSAR classification model) to perform optimally, one must find the natural partitioning of the input manifold. To evaluate the statistical significance of models at varying degrees of binning resolution, *z*-scores were calculated and compared to those derived for random data sets generated by randomly permuting target variables $\{\beta_\kappa\}$. The *z*-score values shown in Table 3 indicate that sSOM models are statistically better than those obtained for random data sets at a 95% significance level. Additionally, the χ^2 statistics reported in Table 3 indicate that the null hypothesis (i.e., that sSOM classification is not better than random classification) can be rejected at a 99% confidence level on both training and test results for all levels of classification resolution. Such a high level of confidence demonstrates the value of sSOMs in QSAR modeling. Finally, a confusion matrix based upon the aggregate classification data from all 10 working set models (Table 4) demonstrates a robust classification. In most cases, classification errors are localized to neighboring categorical bins.

Supervised SOMs with Descriptor Selection. Data sets with excessive noise or redundancy of information are known to be problematic for most chemometric methods because of the curse of dimensionality. Therefore, descriptor selection is commonly coupled with the statistical procedure to surmount this issue. We combined a forward stepwise descriptor selection algorithm (described above) with sSOMs to explore the possibility that they would perform better with fewer dimensions and less noise. An additional motivation for adding descriptor selection was the concern that a sSOM with a large number of prototype vector dimensions might be difficult to interpret.

The results of our experiments with stepwise sSOM models (Figure 2) indicate that this form of descriptor selection provides no improvement in %train or %test. However, the number of descriptors used in the resulting models is dramatically lower in the resulting models (≈12-fold reduction), which may simplify interpretation of the resulting sSOM. The descriptors selected by sSOM training of each working model were used to train PLS models. Poor performance among the PLS models (see Figure 2 as well as Supporting Information for further details) indicates that PLS is unable to relate the target activity to (the nonlinear) descriptors chosen in these sSOM models.

sSOM Robustness in the Presence of Non-Orthogonal Descriptors and Noise. To investigate the redundancy present in our data sets and to inspect the capability of sSOMs to accurately classify unknowns in the presence of redundant information, we designed the following experiments. We generated and analyzed 40 stepwise sSOM models (derived by random partitioning of the data set into 10 training and test sets using 3-, 4-, 5-, and 6-way binning with each) and identified the 40 most frequently selected, or "delegate descriptors", among all 40 models. Since sSOMs are nonlinear, individual delegate descriptors do not necessarily possess a high degree of linear correlation with the targeted biological activities (see Figure 3). We excluded the delegate descriptors from the original feature space one at a time and built 3-way sSOM models for each increasingly reduced data set. The results are shown in Figure 4. There is no indication that predictive power, estimated by %test, decreases with the elimination of delegate descriptors. Even after excluding all 40 delegate descriptors, representing almost a quarter of the original descriptor space, the average %test is around 67%. These results indicate a high degree of redundant information in the data set. Furthermore, three-way sSOM models obtained using only a selected subset of the 40 delegate descriptors show an average %test of 65,

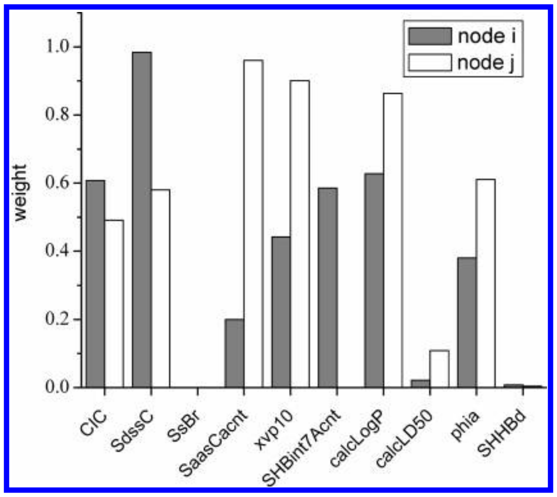

Figure 8. The distinct contribution of descriptors to different subclasses of the most active (bin 3) DHFR inhibitors. Selected nodes *i* and *j* are shown in Figure 7. Key: CIC − complementary information content; SdssC, SsBr SaasCacnt, SHBint7Acnt SHHBd, xvp10 and phia − QSARIS[29] E-state, connective valence and flexibility indices; calcLogP and calcLD50 − calculated LogP and LD 50.

demonstrating that most of the feature information present in the original data set is contained in these 40 selected descriptors—further evidence of extensive redundancy in the data set. These data demonstrate that sSOMs handle feature redundancy quite well.

In addition to exploring feature redundancy, we assessed sSOM's sensitivity to noise. We used the delegate feature set described above to study the performance of sSOMs under conditions of increasing noise in the feature manifold. Two types of noise were added to the delegate feature set in separate experiments: Gaussian noise and "shuffling", or randomly permuted, noise. For Gaussian noise, we created additional descriptors containing a random distribution of numbers normalized to unit variance and following a Gaussian distribution. Figure 5A shows the effect of adding Gaussian noise to the data set. We found the sSOM performance to be very robust. Only when the noise is higher than 80% (40 real descriptors, 160 noise descriptors), do we notice a dramatic decrease in predictive ability. In an effort to further understand susceptibility to noise, we then diluted the descriptor manifold with noise descriptors that possessed variance and distribution identical to the real data set, i.e. shuffling noise. To generate this type of noise, groups of five descriptors were selected randomly from the whole data set, and the feature vectors were randomized with respect to observations. As shown in Figure 5B, if the noise is lower than 60% (40 real descriptors, 60 noise descriptors), the sSOM models still maintain their high level of predictivity.

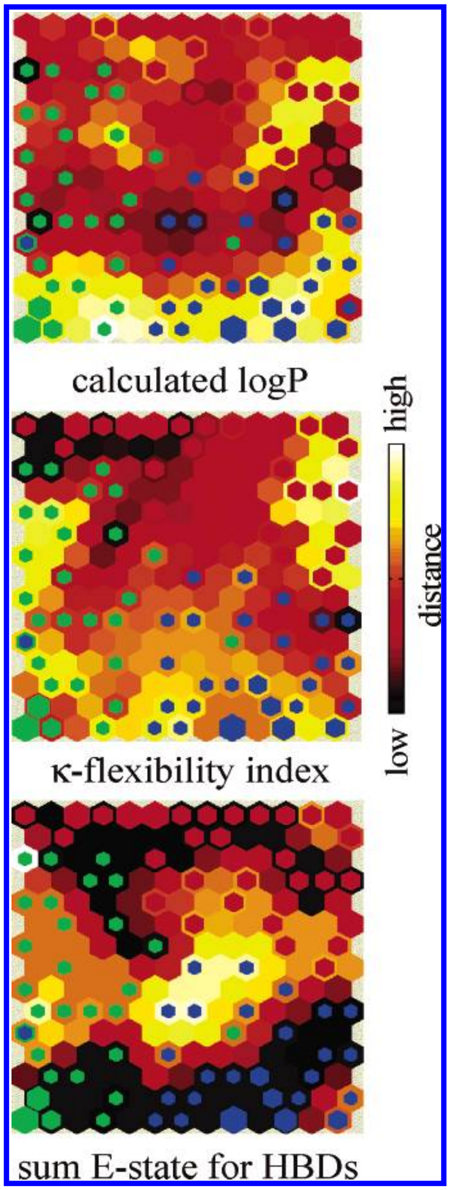

Figure 9. Contribution of three physically meaningful descriptors to clustering in the best DHFR sSOM model. Separate feature plane representations of the contribution of calculated logP, the κ-flexibility index, and the sum of electrotopological indices for hydrogen-bond donors.

Supervised *k*-Means Classification. To evaluate the influence of the SOM neighborhood kernel, we first implemented a *k*-means like classification by performing sSOM training, setting the initial training radius to zero. Because classification information was utilized throughout the training procedure, we call this "supervised *k*-means" classification. We also performed what we refer to as "partial *k*-means" by stopping training only when the neighborhood radius reduces to zero in the fine-tuning phase. Note that the other experiments described above use the typical SOM convergence criteria defined by stationary state: $\partial\mu_i/\partial t = 0$. Figure 6 shows the classification performance of sSOMs in comparison to "supervised *k*-means" and "partial *k*-means" using the descriptors selected by the best models of 3-, 4-, 5-, and 6-way binning plus the delegate descriptors described above. Clearly, the predictive power decreases as one gets nearer to the *k*-means limit. The loss of neighborhood information that occurs when the neighborhood radius is set to 0 forces a sparse and "stiff" classification in which each prototype vector must go to the centroid of the input samples of its class.

SOM Visualization. SOMs provide unique model and data visualization capabilities. By mapping an ordering of the prototype vectors to a much lower dimensional output space, sSOMs preserve the topological relationship in a manner otherwise impossible to visually assess. Figure 7A,B shows the U-matrix and D-matrix for stepwise sSOMs utilizing 3-way binning (U- and D-matrix definitions are in the figure legend). Observing the colors of the U-matrix and

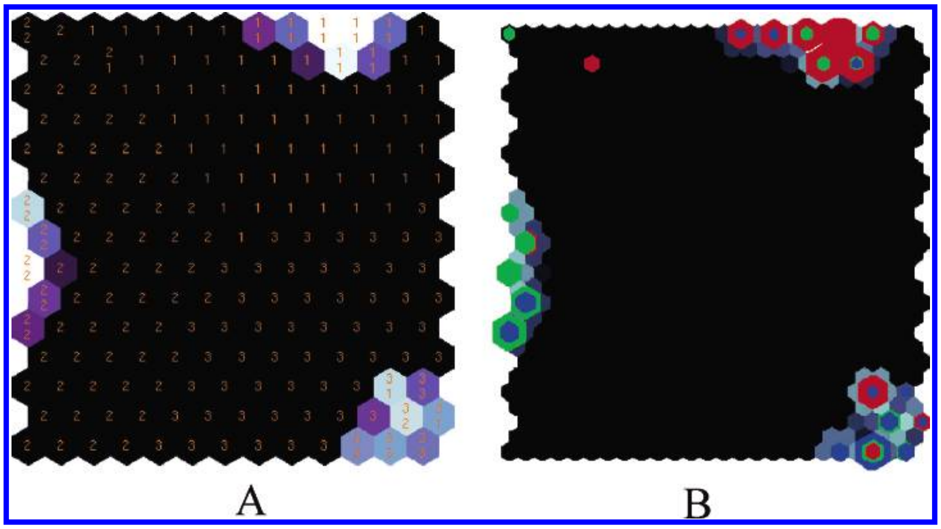

Figure 10. A: D-matrix and B: U-matrix for *k*-means limit (supervised *k*-means), showing a sparse and stiff distribution of clusters.

the assigned numbers on the D-matrix, we immediately conclude that the top right block forms the inactive categorical bin (class 1), the bottom right forms the most active categorical bin (class 3), and the left portion forms the categorical bin of intermediate activity (class 2). Also obvious for each block is that the compounds are distributed among several centers instead of only one center because SOMs can deconvolve complex target-descriptor relationships, such as in cases where more than one combination of determinants lead to a given classification. In the example provided, each categorical bin consists of several subcategories, and there is a different distribution of descriptors to satisfy each subcategory in the same categorical bin. For instance, the two neurons around compounds 173 (Figure 7B, lower center) and 250 (Figure 7B, center left) in the most active category field have different feature weights, as shown in Figure 8. Examining descriptors SaasCacnt (substituted aromatic C E-state index) and SHBint7Acnt (internal H-bonds E-state index), it is clear that compound activity can be driven by the value of certain descriptors and that activity increases or decreases according to the value of those descriptors, depending on the subcategory. The ability to visualize the relationship of descriptor values to target property, such as bioactivity, offers an important advantage for lead optimization and elucidating the determinants of molecular recognition. It is important to emphasize that sSOMs can map nonlinear behavior. So while they indicate which descriptors may influence the target property, we are not necessarily informed as to how the associated characteristics need to be modified to achieve a desired change. Figure 9 shows three topologic/clustering maps that demonstrate the contribution of a selected descriptor to sSOM clustering. These “descriptor planes” demonstrate that no linear correlation likely exists among selected descriptors and that they contribute differently to the data set partitioning. Figure 10 displays the D-matrix and U-matrix representations from the execution of “supervised *k*-means”. Contrast the wide distribution across the sSOM U-matrix in Figure 7b to the supervised *k*-means U-matrix that shows a distribution of clusters that is sparse and stiff, implying the loss of vital neighboring cell information. Any compound that falls outside one of these clusters will not be properly classified.

CONCLUSIONS

In this study, we have demonstrated that sSOMs have the ability to produce overall more accurate predictions than traditional linear QSAR architectures such as PLS. Experiments conducted with various binning schemes show that it is possible to use sSOMs to reveal the natural distribution and topology of data sets. Preliminary investigation of a simple forward stepwise feature selection sSOM was compared to a stand-alone sSOM implementation. Results show no significant statistical difference and indicate that the sSOM is robust (i.e. it is insensitive to information redundancy and noise), with the ability to identify relevant information in the input manifold. Because the classifier function (such as the decision function in Fisher discriminant analysis and the distance function in *k*-nearest-neighbor classification) of many other techniques is influenced by the entire data set, the sSOM is an attractive technique for analysis on high-dimensional chemical spaces and can be likened to an aggregate of numerous “local” PLS models. Another advantage of the sSOM is its visualization capability, allowing the user to visualize the clustering of the entire feature set as well as that of specific descriptors, thereby enhancing understanding of the input data structure and the role or each feature in the clustering process.

The utility of the sSOM methodology is clearly demonstrated here. In an upcoming manuscript, we address a number of enhancements that will extend the range of application and predictive power of the technique. Chief among these is a further simplification of the final model by incorporating an efficient descriptor selection system, such as hill climbing, simulated annealing, or genetic algorithms. Simpler models will allow easier extraction of chemically relevant information and will be more readily interpretable, thereby helping to elucidate the process of drug lead discovery and optimization. Simpler models, because they require the calculation of fewer descriptors, will also make the time-consuming screening process of large libraries more efficient. Finally, in the companion paper we demonstrate the use of sSOMs in a variety of data sets using a more robust modeling protocol including the use of validation sets.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 0139787. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Supporting Information Available: Parameters used for GFA, GPLS, and SR (Table A), measures of sSOM quality with varied binning resolution: no variable selection (Table B), sSOM and PLS results with varied binning resolution: stepwise variable selection (Table C), effect of removing delegate features from feature manifold on sSOM training: 3-way (Table D), and assessment of the neighborhood kernel on sSOMs: DHFR data set (Table E). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hansch, C. A Quantitative Approach to Biochemical Structure−Activity Relationships. *Acc. Chem. Res*. **1969**, *2*, 232−39.

(2) Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons: New York, 1966; p 407.

(3) Lindberg, W.; Persson, J.; Wold, S. *Anal. Chem.* **1983**, *55*, 643.

(4) Geladi, P.; Kowalski, B. R. Partial Least Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(5) Bellman, R. *Adaptive Control Processes: A Guided Tour*; Princeton University Press: Princeton, NJ, 1961.

(6) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom*. **1992**, *6*, 267−281.

(7) Bayram, E.; Santago, P., II; Harris, R.; Xiao, Y.; Clauset, A. J.; Schmitt, J. D. Genetic algorithms and self-organizing maps: a powerful combination for modeling complex QSAR and QSPR problems. *J. Comput.-Aided Mol. Des.* **2004**, 483−493.

(8) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 4−866.

(9) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−527.

(10) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Sadowski, J.; Tackentrup, A.; Wagener, M. *The Use of Self-Organizing Neural Networks in Drug Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer Academic: Boston, 1998; pp 273−99.

(11) Kovalishyn, V. V.; Tetko, I. V.; Luik, A. I.; Ivakhnenko, A. G.; Livingstone, D. J. Application of Self-Organizing Neural Networks with Active Neurons for QSAR Studies. *Proceedings of the European Symposium on Quantitative Structure−Activity Relationships: Molecular Modeling and Prediction of Bioactivity*; 2000; Vol. 12, pp 444−45.

(12) Larry, L.; Curt, B. The use of 2D, 3D, TAE and wavelet coefficient descriptors (WCDs) for generatingself-organizing Kohonen maps for QSAR, QSPR and ADME Analyses; ACS National Meeting, Washington, D.C., August 2000.

(13) Rose, V. S.; Croall, I. F.; MacFie, H. J. H. An Application of Unsupervised Neural Network Methodology Kohonen Topology-Preserving Mapping to QSAR Analysis. *Quant. Struct.-Act. Relat.* **1991**, *10* (1), 6−15.

(14) Polanski, J. Self-organizing neural network for modeling 3D QSAR of colchicinoids. *Acta Biochim. Pol.* **2000**, *47* (1), 37−45.

(15) Espinosa, G.; Arenas, A.; Giralt, F. An integrated SOM-fuzzy ARTMAP neural system for the evaluation of toxicity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 343−59.

(16) Zupan, J.; Gasteiger, J. *Neural Network and Drug Design for Chemists*, 2nd ed.; VCH: Weinheim, 1999.

(17) Polanski, J.; Bak, A. Modeling Steric and Electronic Effects in 3D- and 4D-QSAR Schemes: Predicting Benzoic pK_a Values and Steroid CBG Binding Affinities, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2081−2092.

(18) Kohonen, T. The self-organizing map. *Neurocomputing* **1998**, *21* (1−3), 1−6.

(19) Kohonen, T. *Self-Organizing Maps*, 3rd ed.; Springer: Berlin, 2001.

(20) Van der Putten, P. Utilizing the Topology Preserving Property of Self-Organizing Maps for Classification. M.Sc. Thesis, Cognitive Artificial Intelligence, Utrecht University, NL, 1996.

(21) Anderberg, M. R. *Cluster Analysis for Applications*; Academic Press: New York, 1973.

(22) Kaski, S. Data exploration using self-organizing maps. *Acta Polytech. Scand., Math., Comput. Manage. Eng. Ser. No. 82* **1997**, 57.

(23) Kohonen, T.; Makisara, K.; Saramaki, T. *Phonotopic Maps − Insightful Representation of Phonological Features for Speech Recognition. Proceeding of the IEEE Seventh International Conference on Pattern Recognition*; 1984; pp 182−185.

(24) Matlab 6.5; The Mathworks, Inc.: 3 Apple Hill Dr., Natick, MA.

(25) http://www.cis.hut.fi/projects/somtoolbox/.

(26) Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, 13:21−27.

(27) Accelrys Inc., Cerius^2 Modeling Environment, Release 4.9, Accelrys Inc.: San Diego, 2003.

(28) Silipo, C.; Hansch, C. Correlation Analysis. Its Application to the Structure−Activity Relationship of Triazines Inhibiting Dihydrofolate Reductase. *J. Am. Chem. Soc.* **1975**, *97*, 6849−6861.

(29) QSARIS 1.1; MDL Information Systems Inc.

CI0500839