

# Three-Dimensional Quantitative Structure–Activity Relationships from Molecular Similarity Matrices and Genetic Neural Networks. 1. Method and Validations

Sung-Sau So<sup>\*,†</sup> and Martin Karplus<sup>\*,†,‡</sup>

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, and Laboratoire de Chimie Biophysique, Institut le Bel, Université Louis Pasteur, 4 rue Blaise Pascal, 67000 Strasbourg, France

Received July 28, 1997<sup>⊗</sup>

The utility of genetic neural network (GNN) to obtain quantitative structure–activity relationships (QSAR) from molecular similarity matrices is described. In this application, the corticosteroid-binding globulin (CBG) binding affinity of the well-known steroid data set is examined. Excellent predictivity can be obtained through the use of either electrostatic or shape properties alone. Statistical validation using a standard randomization test indicates that the results are not due to chance correlations. Application of GNN on the combined electrostatic and shape matrix produces a six-descriptor model with a cross-validated  $r^2$  value of 0.94. The model is superior to those obtained from partial least-squares and genetic regressions, and it also compares favorably with the results for the same data set from other established 3D QSAR methods. The theoretical basis for the use of molecular similarity in QSAR is discussed.

## I. Introduction

The use of three-dimensional (3D) molecular fields in quantitative structure–activity relationships (QSAR) has usually been based on the construction of a 3D grid containing known drug molecules, whose interactions with certain probes are evaluated at regularly spaced points on the grid. Most commonly the electrostatic and steric attributes of the known ligands and, by extrapolation, the putative receptor environments are investigated, though other properties can be studied as well. Both the GRID program<sup>1</sup> and the comparative molecular field analysis (CoMFA) package<sup>2</sup> have been used for this purpose. By use of multivariate partial least-squares (PLS) method,<sup>3</sup> QSARs have been derived from the field values, which replace the standard descriptors of 2D QSAR. A growing number of applications based on this method has been reported.<sup>4,5</sup> The approach is the leading 3D QSAR method in the field of drug design.

An alternative way to utilize the 3D molecular fields is to introduce the concept of molecular similarity.<sup>6–8</sup> The molecular similarity concept was first proposed by Carbó,<sup>9</sup> who defined the similarity between two molecules in terms of their electron density distributions (eq 1), where  $P_A$  and  $P_B$  indicate the property of interest for molecule A and B, respectively; in eq 1,  $P_A$  and  $P_B$  represent electron density distributions. Hodgkin and Richards proposed an alternative index, the so-called Hodgkin index,<sup>10</sup> that is more appropriate for properties such as electrostatic potentials (eq 2).<sup>11</sup> Recently new electrostatic similarity indices, the linear and exponential indices, have been suggested (eqs 3 and 4).<sup>11</sup> In the original formulation of these indices, the property ( $P_A$  and  $P_B$  in the equations) of interest was the electron density, but it has been extended to electrostatic potentials and electric fields. For comparison of the shape of two molecules, the Meyer formula has been used (eq 5).<sup>12</sup> This formula is a modified form of the Carbó index;

it is the quotient of the number of grid points that are inside the common volume of the two molecules ( $U_{AB}$ ) and the geometric mean of the number of grid points inside their individual molecular volumes ( $T_A$  and  $T_B$ ). Although it is possible, of course, to define molecular similarity in the context of other global, topological, or even substituent-based parameters,<sup>8</sup> the term “molecular similarity” used in this and the companion study<sup>13</sup> refers to comparisons based on spatial fields, as defined by eqs 1–5.

$$C_{AB} = \frac{\sum P_A P_B}{\sqrt{(\sum P_A^2)(\sum P_B^2)}} \quad (1)$$

$$H_{AB} = \frac{2 \sum P_A P_B}{\sum P_A^2 + \sum P_B^2} \quad (2)$$

$$L_{AB} = \frac{1}{N_{\text{grid points}}} \sum \left( 1 - \frac{|P_A - P_B|}{\max(|P_A|, |P_B|)} \right) \quad (3)$$

$$E_{AB} = \frac{1}{N_{\text{grid points}}} \sum \exp \left( - \frac{|P_A - P_B|}{\max(|P_A|, |P_B|)} \right) \quad (4)$$

$$S_{AB} = \frac{U_{AB}}{\sqrt{T_A T_B}} \quad (5)$$

To build a similarity matrix (SM), each compound in the data set is compared to all the others on the basis of the values of the chosen property at the grid points.<sup>14</sup> The use of such matrices is an efficient means of dimensionality reduction of the raw data. For example, a typical CoMFA analysis may use, depending upon the size of the molecules and the resolution of the grid, on the order of  $10^3$  grid points. Performing PLS on such a large data matrix tends to be slow.<sup>15</sup> The use of molecular similarity compresses the large raw data matrix into a very compact SM, whose dimension depends only on the number of molecules.<sup>7,8</sup> PLS can be applied to this SM to derive QSARs. Several SM/PLS studies have been reported, and some promising results were obtained for standard data sets.<sup>14,16–19</sup>

<sup>†</sup> Harvard University.

<sup>‡</sup> Université Louis Pasteur.

<sup>⊗</sup> Abstract published in *Advance ACS Abstracts*, December 1, 1997.

It has been shown that QSARs based on a selection of a few key variables can be superior than that obtained by performing PLS on all of the variables.<sup>8,20</sup> Encouraged by the success of recent applications of genetic neural networks (GNN) to obtaining QSARs from conventional descriptors,<sup>21,22</sup> we have applied the GNN methodology to the analysis of SM. It appeared likely that the GNN can lead to better optimized QSARs than the traditional linear PLS regression. This may be true because the genetic algorithm can select the most relevant variables and the neural network offers a model-free nonlinear mapping capability to optimize their use.

In this study we propose a QSAR method that analyzes a SM with a GNN. The method is tested and validated with the well-known corticosteroid-binding globulin (CBG) steroid data set. Standard substituent-based analysis is inappropriate for this data set because structural modifications occur at many positions, and furthermore the variety of substitutions at a given position is often limited. This inadequacy, in addition to a relatively uncomplicated molecular alignment, makes a 3D-based method particularly useful in the structure-activity analysis of the data. In fact, since the original CoMFA study,<sup>2</sup> this data set has become a standard in testing novel 3D QSAR methods.<sup>14,16,23-30</sup> The influence of the various SM/GNN parameters on the quality of the QSAR models is investigated. The statistical significance of the final QSARs is evaluated. A comparison of the SM/GNN results with those obtained from other approaches is made.

Section II describes the method. The results are presented and discussed in section III. Section IV outlines the conclusions.

## II. Method

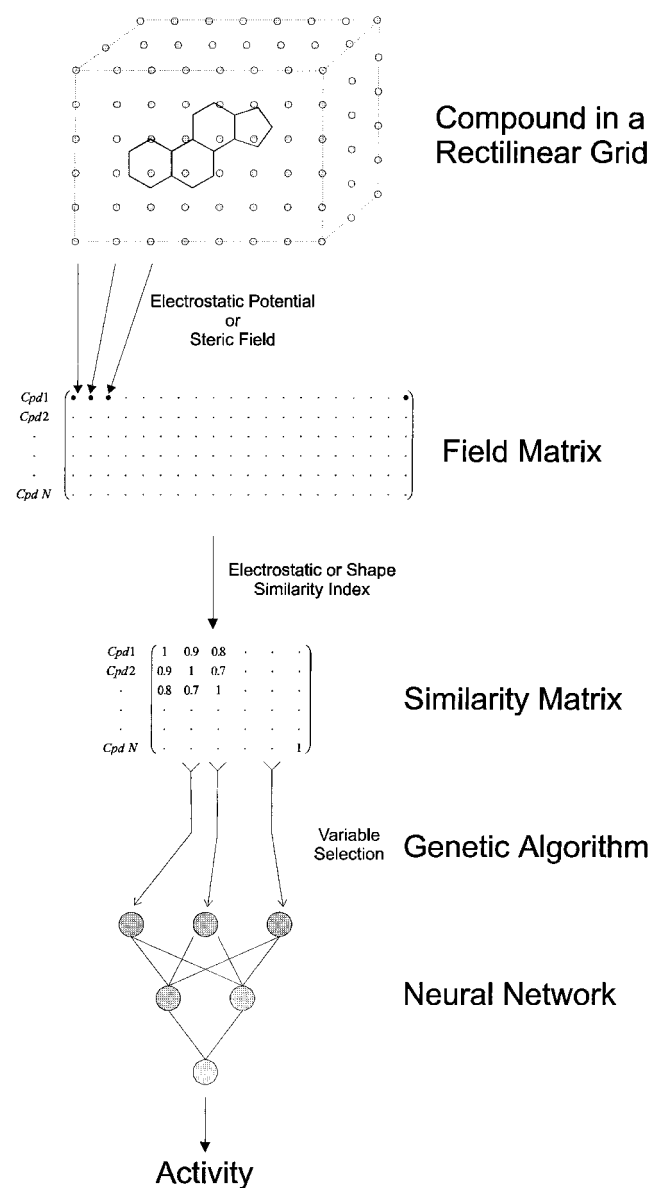
Scheme 1 is a schematic diagram showing the various stages of the SM/GNN QSAR development. The specific details are described in the following subsections.

**Model Building.** The binding affinity ( $\log K$ ) of the 31 steroids (Chart 1) with CBG is shown in Table 1. The coordinates were obtained via anonymous ftp from the Gasteiger group,<sup>31</sup> which recently reported a QSAR study based on neural network and autocorrelation vectors on molecular surface properties as descriptors for this data set.<sup>27</sup> Although their method does not require molecular alignment, the molecular coordinates released by them had the steroidal backbones appropriately aligned. Neither realignment of the structures nor modification of the steroid coordinates was necessary, which made possible a direct comparison with their results. Unless otherwise specified, the Mulliken charges were derived from MOPAC6<sup>32</sup> within the Cerius2 modeling environment,<sup>33</sup> using the AM1 Hamiltonians in a single-point energy calculation.

**Molecular Field and Molecular Similarity Calculations.** Drug-receptor interactions are often separated into electrostatic and steric components. In light of this, two different similarity matrices, that based on the electrostatic and that based on the shape,<sup>14</sup> were used in this investigation. An alternative steric similarity matrix based on a van der Waals (vdW) potential was compared with the shape matrix.

Unless otherwise stated the following protocol was used. The electrostatic similarity matrix (ESM) was derived from the Hodgkin formulation (eq 2). The electrostatic potentials at the grid points were calculated using a unit positive charge probe with a vacuum dielectric constant ( $\epsilon = 1.0$ ). Furthermore, electrostatic potentials greater than +5.0 kcal/mol or less than -5.0 kcal/mol were truncated to the cutoff values.<sup>34</sup> A  $20 \times 30 \times 20 \text{ \AA}^3$  rectilinear box was constructed so that it would extend beyond the atomic coordinates of the entire data

**Scheme 1.** Schematic Diagram for the Construction of a SM/GNN QSAR Model (See Method)



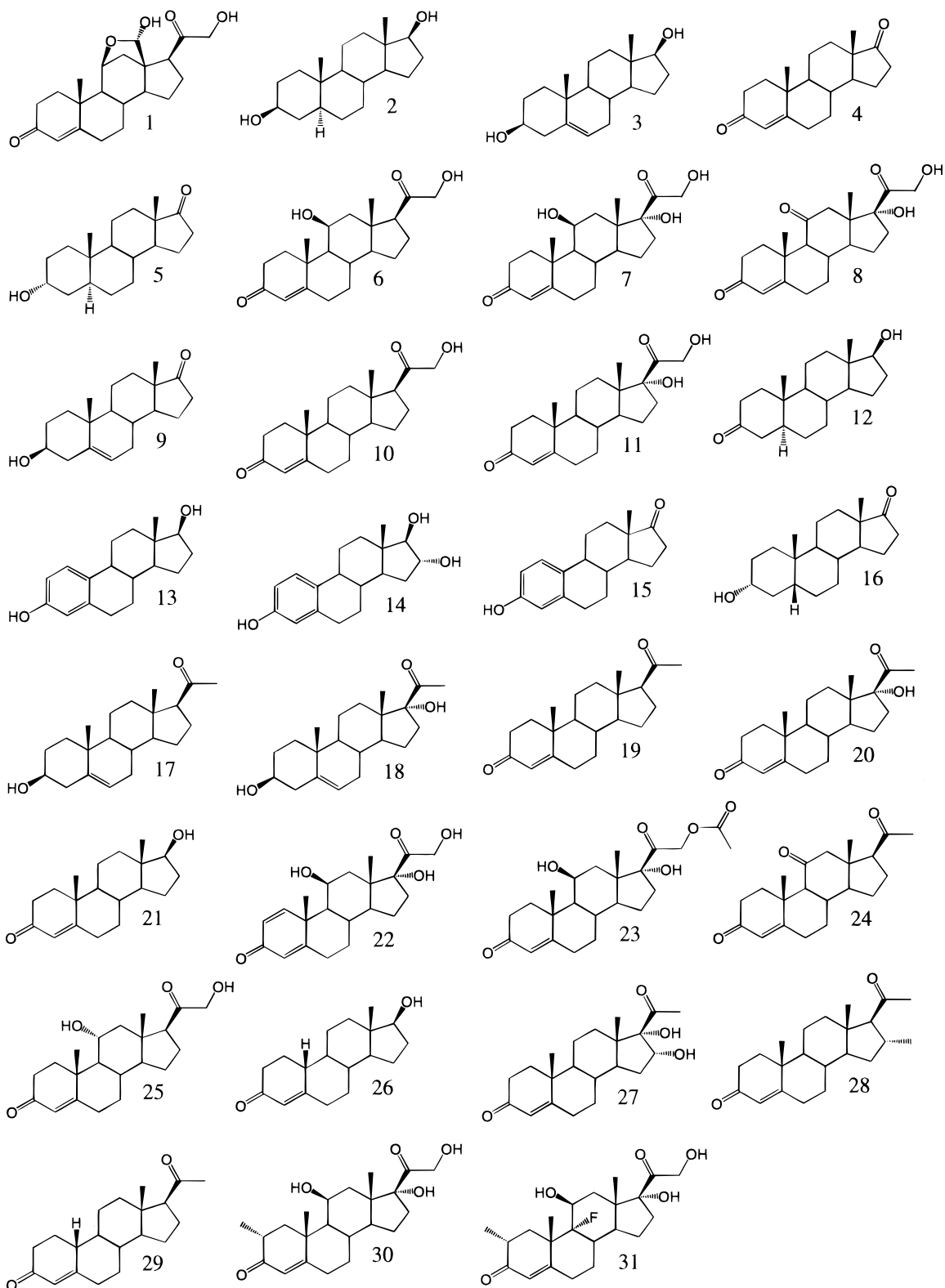
set by at least  $6 \text{ \AA}$  on each side. A  $2.0\text{-\AA}$  grid spacing, a value recommended by Good et al. in a previous SM/PLS study,<sup>16</sup> was employed. To avoid singularities for the electrostatic potential at grid points near the atomic positions, the points within the van der Waals surface<sup>35</sup> of the molecule had their electrostatic potentials set to zero.<sup>11</sup>

The shape field is a binary measure that encodes whether a grid point is within or outside the van der Waals surfaces of the molecules. Since all steric grid points contribute zero to the field beyond the molecular surfaces, a smaller rectilinear box ( $12 \times 20 \times 11 \text{ \AA}^3$ ) was sufficient in the calculation of the shape similarity index. A grid spacing of  $0.5 \text{ \AA}$ , as suggested by Good et al., was employed.<sup>16</sup> The Meyer formula (eq 5) was used to compute the shape similarity matrix (SSM).<sup>12</sup>

The parameters for the vdW interaction energies were based on universal force field (UFF).<sup>36</sup> A comparison of UFF and Merck molecular force field (MMFF)<sup>37-41</sup> showed that the vdW energies derived from the two force fields are highly correlated ( $r^2 = 0.92$ ), though the UFF values are generally larger. The fields were generated using a Csp<sup>3</sup> probe in a  $20 \times 30 \times 20 \text{ \AA}^3$  grid (same size as the electrostatic grid) with a  $0.5\text{-\AA}$  grid spacing, and a  $\pm 5.0$  kcal/mol truncation cutoff was applied. The vdW similarity matrix (VSM) was calculated according to the Carbó formula (eq 2).

**Genetic Neural Networks.** In GNN, selection of descriptors is made using a genetic algorithm<sup>42</sup> and correlation of

Chart 1. Steroid Data Set



biological activities with these descriptors is performed by a neural network.<sup>43</sup> Unless otherwise stated a 6-2-1 scaled conjugate gradient (SCG) neural network,<sup>44</sup> which contained 17 adjustable weights, was used. Empirical studies suggested that with this number of parameters, the neural network would be able to generalize a data set containing 31 compounds. Larger networks were less desirable because of an increasing risk of data overfitting.<sup>45-47</sup> The standard GNN protocol described previously was used in all simulations.<sup>22</sup> The protocol was designed to maximize the efficiency of GNN simulation by taking the computationally intensive cross-

validation procedure only during the final round of simulation. It had been demonstrated that this protocol generated results comparable to a genuine cross-validated GNN simulation.<sup>22</sup> The added efficiency makes possible a more extensive investigation of the data set. In the current work 200 individuals and 50 evolutionary programming<sup>20,22,48</sup> genetic reproduction cycles were used. The correlation coefficients of the training set were used as the fitness function for the first 50 cycles. In the final cycle, leave-one-out cross-validations were performed, and the cross-validated correlation coefficients became the fitness criteria to determine the final ranking of the GNN

**Table 1.** CBG Binding Affinity Data<sup>27</sup>

no.	log <i>K</i>	no.	log <i>K</i>	no.	log <i>K</i>
1	6.279	11	7.881	21	6.724
2	5.000	12	5.919	22	7.512
3	5.000	13	5.000	23	7.553
4	5.763	14	5.000	24	6.779
5	5.613	15	5.000	25	7.200
6	7.881	16	5.225	26	6.144
7	7.881	17	5.225	27	6.247
8	6.892	18	5.000	28	7.120
9	5.000	19	7.380	29	6.817
10	7.653	20	7.740	30	7.688
				31	5.797

models.<sup>49</sup> A typical GNN simulation for the steroid data set required approximately 1 CPU hour on a 175-MHz R4400 Silicon Graphics Indigo2 workstation with this protocol.

### III. Results

**Construction of a Standard Comparison Set.** The electrostatic and the shape SMs (Table 2) for the 31 CBG-binding steroids were obtained using the protocol outlined in the Method section. Six-descriptor GNN QSARs were built from each of the matrices. To investigate the variability of the results obtained by the GNN method, the two sets of simulations (ESM/GNN and SSM/GNN) were repeated 50 times using different random initial seeds. In all cases, they led to QSARs that were excellent in fitting and predicting the biological data. For the ESM, the correlation coefficient ( $r^2_{\text{tm}}$ ) of the fit of the data was  $0.951 \pm 0.004$ ; more importantly the cross-validated correlation coefficient ( $q^2$ ) value was  $0.903 \pm 0.007$ . The very small standard deviations indicated good convergence behavior, despite a small variation of descriptor selections that was observed in the multiple simulations. Each individual model shared, on average, four common descriptors with the others. Good results were also obtained with SSM/GNN. The 50 simulations yielded an average  $r^2_{\text{tm}}$  value of  $0.885 \pm 0.003$  and a  $q^2$  value of  $0.825 \pm 0.013$ . The statistical variables obtained from the two sets of benchmarking simulations served as a standard for comparison in later studies.

**Validation.** It is important that any new QSAR method be extensively validated. If this is not done, there is the possibility that the results may be due to chance correlation.<sup>47,50</sup> Because the GNN selects only a few variables in the final models, there is the possibility that some combination of input variables gives an apparently excellent fit or even good predictivity even though the correlation between the input descriptors and biological responses is not meaningful. The same issue has been discussed by the authors of another variable selection routine, GOLPE,<sup>15</sup> and we use a similar approach to show that the potential problem associated with chance correlation is minimal. We have adopted the randomization test<sup>15,21,22,25,51</sup> that has become a standard for QSAR validation because it provides an estimate of chance correlation. In this technique, the elements of the response vector (here, the binding affinity with CBG) are shuffled by 100 random exchanges in their positions. This is an efficient way to randomize the output values without altering the variance. Given a randomized data set, the same QSAR technique is applied, i.e., using GNN to correlate the real input data with the randomly scrambled activity. The whole procedure is performed on many different

scrambled data sets. If some apparently highly predictive QSARs are still obtained with randomized data, the significance of the real QSAR is suspect. We carried out the 50 ESM/GNN and 50 SSM/GNN simulations with data sets containing different randomized activity vectors. The results are shown in Figure 1, where the 50 artificial QSARs are displayed alongside the 50 true QSARs based on their correlation coefficients. It is evident from the plots that the points corresponding to the real QSARs are well-separated from the random cases. Both the fit of the data ( $r^2_{\text{tm}}$ ) and the cross-validation statistics ( $q^2 = 0.134 \pm 0.229$  and  $0.202 \pm 0.137$  for randomized ESM and SSM) from the artificial QSAR models were much lower than those of the real QSARs. This confirms that the predictive quality of the GNN models is meaningful.

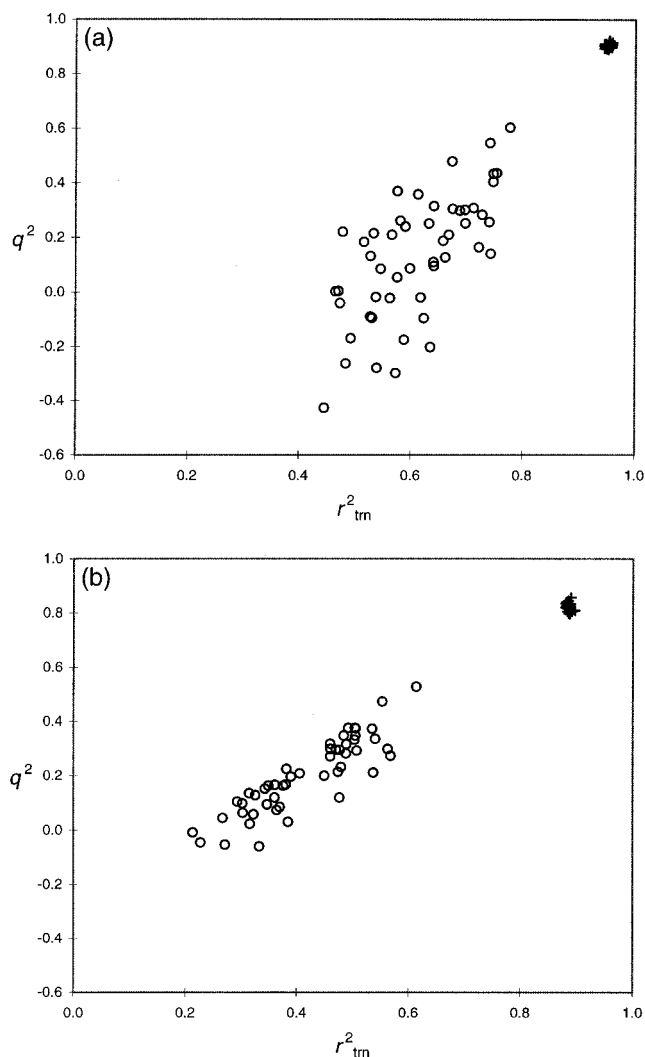
Some researchers have raised a question on the cross-validation process used in the SM/PLS studies.<sup>15,23</sup> Because similarity matrices are symmetric, there is a concern that even though the input vector associated with the cross-validated compound (i.e., the row representing the similarity of that compound with the others) has been removed from the matrix, the information is still present in the corresponding data column (i.e., the column containing the similarity measures of every compound with the cross-validated one). However, we do not think this is a problem in SM/GNN. First, after variable selection with the genetic algorithm, the data matrix that is examined by the neural network is no longer symmetric. Only when a compound that provides the similarity descriptor is being cross-validated can the similarity elements in the removed row be found at some random positions in the descriptor column corresponding to the compound. Further, a very important aspect of these "mirrored" elements is that they come from the input matrix but not the output values (i.e., activities) that cross-validation attempts to predict. There is no reason for this to cause a bias in the QSAR so that better (or worse) predictions are obtained for the compounds which provided the similarity descriptors, relative to the other compounds. To verify this, the following simulations were performed. Each of the 31 steroids in the data set was used in turn to provide electrostatic or shape similarity description, and an individual QSAR was constructed using a 1-2-1 neural network for every SM. We then calculated the cross-validated prediction error for every compound in each model and determined the percentage of compounds that had higher prediction error than the one that provided the similarity descriptor. The simulation results showed that the percentage of cases with higher prediction errors (52%) than that of the similarity template is about the same as those with lower margin of prediction errors (48%).

**Variations of Parameters for Electrostatic SM/GNN QSAR.** Since there are a number of user-defined parameters involved in this method, a systematic study was made to investigate the sensitivity of the results to changes in each parameter. The parameters investigated were (i) how the electrostatic potential was calculated, which included the use of different types of atomic charges, truncation cutoffs, and dielectric constants; (ii) types of similarity indices being used; (iii) the grid parameters including its spacing, size, or location; and (iv) the number of descriptors in the GNN calculation. For each test, three GNN simulations were

Table 2. Molecular Similarity Matrices for the 31 Compounds<sup>a</sup>

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	1	0.862	0.865	0.893	0.850	0.888	0.880	0.871	0.854	0.891	0.883	0.889	0.819	0.813	0.803	0.731	0.856	0.848	0.898	0.889	0.904	0.828	0.834	0.885	0.884	0.880	0.881	0.876	0.874	0.859	0.856
2	0.418	1	0.927	0.922	0.961	0.881	0.873	0.867	0.911	0.887	0.879	0.955	0.843	0.840	0.832	0.776	0.885	0.878	0.897	0.898	0.938	0.836	0.829	0.882	0.880	0.920	0.880	0.875	0.877	0.855	0.852
3	0.445	0.625	1	0.920	0.905	0.875	0.869	0.878	0.982	0.881	0.874	0.940	0.859	0.855	0.847	0.757	0.953	0.944	0.890	0.883	0.931	0.812	0.824	0.893	0.874	0.909	0.874	0.868	0.867	0.849	0.846
4	0.641	0.316	0.349	1	0.932	0.920	0.913	0.904	0.933	0.927	0.920	0.941	0.841	0.837	0.855	0.783	0.883	0.876	0.937	0.929	0.976	0.840	0.866	0.920	0.919	0.951	0.920	0.913	0.911	0.892	0.889
5	0.117	-0.034	-0.093	0.473	1	0.869	0.863	0.855	0.918	0.874	0.869	0.937	0.824	0.820	0.836	0.796	0.868	0.861	0.884	0.877	0.921	0.828	0.818	0.869	0.867	0.901	0.869	0.862	0.863	0.847	0.844
6	0.483	0.425	0.438	0.358	-0.326	1	0.991	0.951	0.864	0.992	0.983	0.896	0.807	0.803	0.794	0.746	0.913	0.906	0.982	0.973	0.937	0.905	0.940	0.951	0.984	0.913	0.963	0.957	0.959	0.967	0.963
7	0.499	0.432	0.450	0.442	-0.269	0.968	1	0.957	0.859	0.983	0.993	0.888	0.799	0.796	0.788	0.743	0.906	0.913	0.973	0.982	0.929	0.912	0.949	0.942	0.975	0.905	0.973	0.951	0.950	0.976	0.972
8	0.662	0.378	0.408	0.382	-0.268	0.785	0.821	1	0.868	0.957	0.964	0.866	0.817	0.814	0.806	0.734	0.912	0.917	0.949	0.955	0.907	0.893	0.910	0.980	0.952	0.890	0.945	0.928	0.932	0.936	0.935
9	0.352	0.149	0.397	0.694	0.666	-0.044	0.016	-0.010	1	0.870	0.864	0.925	0.847	0.843	0.859	0.766	0.943	0.935	0.879	0.873	0.915	0.802	0.815	0.882	0.864	0.892	0.864	0.857	0.855	0.839	0.837
10	0.520	0.459	0.474	0.420	-0.258	0.979	0.964	0.809	-0.007	1	0.991	0.902	0.810	0.807	0.797	0.743	0.919	0.911	0.990	0.980	0.945	0.901	0.933	0.957	0.991	0.920	0.970	0.965	0.966	0.960	0.956
11	0.523	0.453	0.474	0.508	-0.176	0.916	0.973	0.830	0.065	0.956	1	0.894	0.803	0.800	0.792	0.741	0.912	0.919	0.980	0.990	0.936	0.908	0.941	0.948	0.962	0.911	0.980	0.958	0.957	0.969	0.965
12	0.636	0.551	0.601	0.714	0.056	0.623	0.673	0.615	0.276	0.704	0.746	1	0.858	0.855	0.846	0.772	0.898	0.891	0.912	0.903	0.952	0.834	0.842	0.902	0.895	0.934	0.895	0.889	0.892	0.867	0.865
13	0.505	0.341	0.649	0.267	0.028	0.347	0.357	0.442	0.222	0.372	0.370	0.506	1	0.989	0.981	0.708	0.827	0.820	0.819	0.811	0.847	0.752	0.758	0.833	0.806	0.869	0.803	0.798	0.837	0.785	0.788
14	0.349	0.124	0.415	0.374	0.306	0.154	0.174	0.196	0.470	0.176	0.188	0.360	0.775	1	0.970	0.703	0.824	0.817	0.815	0.808	0.844	0.749	0.755	0.829	0.802	0.865	0.806	0.802	0.834	0.782	0.785
15	0.406	0.035	0.273	0.606	0.712	-0.068	-0.016	0.042	0.817	-0.032	0.028	0.245	0.540	0.740	1	0.716	0.814	0.808	0.806	0.800	0.836	0.741	0.748	0.820	0.793	0.857	0.792	0.785	0.824	0.775	0.778
16	0.288	0.018	0.044	0.522	0.868	-0.225	-0.161	-0.051	0.660	-0.145	-0.055	0.178	0.184	0.310	0.726	1	0.732	0.730	0.751	0.748	0.777	0.776	0.705	0.744	0.737	0.747	0.744	0.743	0.721	0.737	0.736
17	0.272	0.462	0.702	0.025	-0.448	0.719	0.685	0.549	-0.048	0.730	0.666	0.417	0.458	0.182	-0.110	-0.318	1	0.991	0.928	0.921	0.889	0.848	0.860	0.926	0.912	0.868	0.912	0.905	0.908	0.886	0.883
18	0.244	0.426	0.719	0.115	-0.325	0.585	0.648	0.531	0.075	0.615	0.676	0.434	0.449	0.186	-0.019	-0.191	0.870	1	0.921	0.928	0.881	0.853	0.867	0.918	0.904	0.861	0.919	0.901	0.892	0.889	
19	0.497	0.455	0.484	0.472	-0.305	0.934	0.944	0.787	-0.028	0.961	0.946	0.771	0.371	0.178	-0.062	-0.192	0.692	0.616	1	0.990	0.954	0.894	0.923	0.966	0.981	0.930	0.981	0.975	0.976	0.950	0.946
20	0.488	0.423	0.460	0.563	-0.199	0.826	0.913	0.778	0.061	0.869	0.951	0.793	0.352	0.180	0.012	-0.081	0.567	0.648	0.941	1	0.945	0.902	0.932	0.956	0.972	0.921	0.990	0.968	0.967	0.959	0.955
21	0.677	0.526	0.563	0.738	-0.063	0.739	0.786	0.706	0.225	0.794	0.828	0.933	0.470	0.315	0.182	0.059	0.451	0.452	0.872	0.884	1	0.853	0.881	0.924	0.937	0.974	0.936	0.930	0.927	0.907	0.903
22	0.463	0.378	0.422	0.470	-0.258	0.930	0.967	0.819	0.036	0.931	0.949	0.665	0.344	0.173	0.010	-0.117	0.629	0.608	0.930	0.912	0.789	1	0.868	0.880	0.894	0.833	0.896	0.878	0.875	0.892	0.889
23	0.553	0.367	0.397	0.722	0.065	0.575	0.651	0.536	0.337	0.587	0.652	0.686	0.314	0.233	0.274	0.158	0.294	0.353	0.684	0.738	0.773	0.666	1	0.893	0.925	0.858	0.923	0.902	0.901	0.926	0.922
24	0.637	0.373	0.410	0.354	-0.366	0.794	0.799	0.964	-0.082	0.805	0.789	0.630	0.443	0.184	-0.027	-0.149	0.557	0.492	0.825	0.779	0.738	0.804	0.565	1	0.952	0.906	0.947	0.941	0.949	0.921	0.921
25	0.527	0.400	0.439	0.450	-0.254	0.936	0.940	0.865	-0.011	0.962	0.945	0.710	0.367	0.167	-0.024	-0.098	0.649	0.574	0.947	0.889	0.815	0.948	0.607	0.863	1	0.912	0.962	0.956	0.958	0.952	0.948
26	0.692	0.538	0.576	0.720	-0.090	0.755	0.798	0.724	0.213	0.804	0.831	0.922	0.486	0.319	0.174	0.037	0.476	0.468	0.877	0.882	0.995	0.796	0.771	0.757	0.820	1	0.912	0.906	0.952	0.883	0.879
27	0.403	0.336	0.364	0.638	-0.055	0.704	0.809	0.643	0.183	0.753	0.858	0.743	0.260	0.292	0.118	-0.011	0.423	0.539	0.842	0.930	0.824	0.817	0.709	0.638	0.777	0.813	1	0.978	0.957	0.950	0.946
28	0.497	0.450	0.479	0.468	-0.304	0.931	0.944	0.789	-0.023	0.967	0.945	0.763	0.367	0.192	-0.056	-0.192	0.688	0.617	0.993	0.938	0.863	0.929	0.679	0.823	0.944	0.868	0.852	1	0.952	0.928	0.924
29	0.507	0.459	0.489	0.453	-0.324	0.942	0.947	0.797	-0.036	0.962	0.941	0.754	0.380	0.178	-0.066	-0.208	0.706	0.622	0.996	0.930	0.860	0.929	0.677	0.836	0.944	0.874	0.825	0.989	1	0.927	0.923
30	0.483	0.446	0.450	0.421	-0.272	0.963	0.995	0.813	0.002	0.958	0.966	0.656	0.358	0.169	-0.026	-0.175	0.695	0.657	0.936	0.903	0.771	0.955	0.636	0.789	0.930	0.783	0.797	0.936	0.940	1	0.996
31	0.401	0.356	0.385	0.466	-0.156	0.845	0.917	0.741	0.016	0.877	0.943	0.671	0.324	0.169	0.002	-0.066	0.567	0.617	0.891	0.926	0.779	0.921	0.619	0.706	0.904	0.775	0.860	0.889	0.881	0.918	1

<sup>a</sup> The electrostatic similarity index is given on the left of the diagonal, and the shape similarity index is given on the right.



**Figure 1.** Scatter plot for  $q^2$  against  $r^2_{tm}$  for the real QSAR (+) and those with randomized activity values (○): (a) electrostatic similarity matrix; (b) shape similarity matrix.

made to estimate the statistical variance. The average and standard deviation of the cross-validated correlation coefficients ( $q^2$ ) are reported alongside with the standard comparison set that was obtained from the earlier validation run (see above).

**(i) Electrostatic Potential Calculation. (a) Effect of Atomic Charges.** AM1 Mulliken charges, which required a few CPU seconds to calculate on a R4400 Silicon Graphics workstation, were used for electrostatics calculation in the standard comparison set. In the validation, four other types of charging algorithms were tried, all of which were available within the Cerius2 modeling environment.<sup>33</sup> They were the AM1 ESP-fitted charges from MOPAC6,<sup>32</sup> the Gasteiger charges,<sup>52</sup> the charge equilibration (QEq) charges,<sup>53</sup> and the MMFF charges.<sup>37–41</sup> Table 3a shows the cross-validated results for the various types of charges. The AM1 Mulliken charges and the QEq charges give similar results, which is consistent with the fact that the atomic charges calculated by the two algorithms are highly correlated ( $r^2 = 0.88$ ; Table 4a). The use of ESP-fitted charges, which are more different from the other four charge types, leads to a small decrease in predictivity. The Gasteiger and the MMFF charges, the second most correlated pair ( $r^2 = 0.84$ ), produce the worst results in this case. On the basis of this comparison, the AM1

**Table 3.** Statistical Data for GNN QSARs Derived from Electrostatic Similarity Matrices

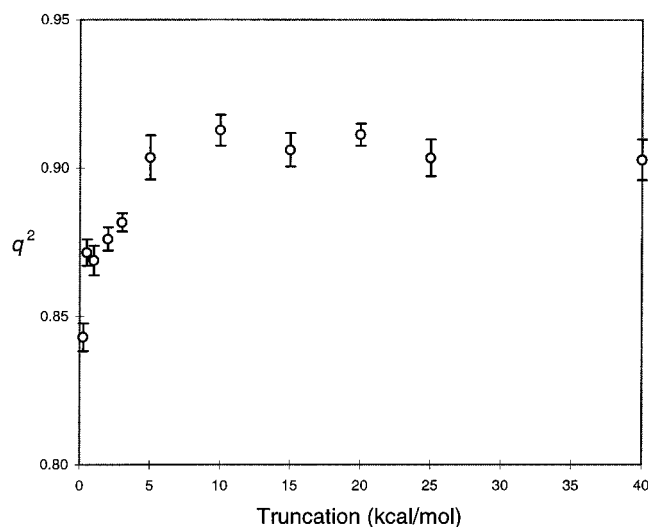
	$q^2$	$q^2$
(a) Charge		
AM1/Mulliken <sup>a</sup>	0.903 ± 0.007	QEq 0.913 ± 0.021
AM1/ESP	0.841 ± 0.012	MMFF 0.819 ± 0.032
Gasteiger	0.801 ± 0.010	
(b) Truncation		
±0.25	0.843 ± 0.005	±10 0.913 ± 0.005
±0.50	0.871 ± 0.004	±15 0.906 ± 0.006
±1	0.869 ± 0.005	±20 0.911 ± 0.004
±2	0.876 ± 0.004	±25 0.903 ± 0.006
±3	0.882 ± 0.003	none 0.903 ± 0.007
±5 <sup>a</sup>	0.903 ± 0.007	
(c) Dielectric		
1 <sup>a</sup>	0.903 ± 0.007	40 0.907 ± 0.005
4	0.908 ± 0.008	80 0.903 ± 0.007
(d) Similarity Index		
Hodgkin <sup>a</sup>	0.903 ± 0.007	exponential 0.914 ± 0.009
Carbó	0.886 ± 0.014	linear 0.898 ± 0.007
(e) Grid Spacing		
0.5	0.890 ± 0.004	3 0.874 ± 0.006
1	0.895 ± 0.006	4 0.848 ± 0.011
2 <sup>a</sup>	0.903 ± 0.007	
(f) Grid Size		
-2	0.832 ± 0.004	+6 <sup>a</sup> 0.903 ± 0.007
0	0.903 ± 0.007	+8 0.899 ± 0.019
+2	0.907 ± 0.008	+10 0.900 ± 0.012
+4	0.901 ± 0.006	
(g) Grid Shift		
+x	0.899 ± 0.005	-x 0.895 ± 0.007
+y	0.898 ± 0.010	-y 0.894 ± 0.008
+z	0.890 ± 0.003	-z 0.893 ± 0.006
standard <sup>a</sup>	0.903 ± 0.007	
(h) Number of Descriptors		
1	0.677 ± 0.010	11 0.883 ± 0.010
2	0.768 ± 0.009	12 0.859 ± 0.022
3	0.851 ± 0.002	13 0.858 ± 0.008
4	0.888 ± 0.007	14 0.852 ± 0.016
5	0.896 ± 0.003	16 0.857 ± 0.024
6 <sup>a</sup>	0.903 ± 0.007	18 0.840 ± 0.005
7	0.899 ± 0.010	20 0.836 ± 0.011
8	0.899 ± 0.011	25 0.824 ± 0.007
9	0.897 ± 0.007	31 0.712 ± 0.014
10	0.879 ± 0.016	

<sup>a</sup> Standard comparison set.

**Table 4.** Correlation Coefficients ( $r^2$ )

(a) Among Different Types of Atomic Charges				
	AM1/ESP	Gasteiger	QEq	MMFF
AM1/Mulliken	0.61	0.61	0.88	0.36
AM1/ESP		0.26	0.55	0.13
Gasteiger			0.76	0.84
QEq				0.46
(b) Among Different Types of Similarity Indices				
	Carbó	linear	exponential	
Hodgkin	0.99	0.89	0.80	
Carbó		0.86	0.75	
linear			0.98	

Mulliken charges set appears to be a good balance between computational efficiency and accuracy in the derivation of atomic charges for the purpose of obtaining electrostatic similarity descriptors. The ESP-fitted charges, which took more than 1000 s to generate, are less desirable due to their high computational cost. The fact that the QEq algorithm derives good charges for similarity calculations is also significant. Because the charges can be obtained very rapidly (estimated at 20 compounds/s), they are well-suited to problems that



**Figure 2.**  $q^2$  as a function of truncation cutoff for the electrostatic potential calculations.

require the handling of a vast number of structures, for example, in database searches or conformational analysis for compounds with many degrees of freedom.

**(b) Truncation of Electrostatic Potential.** A cutoff is often used in the generation of electrostatic potential fields.<sup>14,16</sup> This is done to prevent the similarity index from being dominated by a few grid points with a large magnitude for the electrostatic potential. Any value that was beyond some predefined cutoff was set to the maximum or minimum values. Table 3b and Figure 2 show the cross-validated statistics for simulations performed with different values for the cutoff. Use of a very small value clearly degrades the results. Comparison of the results with cutoffs larger than  $\pm 5.0$  kcal/mol indicates that there is little effect for the present case. This is not surprising since the grid values generally vary within this range and only a few grid points (about 0.6%) have values that are greater than 10.0 kcal/mol in magnitude.

**(c) Dielectric Constant.** Because the similarity index is a ratio between electrostatic potentials, the dielectric terms simply cancel when the index is computed (eq 1). Table 3c shows the variation in the predictivity of the QSARs that used different dielectric constants ranging from 1 (vacuum) to 80 (aqueous) in the electrostatic potentials calculation. As expected, the changes in cross-validated results with dielectric constants are minimal, and the small observed deviations are an artifact of the truncation cutoffs.

**(ii) Similarity Index.** Three other similarity indices were investigated. They were the Carbó index (eq 1)<sup>9</sup> and the linear and exponential indices proposed by Good (eqs 3 and 4).<sup>11</sup> Table 4b shows the correlation among the four electrostatic SMs calculated with different indices based on the same grid potentials. It suggests that the similarity measures from these indices in a congeneric data set are highly correlated, and not surprisingly GNN gives comparable results when applied to the four SMs (Table 3d). There is no apparent advantage of using one index over any other in this test case.

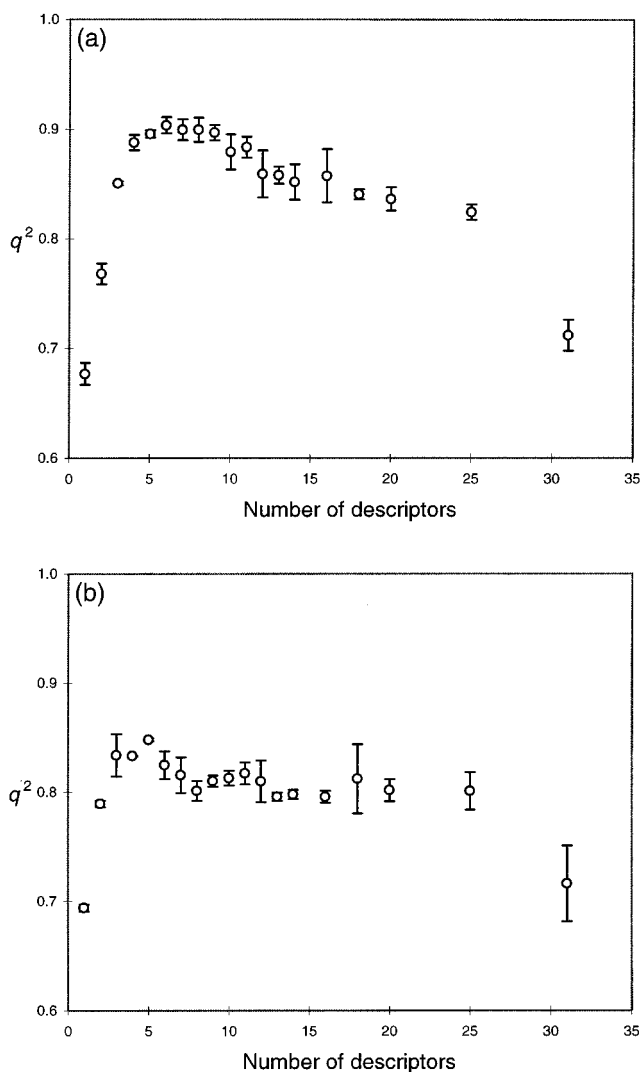
**(iii) Grid Parameters. (a) Spacing.** The effect of different grid spacing on the ESM/GNN QSAR was explored. Five grids, which were approximately equal

in size and had grid spacing ranging from 0.5 to 4.0 Å, were used in the calculation. The results are shown in Table 3e. They indicate that a spacing of 2.0 Å is sufficient since 0.5 or 1.0 Å gave very similar results, in accord with an earlier study.<sup>16</sup> Since a 2-fold decrease in grid spacing leads to an 8-fold increase in the number of grid points, this is important for an efficient method. Further increase of grid spacing beyond 2.0 Å leads to a gradual decrease in predictivity, though it is interesting that a satisfactory model can still be made using a value as large as 4.0 Å. This may be due to the long-range nature of the electrostatic interactions.

**(b) Size.** A grid of size  $8 \times 18 \times 8$  Å<sup>3</sup> was the smallest rectilinear grid (with a 2.0-Å grid spacing) that contained all steroid molecules. A grid size parameter,  $s$ , was defined to give a grid that would extend beyond all molecules by at least  $s$  Å along all axes; i.e., a grid of dimension  $(8+2s) \times (18+2s) \times (8+2s)$  Å<sup>3</sup> was used. In the standard set of simulations, a grid size  $s = 6$  was used. This ensured that the electrostatic field values had decayed to near zero at the grid surfaces. In the present validation, grids with  $s$  from  $-2$  to 10 and spacing fixed at 2.0 Å were examined. The results shown in Table 3f indicate that the quality of the QSAR models is insensitive to grid size, a parameter that is often chosen arbitrarily, as long as the grid contains all molecules, i.e.,  $s \geq 0$ . They also suggest that the dominant component in similarity calculations is derived from the grid points closest to the molecules. It seems reasonable because this region is where hydrogen bonding and the strongest charge interactions take place.

**(c) Location.** Another variable in any grid-based calculation is the exact location of the rectilinear grid, relative to the molecules. For a robust QSAR method, the performance of the model will not depend on the exact location of the grid, as long as all the molecules of interest are contained in it. To monitor this effect, the same electrostatic SM/GNN calculations were performed with the grids shifted by  $\pm 0.2$  Å in the  $x$ -,  $y$ -, and  $z$ -directions relative to the standard set of simulations. The results in Table 3g show that the location of the grid has little or no effect on the quality of QSAR.

**(iv) Number of Similarity Descriptors in QSAR.** The effect of the number of descriptors ( $n$ ) used in ESM/GNN QSAR on the quality of model was investigated. In the standard set of simulations, six similarity descriptors were chosen by the GA and were utilized to correlate with binding affinity using a 6-2-1 neural network. In this test study 19 different values of  $n$ , ranging between 1 and 31, were examined in conjunction with an  $n$ -2-1 neural network. Table 3h and Figure 3a show the cross-validation results for the various GNN simulations. The best single-descriptor model ( $n = 1$ ) yielded a  $q^2$  value of 0.677. This mimicked the QSAR study of a set of bisamidines by Montanari et al., who reported a respectable regression model using only the Carbó index associated with the most potent analogue.<sup>19</sup> In the present study, use of additional similarity descriptors led to a rapid increase in predictivity. However, this effect leveled off at 6–8 descriptors, and then  $q^2$  began to decrease steadily upon further increase of  $n$ . This decline in predictivity might be due to data overfitting associated with the use of larger neural networks.<sup>45,46</sup>



**Figure 3.** Variation of  $q^2$  as a function of the number of input nodes used in the GNN simulations: (a) electrostatic similarity matrix; (b) shape similarity matrix.

To summarize, this part of the validation study demonstrated that the resulting QSAR obtained from the ESM/GNN was very robust with respect to variations in most of the user-specified, and often arbitrarily chosen, parameters. Because the various similarity indices are highly correlated, the particular choice of index had little influence in the resulting QSARs. The grid-related settings such as its spacing, size, or location had relatively little or no impact on the overall result. In addition, the numerical values of the dielectric in the electrostatics calculation played no role, as expected from the functional form of the similarity index. It appears that too small a value for the truncation cutoff could result in very flat similarity indices. The study of different atomic charges suggested that the MOPAC AM1 Mulliken charges and QEq charges were suitable for the calculations of electrostatic potentials. It was shown that the number of descriptors had a strong effect on the resulting QSAR models. It was important to have enough descriptors to reflect the data but not so many that overfitting could arise.

**Variations of Parameters in Shape SM/GNN QSAR.** Similar validations were performed on the SSM/GNN QSARs. Again three simulations were done on the different settings of parameters. There are considerably fewer user-adjustable parameters in this

**Table 5.** Statistical Data for GNN QSARs Derived from Shape Similarity Matrices

	$q^2$		$q^2$	
	(a) Grid Spacing			
0.25	0.828 ± 0.003	2	0.809 ± 0.011	
0.50 <sup>a</sup>	0.825 ± 0.013	3	0.601 ± 0.029	
1	0.821 ± 0.003	4	0.359 ± 0.059	
	(b) Grid Shift			
+x	0.838 ± 0.013	-x	0.831 ± 0.001	
+y	0.822 ± 0.014	-y	0.826 ± 0.007	
+z	0.847 ± 0.029	-z	0.832 ± 0.004	
standard <sup>a</sup>	0.825 ± 0.013			
	(c) Number of Descriptors			
1	0.694 ± 0.003	11	0.817 ± 0.010	
2	0.789 ± 0.003	12	0.810 ± 0.019	
3	0.834 ± 0.019	13	0.796 ± 0.004	
4	0.833 ± 0.001	14	0.798 ± 0.004	
5	0.848 ± 0.001	16	0.796 ± 0.005	
6 <sup>a</sup>	0.825 ± 0.013	18	0.812 ± 0.032	
7	0.816 ± 0.016	20	0.802 ± 0.010	
8	0.801 ± 0.009	25	0.801 ± 0.017	
9	0.810 ± 0.005	31	0.716 ± 0.035	
10	0.813 ± 0.007			

<sup>a</sup> Standard comparison set.

case: they are (i) the grid parameters (spacing and location) and (ii) the number of shape similarity descriptors that are used.

**(i) Grid Parameters. (a) Spacing.** The effect of different grid spacings on the calculation of the SSM/GNN was explored. Six different spacing settings, ranging from 0.25 to 4.0 Å, were used. Table 5a shows the cross-validated statistics for the GNN QSARs. The results indicate that a grid spacing between 0.25 and 1.0 Å gives similar results. However, in contrast to the electrostatic results, when the resolution of the grid is beyond 1.0 Å, the predictivity of resulting QSARs drops rapidly. This is in accord with the shorter range, steeper variation of the shape parameters.

**(b) Location.** An offset of ±0.05 Å was applied to the standard grid at each of the *x*-, *y*-, and *z*-directions. As shown in Table 5b, none of the alternative QSAR models had cross-validated correlation coefficients that were significantly different than the standard simulations. This result verified that the exact location of the grid had little influence on the overall result.

**(ii) Number of Similarity Descriptors in QSAR.** As for the ESM/GNN, 19 SSM/GNN simulations using different numbers of similarity descriptors were performed. The results are shown in Table 5c and also in Figure 5b. As with the electrostatic case, the quality of QSAR increased when more than one descriptor was used. There was no gain in predictivity for models using more than five similarity descriptors.

To summarize, like the previous case with electrostatic SM, the exact location of the rectilinear grid did not appear to influence the quality of the QSAR model. However, unlike the electrostatic case, a finer grid between 0.25 and 1.0 Å was required. Larger grid spacing gave inadequate shape description that subsequently made discrimination among molecules more difficult. Again the number of descriptors was important.

**Combined Electrostatic and Shape SM/GNN QSAR.** We have shown that the application of GNN to either the electrostatic or shape similarity matrix alone leads to highly predictive QSARs. This suggested that both electrostatic and steric factors were important



in determining the binding affinity of the steroid data set to CBG. Consequently, an attempt was made to improve the QSAR by using both types of information simultaneously.

A  $31 \times 62$  similarity matrix was constructed by merging the electrostatic and shape similarity matrices. With this matrix, the GNN simulation yielded a highly predictive QSAR model using three electrostatic and three shape descriptors. This model yielded a  $q^2$  value of 0.941, significantly higher than those obtained by using either electrostatic or shape descriptors alone. Interestingly this behavior is quite different from an earlier SM/PLS study on the same data set,<sup>14</sup> which reported worse results with the combined matrix than the shape matrix alone.

#### Comparison of Shape and van der Waals Fields.

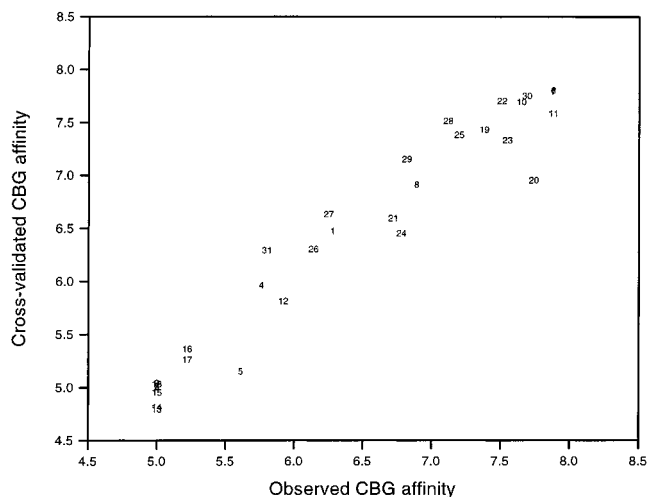
In this and the other SM QSAR studies, a binary shape field is used to provide the steric description of the ligand–receptor environment. In CoMFA, the steric field is a numeric measure which is commonly derived from vdW interactions between the molecules and a methyl probe. The vdW potentials allow a softer variation in the steric field values in the regions that are close to the molecular surfaces. In this study we investigated whether this alternative steric field would lead to a change in QSAR predictivity.

VSM was obtained (see Method). It was very similar to SSM; the correlation coefficient ( $r^2$ ) between the two steric matrices was 0.92. Based on 20 multiple runs, the six-descriptor VSM/GNN simulations gave a  $q^2$  value of  $0.843 \pm 0.009$  which was comparable with the SSM/GNN results ( $q^2 = 0.825 \pm 0.013$ ). Further, running GNN on the combined vdW and electrostatic SM yielded a  $q^2$  value of 0.941, which was virtually the same as the corresponding value obtained from the combined shape and electrostatic matrix. The results demonstrate the practical equivalence of the two steric fields. Thus, the use of the softer vdW field, which came at the expense of a few extra user parameters (e.g., more vdW interaction parameters, truncation cutoff, and grid size), does not appear to be justified, at least for the present case.

#### Comparison with Other Statistical Methods.

PLS<sup>3</sup> was performed with the combined electrostatic and shape similarity matrix. Cross-validation was performed to estimate the optimal number of components to be used in PLS. The final model contained five components and had a  $q^2$  of 0.707. A linear six-descriptor genetic regression model was also developed for comparison. Interestingly, the descriptors in this linear model were very different from the GNN selection; i.e., there was no common descriptor in the two sets. This model contained two electrostatic and four shape descriptors and yielded a  $q^2$  value of 0.819. This increase in predictivity from the PLS regression suggested that a GA-based descriptor selection is more appropriate for the analysis of similarity matrices. Furthermore, the ability of neural networks in handling nonlinearity implicitly<sup>54</sup> led to better model optimizations in GNN over genetic multiple linear regression methods. This is clear from the GNN results, relative to the others just described.

**Comparison with Other Studies.** Comparison with earliest studies had to be made with caution because there were a number of errors in the steroid



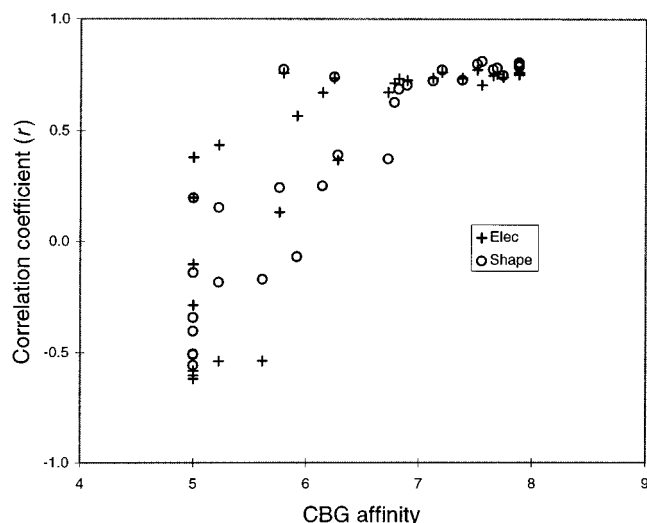
**Figure 4.** Plot of the cross-validated CBG affinity against the observed values. See Chart 1 for the numbering of the 31 steroids.

structures.<sup>27</sup> The earliest 3D QSAR work on this data set was made by Cramer et al. who proposed a CoMFA model that gave a  $q^2$  of 0.66 with the first 21 steroids.<sup>2</sup> For comparison, they also reported a regression equation based on molar refractivity ( $q^2 = 0.31$ ) as well as a molecular shape analysis<sup>55</sup> ( $q^2 = 0.56$ ) in the same paper. Good et al. performed PLS regression on molecular similarity matrices and obtained a QSAR with a  $q^2$  of 0.76.<sup>14</sup> The COMPASS program, based on an iterative procedure to suggest bioactive conformations and a neural network to correlate molecular surface properties, led to a highly predictive QSAR model with a  $q^2$  of 0.89.<sup>23</sup> Comparative molecular similarity indices analysis (CoMSIA), a variant of CoMFA, correlated a combination of electrostatic, steric, and hydrophobic “similarity” fields with binding affinity. It produced a five-component PLS model that yielded a  $q^2$  of 0.67.<sup>24</sup> Hahn and Rogers constructed a receptor surface model (RSM) and built a simple QSAR equation based on just the pseudo-drug–receptor interaction energy. Their model gave a  $q^2$  value of 0.63.<sup>25</sup> In a recent study, Silverman and Platt proposed a new 3D QSAR method, the comparative molecular moment analysis (CoMMA), and obtained a  $q^2$  value of 0.83 for the first 21 steroids and 0.69 for all 31.<sup>29</sup>

Wagener et al. noticed the error in the secondary sources and recompiled the data set from the original literature.<sup>27</sup> They reported a QSAR model based on the autocorrelation vector of molecular surface properties and neural networks. Their initial 12-descriptor model had a  $q^2$  value of 0.63 using all 31 steroids. Later, they assumed one compound (**31**) as an outlier in the data set, and upon its removal their QSAR model had a much higher predictive power ( $q^2 = 0.84$ ).

In the current study, we obtained a substantial improvement over the published QSARs on the data set. By applying the GNN methodology on a combined electrostatic and shape similarity matrix, we were able to derive a six-descriptor QSAR model that yielded a  $q^2$  of 0.94, using the entire data set. Figure 4 shows a scatter plot of the cross-validated CBG affinity against the observed values for the 31 steroids. In contrast to the results of Wagener et al., no outlier is identified.

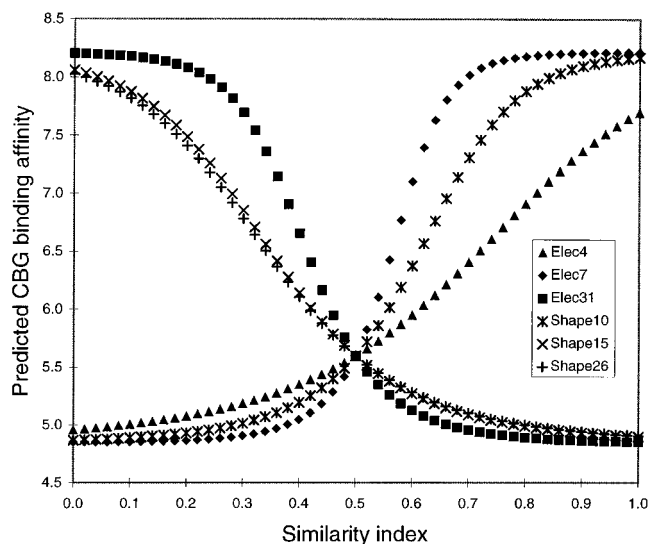
**Theoretical Basis of Similarity in QSAR.** A molecular similarity index is a different kind of QSAR



**Figure 5.** Correlation coefficient ( $r$ ) between similarity descriptor and activity plotted against the activity of the compound providing the descriptor: (+) electrostatic similarity descriptor; (O) shape similarity descriptor.

descriptor from conventional parameters (e.g.,  $\pi$ ,  $\sigma$ , and MR) because it does not encode physicochemical properties that are specific to molecular substituents. The index is derived from numerical integration and normalization of the field values (eqs 1–5), and it represents a global measure of the resemblance between a pair of molecules based on their spatial or electrostatic attributes, or a combination of the two. The use of molecular similarity offers a new perspective in QSAR. Instead of a correlation between substituent properties and activities, a similarity-based QSAR method establishes an association between global properties and activity variation among a series of molecules. The implicit assumption is that globally similar compounds have similar activities.<sup>8</sup> To test this, we calculated the Pearson correlation coefficient ( $r$ ) between similarity descriptors and activity for the steroid data set. The coefficient is plotted against the activity of the compound with which the similarity index is associated (Figure 5). The plot shows that all of the most active compounds have a large positive correlation and the majority of the least active ones have an anticorrelation. The positive correlation is, of course, the impetus for computer-assisted lead finding and optimization in drug design.

The above analysis provides the basis for discussing the selection of descriptors in the GNN QSAR. Six descriptors were obtained as optimal from the final application of GNN on the combined electrostatic and shape SM. They were **4**, **7**, and **31** for the electrostatic similarity, and **10**, **15**, and **26** for the shape. A functional dependence plot for these six descriptors is shown in Figure 6, which was made by keeping all but one of the similarity descriptors fixed at a constant value (0.5) while scanning the variation in the binding affinity with respect to changes in one descriptor.<sup>22,46</sup> It was not surprising to see compounds were predicted to have high CBG binding affinity when they were structurally similar to **7** or **10**, which are two of the more active compounds in the data set. Similarly, binding affinity of a compound decreased if it was close to either **15**, **26**, or **31**, three compounds with rather low binding affinity. However, it came as an initial surprise to see that

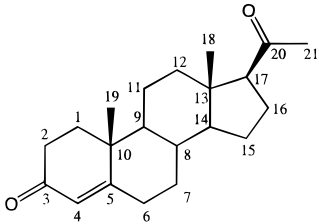


**Figure 6.** Functional dependence for the three electrostatic and three shape similarity descriptors (see text).

compound **4**, a compound with relatively low binding affinity, had a positive slope in the dependence plot. Examination of the electrostatic similarity matrix revealed the following: the two compounds most similar to **4** (**21** and **23**) had fairly high binding affinity, and the five most dissimilar compounds (**2**, **3**, **13**, **17**, and **18**) bound to CBG very weakly. This may also be elucidated at a structural level. Compound **4**, despite its low activity, contains both a 3-oxo group and a  $\Delta^4$  double bond, which are two of the essential features for optimal binding.<sup>56–59</sup> The presence of these structural features in the congeners most similar to **4** and the absence in the least ones seems to be the basis of the present similarity–activity relation.

**Consistency with Established SAR.** In this section we demonstrated that the final SM/GNN QSAR model with both electrostatic and steric properties was consistent with the known facts on the steroid–CBG interactions.<sup>56–59</sup> Previous SAR analysis on this data set revealed that the carbonyl groups at both C3 and C20 were essential for optimal binding, whereas the introduction of a carbonyl group at C11 decreased the binding affinity. It also was suggested that a hydroxyl group at the 11 $\beta$ , 17 $\alpha$ , or 21 position did not affect binding significantly, though such a group at the 11 $\alpha$  position would weaken binding (see Table 6 for numbering). A hydroxyl group was also found to impair binding at the following positions: 6 $\alpha$ , 6 $\beta$ , 11 $\alpha$ , 12 $\alpha$ , 14 $\alpha$ , 16 $\alpha$ , and 19. The introduction of a methyl group decreased binding in the 6 $\alpha$  and 16 $\alpha$  positions, though less so than a hydroxyl group would. The methyl group at the axial 10 $\beta$  position was important for optimal binding, presumably through some favorable van der Waals interactions with the receptor. Reduction of the  $\Delta^4$  double bond could lead to either *cis*- or *trans*-dihydro compounds, and both derivatives had a similar decrease in CBG affinity relative to the parent molecule. A 9 $\alpha$ -fluoro group was found to weaken binding significantly.<sup>25</sup>

A straightforward way to test the ability of the SM/GNN QSAR to account for the above SAR was to use a model compound and make structural modifications in accord with the above observations. We then ran the new analogues through the QSAR model to obtain the changes in predicted binding associated with any par-

**Table 6.** Consistency with the Established SAR


no.	modification on <b>19</b>	predicted affinity	comparing to <b>19</b> , SAR suggests that binding affinity of the modified analogue will be:	result <sup>a</sup>
<b>19</b>		7.45		
<b>32</b>	3-dihydro	4.90	lower	✓
<b>33</b>	3β-OH	5.06	lower	✓
<b>34</b>	3α-OH	4.87	lower	✓
<b>35</b>	20-dihydro	6.62	lower	✓
<b>36</b>	20β-OH	7.15	lower	✓
<b>37</b>	20α-OH	6.98	lower	✓
<b>38<sup>b</sup></b>	11-oxo	6.45	lower	✓
<b>39</b>	11β-OH	7.71	similar	✓
<b>40<sup>b</sup></b>	17α-OH	6.95	similar	✗
<b>41<sup>b</sup></b>	21-OH	7.69	similar	✓
<b>42</b>	11α-OH	6.72	lower	✓
<b>43</b>	6β-OH	7.64	lower	?
<b>44</b>	12α-OH	5.90	lower	✓
<b>45</b>	14α-OH	6.04	lower	✓
<b>46</b>	19-OH	7.18	lower	?
<b>47</b>	16α-OH	7.15	lower	✓
<b>48</b>	6α-OH	6.35	lower	✓
<b>49<sup>b</sup></b>	16α-CH <sub>3</sub>	7.51	lower, but higher than <b>47</b>	? ✓
<b>50</b>	6α-CH <sub>3</sub>	7.46	lower, but higher than <b>48</b>	? ✓
<b>51<sup>b</sup></b>	nor- <b>19</b>	7.15	lower	✓
<b>52</b>	5β-H (Δ <sup>4</sup> reduction)	6.52	lower	✓
<b>53</b>	5α-H (Δ <sup>4</sup> reduction)	6.85	lower, and similar to <b>52</b>	✓ ✗
<b>54</b>	9α-F	5.46	lower	✓
<b>55</b>	9α-CH <sub>3</sub>	7.43		

<sup>a</sup> A significant change in predicted affinity (PA) is defined by a shift of more than 0.3 log unit, which is approximately 10% of the activity range of the data set. For an analogue **X** that is suggested to have a lower binding affinity than **Y** by SAR, the corresponding GNN prediction is considered consistent (✓) if  $PA_x - PA_y < -0.3$ ; inconsistent (✗) if  $PA_x - PA_y > 0.3$ ; and inconclusive (?) if  $|PA_x - PA_y| \leq 0.3$ . For an analogue **X** whose binding affinity is suggested to be similar to that of **Y** by SAR, the GNN prediction is considered consistent (✓) if  $|PA_x - PA_y| \leq 0.3$ ; and inconsistent (✗) if  $|PA_x - PA_y| > 0.3$ . <sup>b</sup> Five of the modified analogues can be found in the training set (**38** = **24**; **40** = **20**; **41** = **10**; **49** = **28**; **51** = **29**). To avoid duplication, their predicted affinities were taken from cross-validated predictions.

ticular structural perturbation. **19** was chosen as the model template because it could be transformed readily to the structural analogues of interest. Table 6 shows the structural changes of **19** required to make these new analogues together with their predicted binding affinity. It is evident that the SM/GNN predictions are consistent with most of the known SARs.

An additional analogue was made to test whether the decrease in binding affinity for the 9α-fluoro derivative **54** (predicted affinity = 5.46) was due largely to unfavorable electrostatic interactions with the receptor, as suggested by the receptor surface model.<sup>25</sup> A 9α-methyl derivative (**55**) was made from **19**, and its binding affinity was predicted at 7.43 with the SM/GNN. This result seemed to suggest that a bulky hydrophobic group at this position did not affect binding whereas a small polar group had a prominent effect.

## IV. Concluding Discussion

A QSAR approach applying a genetic neural network (GNN) based on molecular similarity matrices (SM) has been described. In this initial application, the affinity of a well-studied set of CBG-binding steroids was examined. Excellent correlation and prediction were obtained from the use of either an electrostatic or a shape matrix alone, though the inclusion of both factors improved the quality of the QSAR. The result of the randomization test indicated that the predictivity of these models is statistically significant. Since the SM/GNN contains a number of user-defined parameters, tests were made to determine optimal values.

The combined SM/GNN QSAR model has been compared to those derived from statistical analysis using PLS and genetic regression methods. It was found that models with variable selection based on a genetic algorithm give better results. Furthermore, the benefit of using a nonlinear method, such as a neural network, in a complex modeling problem was evident. The GNN result was also compared to the benchmarks obtained by other 3D QSAR methods.

The basis for the similarity-based QSARs was discussed, and it was concluded that the essential element is the simple fact that globally similar compounds have similar activities. We also showed that the SM/GNN QSAR was consistent with the known SAR for CBG-steroid binding.

The SM/GNN method is not without its shortcomings. For example, the model is more difficult to interpret because conventional 2D descriptors are not involved in the modeling, its generation is computationally intensive, and its quality is dependent on a good molecular alignment. The other 3D QSAR techniques have much to offer, though most of them also require alignment. CoMFA and CoMSIA provide a convenient visual analysis based on the property contour maps in the important interaction regions. Performing PLS regression on similarity matrices is computationally inexpensive. The COMPASS program gives good statistics and offers an automatic way to suggest bioactive conformations and pick an alignment. The RSM approach leads to a simple regression equation that is easy to interpret and gives clues to the putative receptor environment. The use of autocorrelation vectors and molecular quadrupolar moments as descriptors is innovative, and these approaches can be especially useful in cases where the exact molecular alignment is far from obvious. We believe that by combining the important attributes of different 3D QSAR methods new insight can be sought and rapid progress will be made.

To demonstrate the general utility of the new SM/GNN approach, additional applications of this method have been made on eight data series. This study is reported in the companion paper.<sup>13</sup>

**Acknowledgment.** S.-S. S. is grateful to Prof. Graham Richards for introducing him to the molecular similarity concept. The steroid coordinates were obtained via anonymous ftp from the Gasteiger group. We thank Molecular Simulations for providing software including Cerius2, which was extensively used in the present work. This work is supported in part by a GAOLI grant from the National Science Foundation.

## References

- (1) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important molecules. *J. Med. Chem.* **1985**, *28*, 849–857.
- (2) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (3) Dunn, W. J.; Wold, S.; Edlund, U.; Hellberg, S. Multivariate structure–activity relationships between data from a battery of biological tests and an ensemble of structure descriptors: the PLS method. *Quant. Struct.–Act. Relat.* **1984**, *3*, 131–137.
- (4) Kubinyi, H. *3D QSAR in Drug Design: Theory, Methods and Applications*; ESCOM Science Publishers B.V.: Leiden, The Netherlands, 1993.
- (5) Cruciani, G.; Watson, K. A. Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. *J. Med. Chem.* **1994**, *37*, 2589–2601.
- (6) Dean, P. M. Molecular similarity. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers B. V.: Leiden, The Netherlands, 1993; pp 150–172.
- (7) Richards, W. G. Molecular similarity and dissimilarity. In *Modelling of Biomolecular Structures and Mechanisms*; Pullman, A., Jortner, J., Pullman, B., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; pp 365–369.
- (8) Kubinyi, H. A general view on similarity and QSAR studies. In *Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry*; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; VCH: Basel, and Wiley-VCH: Weinheim, 1997; pp 7–28.
- (9) Carbó, R.; Leyda, L.; Arnau, M. An electron density measure of the similarity between two compounds. *Int. J. Quantum Chem.* **1980**, *17*, 1185–1189.
- (10) Hodgkin, E. E.; Richards, W. G. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1987**, *14*, 105–110.
- (11) Good, A. C. The calculation of molecular similarity: alternative formulas, data manipulation and graphical display. *J. Mol. Graph.* **1992**, *10*, 144–151.
- (12) Meyer, A. M.; Richards, W. G. Similarity of molecular shape. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 421–439.
- (13) So, S.-S.; Karplus, M. Three-dimensional quantitative structure–activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications. *J. Med. Chem.* **1997**, *40*, 4360–4371.
- (14) Good, A. C.; So, S.-S.; Richards, W. G. Structure–activity relationships from molecular similarity matrices. *J. Med. Chem.* **1993**, *36*, 433–438.
- (15) Cruciani, G.; Clementi, S. GOLPE: philosophy and applications in 3D QSAR. In *Advanced Computer-Assisted Techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH Publishers, Inc.: New York, 1994; Vol. 3, pp 61–88.
- (16) Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. *J. Med. Chem.* **1993**, *36*, 2929–2937.
- (17) Benigni, R.; Cotta Ramusino, M.; Giorgi, F.; Gallo, G. Molecular similarity matrices and quantitative structure–activity relationships: a case study with methodological implications. *J. Med. Chem.* **1995**, *38*, 629–635.
- (18) Horwell, D. C.; Howson, W.; Higginbottom, M.; Naylor, D.; Ratcliffe, G. S.; Williams, S. Quantitative structure–activity relationships (QSARs) of N-terminus fragments of NK1 tachykinin antagonists: a comparison of classical QSARs and three-dimensional QSARs from similarity matrices. *J. Med. Chem.* **1995**, *38*, 4454–4462.
- (19) Montanari, C. A.; Tute, M. S.; Beezer, A. E.; Mitchell, J. C. Determination of receptor-bound drug conformations by QSAR using flexible fitting to derive a molecular similarity index. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 67–73.
- (20) Kubinyi, H. Variable selection in QSAR studies. I. An evolutionary algorithm. *Quant. Struct.–Act. Relat.* **1994**, *13*, 285–294.
- (21) So, S.-S.; Karplus, M. Evolutionary optimization in quantitative structure–activity relationship: an application of genetic neural network. *J. Med. Chem.* **1996**, *39*, 1521–1530.
- (22) So, S.-S.; Karplus, M. Genetic neural networks for quantitative structure–activity relationships: improvements and application of benzodiazepine affinity for benzodiazepine/GABA<sub>A</sub> receptors. *J. Med. Chem.* **1996**, *39*, 5246–5256.
- (23) Jain, A. N.; Koile, K.; Chapman, D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315–2327.
- (24) Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.
- (25) Hahn, M.; Rogers, D. Receptor surface models. 2. Application to quantitative structure–activity relationships studies. *J. Med. Chem.* **1995**, *38*, 2091–2102.
- (26) Cho, S. J.; Tropsha, A. Cross-validated R<sup>2</sup>-guided region selection for comparative molecular field analysis: a simple method to achieve consistent results. *J. Med. Chem.* **1995**, *38*, 1060–1066.
- (27) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.
- (28) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-state fields: applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513–520.
- (29) Silverman, B. D.; Platt, D. E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* **1996**, *39*, 2129–2140.
- (30) Oprea, T. I.; Garcia, A. E. Three-dimensional quantitative structure–activity relationships of steroid aromatase inhibitors. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 186–200.
- (31) Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Chemical information in 3D space. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030–1037.
- (32) Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–105.
- (33) Cerius2, version 2.0; Molecular Simulations Inc., San Diego, CA.
- (34) Kim, K. H.; Martin, Y. C. Direct prediction of dissociation constants (pK<sub>a</sub>'s) of clonidine-like imidazoles, 2-substituted imidazoles, and 1-methyl-2-substituted-imidazoles from 3D structures using a comparative molecular field analysis (CoMFA) approach. *J. Med. Chem.* **1991**, *34*, 2056–2060.
- (35) The van der Waals radii are taken from WebElements: <http://www.shef.ac.uk/chemistry/web-elements/web-elements-home.html>.
- (36) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 5832–5842.
- (37) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parametrization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.
- (38) Halgren, T. A. Merck molecular force field. II. van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **1996**, *17*, 520–552.
- (39) Halgren, T. A. Merck molecular force field. III. Molecular geometrics and vibrational frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17*, 553–586.
- (40) Halgren, T. A.; Nachbar, R. B. Merck molecular force field. IV. Conformational energies and geometrics for MMFF94. *J. Comput. Chem.* **1996**, *17*, 587–615.
- (41) Halgren, T. A. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem.* **1996**, *17*, 616–641.
- (42) Holland, J. H. *Adaption in Natural and Artificial Systems*; The University of Michigan Press: Ann Arbor, MI, 1975.
- (43) Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Addison-Wesley Publishing Co.: Redwood City, CA, 1991.
- (44) Möller, M. F. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks* **1993**, *6*, 525–533.
- (45) Andrea, T. A.; Kalayeh, H. Applications of neural networks in quantitative structure–activity relationships of dihydrofolate reductase inhibitors. *J. Med. Chem.* **1991**, *34*, 2824–2836.
- (46) So, S.-S.; Richards, W. G. Application of neural networks: quantitative structure–activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl)pyrimidines as DHFR inhibitors. *J. Med. Chem.* **1992**, *35*, 3201–3207.
- (47) Manallack, D. T.; Ellis, D. D.; Livingstone, D. J. Analysis of linear and nonlinear QSAR data using neural networks. *J. Med. Chem.* **1994**, *37*, 3758–3767.
- (48) Luke, B. T. Evolutionary programming applied to the development of quantitative structure–activity relationships and quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279–1287.
- (49) An arguably better way to perform cross-validation, as suggested by a reviewer, is to run a series of GNN calculations from the beginning with each compound removed in turn. It is unlikely that this computationally demanding procedure will have any effect on the results.
- (50) Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure–activity relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.
- (51) Wold, S.; Eriksson, L. Statistical validation of QSAR results. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH Publishers, Inc.: New York, 1995; Vol. 2, pp 309–318.
- (52) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity – rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3288.

- (53) Rappé, A. K.; Goddard, W. A., III. Charge equilibration for molecular dynamics simulations. *J. Phys. Chem.* **1991**, *95*, 3358–3363.
- (54) Maggiora, G. M.; Elrod, D. W. Computational neural networks as model-free mapping devices. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 732–741.
- (55) Hopfinger, A. J. Theory and application of molecular potential energy fields in molecular shape analysis: a quantitative structure–activity relationship study of 2,4-diamino-5-benzylpyrimidines as dihydrofolate reductase inhibitors. *J. Med. Chem.* **1983**, *26*, 990–996.
- (56) Dunn, J. F.; Nisula, B. C.; Rodbard, D. Transport of steroid hormones: binding of 21 endogenous steroids to both testosterone-binding globulin and corticosteroid-binding globulin in human plasma. *J. Clin. Endocrinol. Metab.* **1981**, *53*, 58–68.
- (57) Pugeat, M. M.; Dunn, J. F.; Nisula, B. C. Transport of steroid hormones: interaction of 70 drugs with testosterone-binding globulin and corticosteroid-binding globulin in human plasma. *J. Clin. Endocrinol. Metab.* **1981**, *53*, 69–75.
- (58) Mickelson, K. E.; Forsthoefel, J.; Westphal, U. Steroid–protein interactions. Human corticosteroid binding globulin: some physicochemical properties and binding specificity. *Biochemistry* **1981**, *20*, 6211–6218.
- (59) Westphal, U. Corticosteroid-binding globulin: a review of some recent aspects. *Mol. Cell. Biochem.* **1983**, *55*, 145–157.

JM970487V