

QSARs for 6-Azasteroids as Inhibitors of Human Type 1 5 α -Reductase: Prediction of Binding Affinity and Selectivity Relative to 3-BHSD

Gregory A. Bakken and Peter C. Jurs*

152 Davey Laboratory, Department of Chemistry, The Pennsylvania State University,
University Park, Pennsylvania 16802

Received April 15, 2001

Quantitative structure–activity relationships (QSARs) are developed to describe the ability of 6-azasteroids to inhibit human type 1 5 α -reductase. Models are generated using a set of 93 compounds with known binding affinities (K_i) to 5 α -reductase and 3 β -hydroxy- Δ^5 -steroid dehydrogenase/3-keto- Δ^5 -steroid isomerase (3-BHSD). QSARs are generated to predict K_i values for inhibitors of 5 α -reductase and to predict selectivity (S_i) of compound binding to 3-BHSD relative to 5 α -reductase. $\text{Log}(K_i)$ values range from -0.70 log units to 4.69 log units, and $\text{log}(S_i)$ values range from -3.00 log units to 3.84 log units. Topological, geometric, electronic, and polar surface descriptors are used to encode molecular structure. Information-rich subsets of descriptors are identified using evolutionary optimization procedures. Predictive models are generated using linear regression, computational neural networks (CNNs), principal components regression, and partial least squares. Compounds in an external prediction set are used for model validation. A 10–3–1 CNN is developed for prediction of binding affinity to 5 α -reductase that produces root-mean-square error (RMSE) of 0.293 log units ($R^2 = 0.97$) for compounds in the external prediction set. Additionally, an 8–3–1 CNN is generated for prediction of inhibitor selectivity that produces RMSE = 0.513 log units ($R^2 = 0.89$) for the external prediction set. Models are further validated through Monte Carlo experiments in which models are generated after dependent variable values have been scrambled.

INTRODUCTION

The androgen dihydrotestosterone (DHT) is related to increased facial and body hair, hair line recession, acne, and enlargement of the prostate.¹ Elevated levels of DHT are related to disorders such as benign prostatic hyperplasia,^{2–5} prostatic cancer,^{5–7} acne,^{8–10} and male pattern baldness.^{9–11} DHT is produced from conversion of testosterone by the enzyme 5 α -reductase. When DHT levels are elevated, there is a corresponding increase in the quantity of the enzyme in affected areas. Therefore, inhibition of 5 α -reductase is of paramount importance in treating disorders related to increased levels of DHT.

Two isozymes, denoted type 1 and type 2, of 5 α -reductase are known.¹² To effectively reduce overall DHT levels, a potent, dual inhibitor of both isozymes would be desirable. Alternatively, two inhibitors could be coadministered, each selectively targeting one form of 5 α -reductase.

Much work has been done to identify potent inhibitors of 5 α -reductase.^{1,13–17} Frye and co-workers examined a series of 6-azasteroids (Figure 1) as dual inhibitors of type 1 and type 2 5 α -reductase.^{13,14} Recently, Kurup and co-workers developed a series of quantitative structure–activity relationships (QSARs) for 5 α -reductase inhibitors.¹⁵ Models were developed for groups of compounds with similar substitution patterns, e.g., a QSAR was developed for a group of 15 compounds with varying substituents at the 4 and 6 positions of the steroid skeleton.

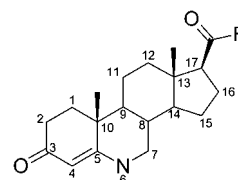


Figure 1. Structure of the 6-azasteroid moiety present in all inhibitors used in this study. Table 2 provides detailed structural information for individual compounds.

In addition to potent inhibition of both isozymes of 5 α -reductase, inhibitor selectivity is also important. To monitor selectivity, Frye and co-workers measured binding affinities for 3 β -hydroxy- Δ^5 -steroid dehydrogenase/3-keto- Δ^5 -steroid isomerase (3-BHSD).^{13,14} Ideal inhibitors are those with low binding affinities to both type 1 and type 2 5 α -reductase and high binding affinity for 3-BHSD.

This work presents the development of QSARs to predict the binding affinity of 6-azasteroids to human type 1 5 α -reductase and to predict selectivity of binding affinity to human adrenal 3-BHSD relative to human type 1 5 α -reductase. The data used are taken from the work of Frye and co-workers.^{13,14} All inhibitors studied were 6-azasteroids (Figure 1). Much of the same data has been used in a recent study to develop QSARs based on steroid substitution pattern.¹⁵ The present work will seek development of single models for compounds with varying substitution patterns. The QSARs developed in this work do not consider binding affinity to type 2 5 α -reductase due to lack of available quantitative data for many of the compounds used. However,

* Corresponding author phone: (814)865-3739; fax: (814)865-3314; e-mail: pcj@psu.edu.

Table 1. Feature Selection and Modeling Building Methodology for the Five Model Types Used

model type	feature selection	model building
1	linear regression	linear regression
2	linear regression	CNN
3	CNN	CNN
4	—	PCR
5	—	PLS

results for type 2 5 α -reductase inhibition would likely be similar to results shown in this paper.

Numerical descriptors were used to encode the structural characteristics of the inhibitors. Information rich subsets of descriptors were examined, and subsets were evaluated with respect to their ability to predict binding affinity and selectivity. Final models were validated with compounds in an external prediction set. Similar methodology has been successfully applied to the prediction of a variety of properties including hydroxyl radical reaction rate constants,¹⁸ inhibition of the sodium ion-proton antiporter,¹⁹ liquid crystal clearing temperature,²⁰ and human intestinal adsorption.²¹ Models developed in this work provide prediction accuracy comparable to that achieved with models developed for small groups of compounds with similar substitution patterns.¹⁵

METHODOLOGY

QSARs were developed using the Automated Data Analysis and Pattern recognition Toolkit (ADAPT) software system,^{22,23} along with feature selection and computational neural network (CNN) routines written in-house. Molecular modeling, feature selection, linear regression, and CNN analysis were performed on a DEC 3000 AXP Model 500 workstation. Principal component regression (PCR) and partial least squares (PLS) routines were written in-house in MATLAB on a 200 MHz Pentium PC with 32 MB of RAM.

In this work, five model types were generated. The model types are summarized in Table 1. Three of the five model types employ linear methods to relate descriptors to binding affinity and two use nonlinear CNNs. Although CNN-based models offer the advantage of being able to model nonlinearities, such models are significantly more computationally intensive than the linear methods. However, the increased computational time is justified if better predictions can be obtained.

Data Set. The compounds used in this study were obtained from the literature.^{13,14} The 93 6-azasteroids with known binding affinities to type 1 5 α -reductase and 3-BHSD are shown in Table 2. Molecular weights ranged from 331 (**75**) to 593 amu (**12**). The dependent variables used in this study were $\log(K_i)$, with K_i measured in nM, and $\log(S_i)$, where S_i denotes the ability to selectively inhibit type 1 5 α -reductase relative to 3-BHSD. The $\log(K_i)$ values for the 93 compounds ranged from -0.70 log units (**23** and **24**) to 4.69 log units (**53**). Compounds with low binding affinities represent potent inhibitors of 5 α -reductase. The $\log(S_i)$ values ranged from -3.00 log units (**42**) to 3.84 log units (**10**). S_i values were computed as the ratio of K_i for 3-BHSD to K_i for type 1 5 α -reductase. Therefore, compounds with high S_i values denote selective inhibitors of type 1 5 α -reductase.

To validate models, a portion of the available data was held out of the model building process. For model types 1,

4, and 5, the training set (TSET) consisted of 84 compounds and the external prediction set (PSET) contained nine compounds. For types 2 and 3, proper CNN training required a cross-validation set (CVSET). Therefore, the original TSET of 84 compounds was further divided into a TSET of 75 compounds and a CVSET of nine compounds. For all model types, the PSET compounds were not used to guide model building but only to validate models. PSET and CVSET compounds were randomly selected with the constraint that dependent variable values for both sets spanned approximately the same range as the TSET compounds for both properties.

Structure Entry and Modeling. Two-dimensional sketches of the compounds were generated using HyperChem (Hypercube, Inc., Waterloo, ON). Information from the sketches, including atom types, bond angles, and bond types, was stored in connection tables. Additionally, estimates of three-dimensional coordinates for all atoms were obtained. The estimated molecular geometries were passed to the semiempirical molecular orbital program MOPAC²⁴ for further refinement. Final geometries were calculated using a PM3 Hamiltonian.²⁵

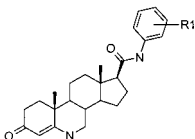
Descriptor Generation. To develop a successful QSAR, it is important to encode the information about the molecular structure responsible for the observed activity. Structure-based descriptors were used in an effort to describe the structure of each compound as completely as possible. A set of 224 descriptors was calculated for each compound. Calculated descriptors were either topological (139), geometric (26), or electronic (11) in nature. Additionally, hybrid descriptors (48) were calculated that combined information from two of these descriptor classes. These hybrid descriptors will be referred to as polar surface descriptors.

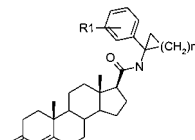
Topological descriptors^{26–31} are generally the least computationally intensive of the descriptor types. A two-dimensional sketch of the compounds is adequate for topological descriptor calculation; no geometry optimization is required. Calculated topological descriptors included kappa indices, molecular connectivity indices, molecular distance edge values, and superpendent indices. Topological descriptors provide information about structural attributes such as atom connectivity, branching, and molecular shape. Additionally, counts of atom types, bond types, etc. are included in this descriptor class.

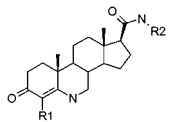
Geometric descriptors^{32,33} require accurate three-dimensional atomic coordinates. Such descriptors are more computationally intensive due to the need for geometry optimized structures. Calculated geometric descriptors included moments of inertia, gravitational index, and molecular surface area and volume. These descriptors attempt to capture information about the three-dimensional structure of the molecule, including size and shape.

Electronic descriptors^{34–36} are used to encode information about electron distribution in the molecule. Some of these descriptors, including the charges on the most positive and most negative atoms, are calculated using an iterative scheme and are not geometry dependent. These electronic descriptors are not very computationally intensive. However, other electronic descriptors do require accurate modeling for proper calculation. Examples of geometry dependent electronic descriptors include dipole moments and frontier molecular orbital energies.

Table 2. Structures and Experimentally Determined K_i and S_i Values for All 93 Compounds Used in the Study

no.	substituents R1		log(<i>K_i</i>) ^c	log(<i>S_i</i>)	ref
					
1			2.38	−1.38	2
2	2- <i>tert</i> -butyl		1.43	0.50	2
3^b	5-chloro, 2- <i>tert</i> -butyl		0.88	2.01	2
4	5-bromo, 2- <i>tert</i> -butyl		0.62	2.29	2
5	4-bromo, 2- <i>tert</i> -butyl		0.72	1.87	2
6	2,5-di- <i>tert</i> -butyl		0.70	2.00	2
7	2- <i>tert</i> -butyl, 5-phenyl		0.66	2.26	2
8	2- <i>tert</i> -butyl, 5-trifluoromethyl		0.94	2.26	2
9	2- <i>tert</i> -butyl, 5-(4-chlorophenyl)		0.53	2.55	2
10	2- <i>tert</i> -butyl, 5-(4- <i>tert</i> -butylphenyl)		0.11	3.84	2
11^a	2,5-bis(trifluoromethyl)		0.60	1.88	2
12	2-(4- <i>tert</i> -butylphenyl), 5-trifluoromethyl		3.04	−1.76	2
13^b	3,5-bis(trifluoromethyl)		1.41	0.35	2
14	3,5-di- <i>tert</i> -butyl		0.90	−0.01	2

no.	substituents		log(<i>K_i</i>) ^c	log(<i>S_i</i>)	ref
	n	R1			
					
15	3	4-chloro	0.83	2.37	2
16^b	4	4-chloro	0.49	2.76	2
17	1	2,4-dichloro	0.83	1.29	2
18	3	4- <i>tert</i> -butyl	0.48	2.70	2
19	4	4- <i>tert</i> -butyl	0.18	3.17	2
20	5	4- <i>tert</i> -butyl	0.56	2.62	2

no.	substituents		log(<i>K_i</i>) ^c	log(<i>S_i</i>)	ref
	R1	R2			
					
21^a	Me	2- <i>tert</i> -butyl-5-(trifluoromethyl)-phenyl	−0.30	2.60	2
22	Cl	2- <i>tert</i> -butyl-5-(trifluoromethyl)-phenyl	−0.22	2.79	2
23	Me	2,5-bis(trifluoromethyl)phenyl	−0.70	1.98	2
24	Cl	2,5-bis(trifluoromethyl)phenyl	−0.70	2.98	2
25^a	Me	<i>N</i> -[1-(4-chlorophenyl)cyclopentyl]	−0.52	2.73	2
26	Cl	<i>N</i> -[1-(4-chlorophenyl)cyclopentyl]	−0.22	2.91	2

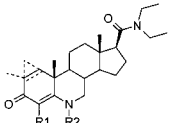
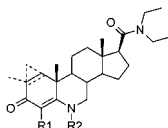
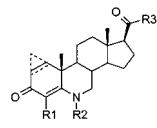
no.	substituents			log(<i>K_i</i>) ^c	log(<i>S_i</i>)	ref
	R1	R2	other			
						
27	H	H		2.88	−1.10	1
28	H	H	Δ ¹	3.76	−1.86	1
29^b	H	H	1,2-α-methano	4.15	−2.17	1
30^b	H	H	2-α,β-Me	3.54	−0.84	1
31	H	COCH ₃		4.60	−0.07	1
32	H	CN		3.92	−0.74	1
33	H	CH ₂ CO ₂ H		4.00	−1.74	1
34	H	Me		2.26	−1.70	1
35	H	Et		3.11	−1.42	1
36	H	Pr		4.38	−1.66	1
37^a	H	<i>iso</i> -Pr		3.53	−0.49	1
38	H	Bu		3.89	−0.34	1
39	H	Hex		3.77	0.27	1

Table 2 (Continued)

no.	substituents			log(<i>K</i> _i) ^c	log(<i>S</i> _i)	ref
	R1	R2	other			
						
40	H	Bn		4.00	0.20	1
41	H	Me	Δ ¹	3.11	−1.77	1
42	H	Me	1,2-α-methano	4.00	−3.00	1
43	Cl	H		1.71	−0.28	1
44	Br	H		1.99	−0.87	1
45	I	H		2.84	−1.66	1
46	CH ₂ NMe ₂	H		4.00	−0.92	1
47	Me	H		1.60	−0.35	1
48 ^a	Me	H	Δ ¹	2.45	−0.49	1
49	Et	H		3.61	−1.94	1
50	Me	Me		1.91	−0.87	1

no.	substituents			log(<i>K</i> _i) ^c	log(<i>S</i> _i)	
	R1	R2	R3			other
						
51	H	H	NH- <i>tert</i> -butyl		2.91	−0.74
52	H	H	NH- <i>tert</i> -butyl	Δ ¹	3.38	−0.04
53	H	H	NH- <i>tert</i> -butyl	1,2-α-methano	4.69	−2.21
54	H	Me	NH- <i>tert</i> -butyl		1.94	−1.12
55	H	Me	NH- <i>tert</i> -butyl	Δ ¹	3.15	−1.35
56 ^a	Me	H	NH- <i>tert</i> -butyl		1.08	−0.12
57	Me	Me	NH- <i>tert</i> -butyl		1.78	−0.42
58	H	H	<i>iso</i> -butyl		0.95	0.05
59 ^b	H	Me	<i>iso</i> -butyl		0.49	−0.45
60	Br	H	<i>iso</i> -butyl		0.51	−0.14
61	Me	H	<i>iso</i> -butyl		−0.40	0.48
62	H	H	NH-1-Ad		1.04	0.85
63 ^b	H	H	NH-1-Ad	Δ ¹	1.87	1.13
64	H	Me	NH-1-Ad		0.93	−0.50
65	H	Me	NH-1-Ad	Δ ¹	2.45	−0.65
66	Br	H	NH-1-Ad		0.65	0.46
67	Me	H	NH-1-Ad		0.04	0.91
68	Br	Me	NH-1-Ad		0.93	−0.02
69	Me	Me	NH-1-Ad		0.79	0.46
70	H	H	NHCHPh ₂		1.48	0.70
71	H	H	NHCHPh ₂	Δ ¹	1.68	1.31
72	H	Me	NHCHPh ₂		0.81	0.24
73	H	Pr	NHCHPh ₂	Δ ¹	2.53	−0.28
74 ^a	Me	H	NHCHPh ₂		0.56	0.70

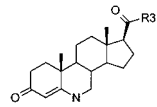
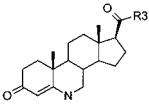
no.	substituents R3	log(<i>K</i> _i) ^c	log(<i>S</i> _i)	ref
				
75	OCH ₃	2.18	−1.10	1
76	O-2-Ad	0.84	1.42	1
77	NMeOMe	3.36	−1.66	1
78 ^a	piperazine	3.93	−1.59	1
79	morpholine	3.34	−1.06	1
80 ^b	thiomorpholine	2.76	−1.26	1
81	piperidine	1.93	0.11	1
82	NHCH(4-fluorophenyl) ₂	1.30	1.20	1
83	NHCH(4-chlorophenyl) ₂	1.30	1.41	1
84	NHNPh ₂	1.15	1.06	1
85 ^b	NOH- <i>tert</i> -butyl	1.58	0.21	1
86	NHCH(cyclohexyl) ₂	1.30	2.30	1
87	NHCPPh ₃	0.91	1.09	1
88	<i>n</i> -Pr	1.08	−0.04	1
89	<i>n</i> -octyl	0.00	0.88	1
90	CH ₂ (cyclohexyl)	0.60	0.60	1

Table 2 (Continued)

no.	substituents R3	$\log(K_i)^c$	$\log(S_i)$	ref
				
91	2,6-difluorophenyl	1.59	-0.55	1
92 ^a	1-naphthyl	1.18	-0.28	1
93	2,4,6-triisopropylphenyl	2.45	0.83	1

^a Cross-validation set member. ^b Prediction set member. ^c K_i values measured in nM.

The final class of descriptors, polar surface descriptors, attempts to combine information from two of the previously defined descriptor classes. Two types of polar surface descriptors were calculated: charged partial surface area³⁷ and hydrogen bonding descriptors.^{38,39} Charged partial surface area descriptors combine information about atomic charges with solvent accessible surface areas. Our charged partial surface area descriptors are closely related to polar surface area descriptors reported in the literature.⁴⁰⁻⁴³ These descriptors provide information about regions of both positive and negative surface area. The same method can be used to encode information about atoms capable of participating in hydrogen bonding. Counts of donor and acceptor atoms are generated for each molecule. Such descriptors describe the ability of a compound to participate in both intermolecular and intramolecular hydrogen bonds.

Objective Feature Selection. As stated previously, 224 descriptors were calculated for each compound. With only 84 compounds in the TSET, selecting descriptor subsets from this large pool would lead to an unacceptably high risk of finding a correlation due to chance. Therefore, it is important to reduce the descriptor pool to a reasonable size with objective means, i.e., without using the dependent variable to guide the elimination of descriptors.

Three steps were taken to objectively eliminate descriptors. First, values for all descriptors were examined, and any descriptor with identical values for more than 85% of the TSET compounds was eliminated. This reduced the descriptor pool from 224 members to 196 members. The next two steps taken to reduce the descriptor pool were used in an effort to reduce correlation among the descriptors. Since PCR and PLS are designed to handle data with inherent correlations present, models formed using these methods were based on the pool of 196 descriptors.

Additional descriptors were eliminated from the pool by removing one of any two descriptors with a pairwise correlation of 0.85 or higher. This step reduced the number of descriptors in the reduced pool to 63. To further reduce the pool, descriptor orthogonality was examined using vector space descriptor analysis, which employs a Gram-Schmidt ranking process.⁴⁴ In short, a single descriptor was randomly selected and designated the basis vector. The stepwise orthogonalization process then proceeded by selecting the descriptor from the 62 remaining that was most orthogonal to the basis vector. These two descriptors were used to define a plane in space. A third vector was selected from the remaining 61 that was most orthogonal to the plane. This process was repeated until all 63 descriptors in the reduced pool were ranked according to orthogonality.

The orthogonalization process in vector space descriptor analysis can be affected by the choice of the descriptor to

serve as the initial basis vector. To reduce the impact of the initial basis vector, the orthogonalization was repeated 10 times, each time with a different randomly selected initial basis vector. Therefore, 10 ordered lists of descriptors were generated. The 40 most orthogonal members were taken from each of the 10 lists. The 10 lists were then compared, and any descriptor not appearing on all 10 lists was eliminated. This final step reduced the pool to 35 descriptors. This pool of 35 descriptors was then passed to the evolutionary optimization algorithms for model formation.

Model Formation and Validation. Five methods were used to generate predictive models. All models were developed using the compounds in the TSET or the TSET and CVSET in the case of CNN models. Once models were finalized, predictions were performed on the PSET compounds to validate the model. Prediction errors for the PSET should be similar to TSET errors. If PSET errors are significantly larger than TSET errors, this means that the model does not generalize well and would not be the best alternative for future predictions.

Type 1. To generate type 1 models, the reduced descriptor pool was screened using simulated annealing.⁴⁵⁻⁴⁷ Descriptor subsets were examined to determine if they could successfully predict K_i (S_i) based on linear regression. For each model examined, root-mean-square error (RMSE) was calculated for the TSET compounds. Additionally, T-values were calculated for the descriptor coefficients to ensure that the coefficients were significantly different than zero. Models were checked for multicollinearities by calculating variance of inflation factors (VIFs) for each descriptor in the model. VIFs are calculated by regressing each descriptor against all others in the model.⁴⁸ According to the relation $VIF = [1 - R^2]^{-1}$, a VIF less than 10 indicates that R , the multiple correlation coefficient, is less than 0.95. Hence, models were deemed to be free of multicollinearities if all VIFs were less than 10. Models of various size were examined, starting with very few descriptors, and increasing model size by one descriptor until RMSE no longer significantly decreased. Once the optimal model was identified, K_i (S_i) values were calculated for the PSET compounds.

Type 2. A type 2 model was generated by using the descriptors identified as optimal for the type 1 model as inputs to a CNN. The CNN used was a three-layer (input-hidden-output), fully connected, feed-forward network.⁴⁹ The number of neurons in the input layer was determined by the model size for the type 1 model, and the output layer contained one neuron representing K_i (S_i). The number of hidden layer neurons was varied to find the optimal network architecture. The optimal network architecture contained the fewest hidden layer neurons without compromising network performance in predicting K_i (S_i) values for the TSET and

CVSET compounds. In general, the number of hidden layer neurons should be kept small so the number of adjustable parameters (weights and biases) does not exceed half the number of TSET compounds. The number of adjustable parameters is computed as $AP = IL \cdot HL + HL \cdot OL + HL + OL$, where AP represents the number of adjustable parameters and IL, HL, and OL denote the number of neurons in the input, hidden, and output layers, respectively. Too many adjustable parameters can lead to a CNN not capable of generalizing to compounds not used in training.

CNN training involved adjusting weights and biases to minimize RMSE for TSET compounds. The quasi-Newton BFGS (Broyden-Fletcher-Goldfarb-Shanno) method was used to train the networks.⁵⁰ The nature of the training process resulted in the RMSE of the TSET continually decreasing toward zero. Therefore, the RMSE for the CVSET was periodically calculated. Training was stopped when the RMSE for the CVSET reached a minimum. This procedure was used to identify the point at which the network begins to lose the ability to generalize. Further training would represent memorization of idiosyncrasies of the TSET compounds instead of general data characteristics.

Network training is typically dependent on the starting weights and biases. To ensure that CNN models produced reliable results, a committee of networks was used to generate final predictions. That is, 10 separate networks, with 10 unique sets of starting weights and biases, were trained. The predicted K_i (S_i) values from the 10 networks were averaged to produce a single estimate of K_i (S_i). The average K_i (S_i) values were used to compute RMSE and R values for type 2 models.

Type 3. Type 3 models were generated using nonlinear methods for feature selection and model building. Candidate subsets of descriptors were identified using simulated annealing^{45–47} and genetic algorithm^{51,52} feature selection routines. Instead of using linear regression for fitness evaluation as was done for type 1 models, a CNN was used. Identified subsets were then used to fully train a committee of CNNs, as described above.

Fitness evaluation of descriptor subsets was accomplished by training a single CNN and computing a COST value. That is, no network averaging was done during the descriptor selection phase of type 3 model formation. The COST function was defined as $COST = TSET_{RMSE} + 0.4 \cdot |TSET_{RMSE} - CVSET_{RMSE}|$, where $TSET_{RMSE}$ and $CVSET_{RMSE}$ denote the RMSE for the training and cross-validation sets, respectively. This COST function helped identify descriptor subsets that produce low RMSEs similar in magnitude for the TSET and CVSET. Previous work has shown that such descriptor subsets are likely to generalize well to compounds not used in training.

Type 4. For type 4 models, the pool of 196 descriptors remaining after removal of descriptors with greater than 85% identical values was used. PCR is designed to work in situations where collinearity exists among independent variables, so further reduction of the pool should not be necessary. Details of PCR can be found in the literature.^{53,54} In short, the matrix of descriptors is decomposed into two matrices of principal components or factors. The factors are ranked according to the amount of variance for which they account. Estimated values for the property of interest can be obtained using a given number of factors. As with other

Table 3. Comparison of $\log(K_i)$ Prediction Results for the Five Model Types Investigated

type	TSET		CVSET		PSET	
	RMSE ^a	R ²	RMSE ^a	R ²	RMSE ^a	R ²
1 (10 descriptors)	0.629	0.80	—	—	0.661	0.73
2 (10–3–1 CNN)	0.315	0.95	0.346	0.97	0.293	0.97
3 (10–3–1 CNN)	0.259	0.97	0.207	0.98	0.486	0.86
4 (19 factors)	0.570	0.84	—	—	0.465	0.86
4 (41 factors)	0.405	0.92	—	—	0.434	0.89
5 (7 factors)	0.516	0.87	—	—	0.574	0.79

^a RMSE values are expressed in log units.

models, RMSE and R values were calculated to determine the goodness of fit of the calculated values relative to the actual values.

Two PCR models, with differing number of factors, were examined. The first model was formed by determining the number of factors necessary to minimize RMSE using a leave-one-out cross-validation of the TSET compounds. That is, one TSET compound was held out, and the remaining compounds were used to build a one-factor model. RMSE was computed for the held-out compound. This was repeated so each TSET compound was held out. This process was then repeated for models with successively more factors. The number of factors necessary for model formation was determined by the minimum RMSE. A second PCR model was generated using the number of factors necessary to account for 95% of the data variance.

Type 5. Type 5 models were developed similarly to type 4 models. The descriptor pool was not reduced to eliminate collinearity since, like PCR, PLS is designed to work with such data. Details of the PLS algorithm used are available in the literature.^{53,54} As with PCR, PLS provides a means of decomposing the descriptor matrix into two matrices of factors. Leave-one-out cross-validation was used with the TSET compounds to determine the number of factors optimal for PLS model formation. RMSE and R values were used to assess model accuracy.

RESULTS AND DISCUSSION

Prediction of K_i Values. Simulated annealing was used to screen descriptor subsets of various size. The best type 1 linear model contained 10 descriptors. All descriptor T-values were greater than 3, and no VIF exceeded 10, indicating that no multicollinearities were present. For the 84 TSET compounds, the model produced $RMSE = 0.629$ log units and $R^2 = 0.80$, where R is the correlation coefficient obtained by comparing calculated K_i values to actual K_i values. This model produced $RMSE = 0.661$ log units and $R^2 = 0.73$ for the nine PSET compounds (see Table 3).

The 10 descriptors chosen for the type 1 model are shown in Table 4. Six of the descriptors are topological, one geometric, one electronic, and two are polar surface descriptors. Pairwise correlations among the descriptors range from 0.005 to 0.386, with an average of 0.127.

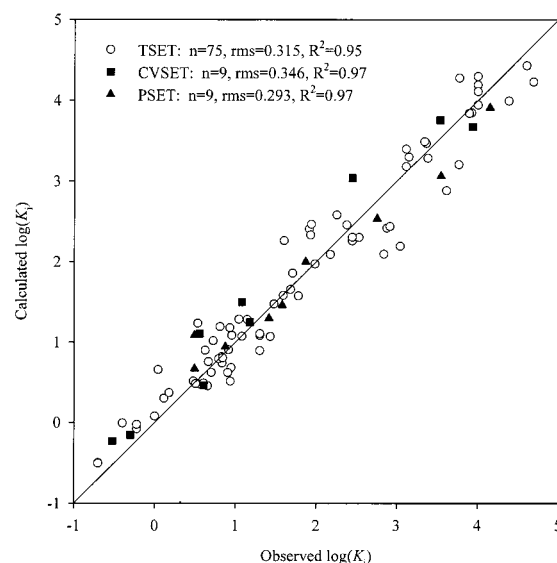
Although there is no causal relationship between the descriptors chosen and the K_i values, the information probably encoded by each descriptor can be surmised. The two molecular connectivity indices, V6C11 and N6CH16, denote the sixth-order valence corrected connectivity index for clusters ($^6\chi_c$) and the number of sixth-order chains,

Table 4. Ten-Descriptor Type 1 Linear Regression Model Selected by Simulated Annealing for Prediction of $\log(K_i)$

descriptor ^a	coefficient	error estimate	range	type
V6C11	-1.10E+01	2.19	0.14–0.33	topological
N6CH16	1.13E-01	3.24E-02	7–21	topological
NN4	7.28E-01	1.96E-01	1–3	topological
WTPT4	3.16E-01	9.20E-02	5.37–10.24	topological
2SP3	-2.19E-01	3.71E-02	4–16	topological
MDE33	-7.63E-03	2.53E-03	6.07–156.30	topological
GEOM6	1.86E-01	5.86E-02	1.31–8.87	geometric
DPOL	2.70E-01	6.88E-02	5.63–11.93	electronic
DPSA1	5.70E-03	7.14E-04	-39.97–579.50	polar surface
RNCS1	3.80E-01	8.49E-02	5.04–10.75	polar surface
constant	-6.26	1.30		

^a Explanation: V6C11, sixth-order valence corrected connectivity index for clusters (χ_6);²⁸ N6CH16, number of sixth-order chains;²⁸ NN4, number of nitrogen atoms; WTPT4, sum of all path weights starting from oxygen atoms;²⁶ 2SP3, number of sp^3 hybridized carbon atoms bonded to two other carbon atoms and two hydrogen atoms; MDE33, molecular distance edge term between pairs of tertiary carbons;²⁷ GEOM6, ratio of molecular width to thickness; DPOL, electric dipole moment; DPSA1, difference between partial positive surface area and partial negative surface area;³⁷ RNCS1, relative negatively charged surface area [RNCS1 = (surface area of most negative atom)(charge of most negative atom)/(total negative charge)].³⁷

respectively.²⁸ These two descriptors encode the size and branching of the compounds. NN4 simply provides the number of nitrogen atoms. All compounds in Table 2 are 6-azasteroids and therefore have a nitrogen at position 6 of the steroid skeleton. For 12 compounds, this is the only nitrogen present. The remaining 81 compounds have an additional nitrogen linked to the carbonyl group attached at position 17. Four of these 81 have a third nitrogen in one of the substituted portions. It may be that NN4 serves as an indicator regarding substitution of the carbonyl group attached to position 17. The descriptor WTPT4 denotes the sum of all path weights starting from oxygen atoms.²⁶ The number and position of oxygen atoms are encoded here. 2SP3 provides a count of sp^3 hybridized carbons bonded to two other carbon atoms and two hydrogen atoms. This descriptor captures information about carbons in the steroid skeleton (specifically substitution or unsaturation at positions 1 and 2) and carbons in side chains. The descriptor MDE33 represents the distance edge term between pairs of tertiary carbons, where a tertiary carbon is defined as a carbon with three bonds to non-hydrogen atoms.²⁷ This descriptor probably encodes information about branching as well as complementing the information provided by 2SP3 about the steroid skeleton. 2SP3 and MDE33 provide information about unsaturation or substitution at position 2 of the steroid skeleton which was found by Frye and co-workers to influence the potency of inhibitors.¹⁴ The geometric descriptor GEOM6 encodes information about molecular size and shape. This descriptor represents the ratio of the width of the compound to the thickness. DPOL1 represents the dipole moment and provides information about the ability of the compound to interact with other polar molecules. The two polar surface descriptors, DPSA1 and RNCS1, denote the difference between partial positive and negative surface area and the charge weighted percent surface area of the most negatively charged atom, respectively.³⁷ Both descriptors further encode the ability of the compound to interact with other molecules through polar interactions. Interestingly, the

**Figure 2.** Plot of calculated versus observed $\log(K_i)$ values for the training, cross-validation, and prediction set compounds using the type 2 model.

carbonyl oxygen at position 3 in the steroid skeleton is the most negatively charged atom for 92 of the 93 compounds in the data set. Analysis of this descriptor reveals that it is providing information about the solvent accessible surface area for the carbonyl oxygen. One of the key factors affecting the accessible surface area is substitution at the 4 position of the steroid skeleton. Frye and co-workers found that small groups at the 4 position increase the potency of inhibitors,¹⁴ and RNCS1 is describing relevant information about substitution at the 4 position.

The 10 descriptors forming the optimal type 1 model were used as input to a fully connected, feed-forward, three-layer CNN to generate a type 2 model. Initial network architecture was 10–2–1; however, addition of a hidden layer neuron to form a 10–3–1 network significantly reduced RMSE for the TSET and CVSET. This network had 37 adjustable parameters, which is very close to half the number of TSET observations (75). Therefore, no additional hidden layers neurons could safely be added, and a 10–3–1 network architecture was selected as optimal.

The TSET compounds were used to adjust weights and biases to minimize RMSE during the CNN training phase. The CVSET RMSE was computed periodically and used to determine the appropriate point to stop network training. Predicted K_i values were determined as the average of 10 trained CNNs. As shown in Table 3, the type 2 model produced TSET RMSE = 0.315 log units and $R^2 = 0.95$ and CVSET RMSE = 0.346 log units and $R^2 = 0.97$. The PSET RMSE was 0.293 log units ($R^2 = 0.97$), which demonstrates that the type 2 model can generalize to compounds not used in training. The type 2 model represents substantially improved predictions relative to the type 1 model.

A plot of calculated versus observed K_i values for the type 2 model is shown in Figure 2. Recall that the calculated values represent the average K_i values for 10 fully trained CNNs. As Figure 2 shows, most of the points fall very close to the one-to-one correlation line. More importantly, the PSET compounds behave similarly to the TSET and CVSET

compounds, indicating the potential utility and reliability of this model in future predictions.

To generate type 3 models, both simulated annealing and genetic algorithm feature selection routines were employed to select subsets of 10 descriptors. The use of CNNs as fitness evaluators is computationally intensive. In this case, subsets identified using simulated annealing produced lower RMSE for the TSET and CVSET. Therefore, the 10 descriptors used to form the optimal type 3 model were those identified as best using simulated annealing.

As with the type 2 model, the 10 descriptors were used to train ten CNNs from different starting points. Table 3 shows that the type 3 model produced TSET RMSE = 0.259 log units ($R^2 = 0.97$) and CVSET RMSE = 0.207 log units ($R^2 = 0.98$). This represents an improvement over the type 2 model. However, model validation is suspect with PSET RMSE = 0.486 log units ($R^2 = 0.86$). This somewhat poor validation may be related to the small body of available training data. A general lack of validation was observed for descriptor subsets identified for type 3 model formation. This finding suggests that, for small data sets, the use of evolutionary optimization with a CNN fitness evaluator may select descriptor subsets prone to over-training in CNN model formation. The CVSET, which is not involved in the descriptor selection phase, may not be adequate to ensure that type 3 models are capable of generalizing. However, the present work does not provide definitive evidence that such a conclusion can be drawn.

To form type 4 models, PCR was performed using the pool of 196 descriptors. The data were autoscaled prior to PCR such that the average and standard deviation were 0 and 1, respectively, over all TSET compounds for each descriptor. Using leave-one-out cross-validation of the TSET compounds, a model of 19 factors (79.9% variance) was selected. For this model, TSET RMSE = 0.570 log units ($R^2 = 0.84$) and PSET RMSE = 0.465 log units ($R^2 = 0.86$). As shown in Table 3, this model is better than the type 1 model but demonstrates less predictive ability than the type 2 model.

A second PCR model was formed by considering the importance of capturing the majority of the data variance. A model with 41 factors was generated since this accounted for 95.1% of the data variance. This 41-factor PCR model produced a TSET RMSE = 0.405 log units and $R^2 = 0.92$. This model generalized well as shown by the PSET RMSE = 0.434 log units and $R^2 = 0.89$.

Clearly, the 41-factor PCR model is superior to the 19-factor PCR model and to the type 1 model. However, the large number of factors required for this model suggests that PCR may not be the most appropriate analysis method. Generally, PCR is best used in situations involving high collinearity among the independent variables. The large number of factors required indicates that these data are not very well suited for PCR analysis.

As with type 4 models, the pool of 196 descriptors was autoscaled prior to generation of type 5 PLS models. Leave-one-out cross-validation of the TSET compounds indicated that a seven-factor model was appropriate. This finding suggested that the PLS decomposition is better able to capture relevant information from the data compared to PCR.

The seven-factor PLS model produced TSET RMSE = 0.516 log units and $R^2 = 0.87$. For the PSET, RMSE = 0.574

Table 5. Comparison of Monte Carlo $\log(K_i)$ Prediction Results for the Five Model Types Investigated^b

type	TSET		CVSET		PSET	
	RMSE ^a	SD	RMSE ^a	SD	RMSE ^a	SD
1 (10 descriptors)	1.184	0.044	—	—	1.602	0.305
2 (10–3–1 CNN)	0.768	0.094	0.915	0.219	1.600	0.211
3 (10–3–1 CNN)	0.661	0.110	0.856	0.140	1.548	0.260
4 (19 factors)	1.239	0.047	—	—	1.490	0.301
4 (41 factors)	1.019	0.070	—	—	1.821	0.436
5 (7 factors)	1.103	0.045	—	—	1.546	0.273

^a RMSE values are expressed in log units. ^b The RMSE is the average of ten scrambling experiments, and SD is the standard deviation of the ten experiments.

log units and $R^2 = 0.79$. Table 3 demonstrates that this type 5 model is better than the type 1 model, equivalent to the type 4 19-factor model, and inferior to the type 4 41-factor model in terms of RMSE and R values for the TSET. However, the PLS model requires far fewer factors, thus demonstrating that this methodology is probably more suitable than PCR for the present application.

Monte Carlo Experiments. Monte Carlo experiments were conducted to verify that results were not due to chance. In each Monte Carlo experiment, the same set of independent variables was used. The dependent variable, $\log(K_i)$, was randomly scrambled. Evolutionary optimization procedures were used to select subsets of descriptors best able to reproduce the scrambled K_i values. The randomization experiments were repeated 10 times, each time with a different scrambling of the dependent variable. RMSE values were averaged over the 10 experiments for comparison with models previously developed.

Table 5 shows the results of the Monte Carlo experiments for all five model types. As can be seen, the RMSE values for all model types are significantly higher than the analogous models using the actual dependent variable. Additionally, plots of calculated versus observed K_i values for the Monte Carlo experiments confirm the lack of correlation (results not shown). These results demonstrate that the original models formed are not due to the finding of chance correlations.

Prediction of S_i Values. The five model types shown in Table 1 were used to generate models for the prediction of S_i values. For type 1 models, descriptor subsets of various size were screened with simulated annealing. The optimal linear model contained eight descriptors. For all model coefficients, T -values were greater than 3. Additionally, all VIFs among the descriptors were less than 10, demonstrating that the model was free of multicollinearities. As shown in Table 6, the RMSE for the 84 TSET compounds was 0.796 log units and $R^2 = 0.73$. For the nine PSET compounds, RMSE = 0.781 log units and $R^2 = 0.73$, demonstrating the ability of the model to generalize to compounds not involved model formation.

Table 7 shows the eight descriptors selected to form the type 1 model. Interestingly, 5 of the descriptors are topological, 3 are polar surface, and no geometric or electronic descriptors were chosen. Pairwise correlations for the descriptors ranged from 0.001 to 0.785, with an average value of 0.232.

The first two descriptors in Table 7, V6C11 and MOLC7, denote molecular connectivity indices for clusters.²⁸ These

Table 6. Comparison of $\log(S_i)$ Prediction Results for the Five Model Types Investigated

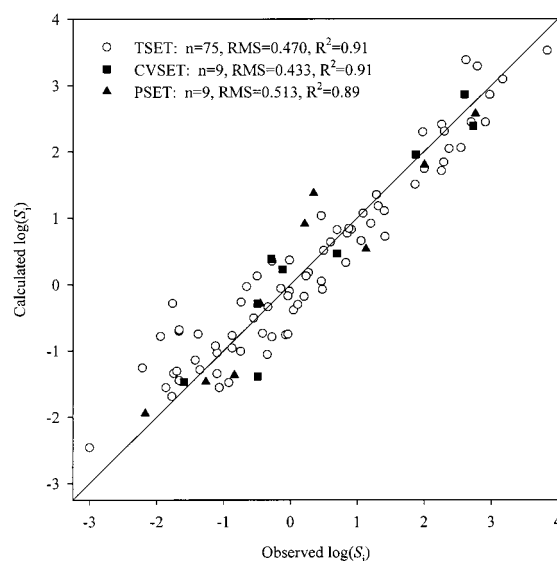
type	TSET		CVSET		PSET	
	RMSE ^a	R ²	RMSE ^a	R ²	RMSE ^a	R ²
1 (8 descriptors)	0.796	0.73	—	—	0.781	0.73
2 (8–3–1 CNN)	0.470	0.91	0.433	0.91	0.513	0.89
3 (8–3–1 CNN)	0.412	0.93	0.380	0.93	0.610	0.84
4 (17 factors)	0.741	0.76	—	—	0.653	0.85
4 (41 factors)	0.543	0.87	—	—	0.692	0.90
5 (4 factors)	0.718	0.78	—	—	0.783	0.79

^a RMSE values are expressed in log units.**Table 7.** Eight-Descriptor Type 1 Linear Regression Model Selected by Simulated Annealing for Prediction of $\log(S_i)$

descriptor ^a	coefficient	error estimate	range	type
V6C11	1.46E+01	3.92	0.14–0.33	topological
MOLC7	−9.02E−01	2.96E−01	2.00–4.90	topological
1SP3	6.51E−01	1.40E−01	2–8	topological
2SP3	5.40E−01	6.41E−02	4–16	topological
MDE33	1.67E−02	3.61E−03	6.07–156.30	topological
DPSA1	−8.73E−03	1.28E−03	−39.97–579.50	polar surface
CHDH3	3.32E+03	8.71E+02	0–5.48E−04	polar surface
SCDH2	−2.71E−01	8.70E−02	0–8.54	polar surface
constant	−3.99	0.55		

^a Explanation: V6C11, sixth-order valence corrected connectivity index for clusters (χ_{C});²⁸ MOLC7, molecular connectivity index for third-order clusters;²⁸ 1SP3, number of sp^3 hybridized carbon atoms bonded to one other carbon atom and three hydrogen atoms; 2SP3, number of sp^3 hybridized carbon atoms bonded to two other carbon atoms and two hydrogen atoms; MDE33, molecular distance edge term between pairs of tertiary carbons;²⁷ DPSA1, difference between partial positive surface area and partial negative surface area;³⁷ CHDH3, $[\text{CHDH3} = (\text{sum of charges on donatable hydrogens})/(\text{total molecular surface area})]$;^{38,39} SCDH2, $[\text{SCDH2} = (\text{sum of (surface area} \times \text{charge)}) \text{ on donatable hydrogens})/(\text{total number of donatable hydrogens})]$.^{38,39}

two descriptors provide information about size and branching. The next two descriptors, 1SP3 and 2SP3, provide counts of sp^3 hybridized carbons in slightly different bonding environments. The descriptor 1SP3 describes sp^3 carbons bonded to one carbon and three hydrogens, and 2SP3 signifies sp^3 carbons bonded to two carbons and two hydrogens. 1SP3 provides a count of the methyl groups always present at positions 10 and 15 of the steroid skeleton (Figure 1) and sometimes present as part of substituent groups. 2SP3 provides a count of methylene groups always present at positions 11, 12, 15, and 16 (Figure 1), often present at positions 1 and 2, and sometimes present as part of substituent groups. MDE33 denotes the distance edge term between pairs of tertiary carbons.²⁷ This descriptor provides information about positions 8, 9, 14, and 17 (Figure 1) and positions 2 and 4 with appropriate substitution as well as tertiary carbons in substituent groups. These two descriptors are describing the trend observed by Frye and co-workers¹⁴ which suggested substitution or unsaturation at the 2 position directly affects selectivity of inhibitors. The final three descriptors in Table 7 are polar surface descriptors.^{37–39} DPSA1 encodes the difference between the partial positive and partial negative surface area for each compound. This descriptor provides information about the ability of a compound to interact with other polar compounds. The final two descriptors, CHDH3 and SCDH2, provide information about hydrogens available for hydrogen bonding. CHDH3

**Figure 3.** Plot of calculated versus observed $\log(S_i)$ values for the training, cross-validation, and prediction set compounds using the type 2 model.

is the summation of charges on all donatable hydrogens divided by the total molecular surface area. SCDH2 denotes the summation of charge weighted surface areas of all donatable hydrogens divided by the total number of donatable hydrogens. These two descriptors encode information about the ability of compounds to hydrogen bond and interact with other polar compounds.

The eight descriptors from the type 1 model were used to train fully connected feed-forward neural networks to form type 2 models. An 8–3–1 network architecture was selected as optimal. This network architecture contains 31 adjustable parameters. Ten networks were trained, and calculated S_i values from each network were averaged to produce predictions. For the TSET and CVSET, RMSE = 0.470 log units ($R^2 = 0.91$) and RMSE = 0.433 log units ($R^2 = 0.91$), respectively. As shown in Table 6, the model generalized well to PSET compounds (RMSE = 0.513 log units, $R^2 = 0.89$). Results for the type 2 model represent a significant improvement over the type 1 model.

A plot of calculated versus observed S_i values is shown in Figure 3. All compounds lie reasonably close to the one-to-one correlation line. Calculated S_i values tend to be high toward the lower left of the plot. This behavior is seen frequently with neural networks. However, it is important to note that there is no appearance of consistently low calculated S_i values at the upper right of the plot. This is the region where the compounds with highest selectivity appear. Thus, compounds with the most promising selectivity are likely to be more accurately predicted than less selective compounds.

Evolutionary optimization with a CNN fitness evaluator was used to select eight descriptors for a type 3 model. The optimal subset of eight descriptors was selected using simulated annealing with an 8–3–1 network architecture. Ten CNNs were trained from different starting points, and predictions from the 10 networks were averaged to produce calculated S_i values. TSET RMSE was 0.412 log units ($R^2 = 0.93$), and CVSET RMSE was 0.380 log units ($R^2 = 0.93$). However, PSET RMSE was 0.610 log units ($R^2 = 0.84$). As with the type 3 model for K_i prediction, validation with

Table 8. Comparison of Monte Carlo $\log(S_i)$ Prediction Results for the Five Model Types Investigated^b

type	TSET		CVSET		PSET	
	RMSE ^a	SD	RMSE ^a	SD	RMSE ^a	SD
1 (8 descriptors)	1.331	0.050	—	—	1.686	0.309
2 (8–3–1 CNN)	0.959	0.105	1.228	0.330	1.705	0.232
3 (8–3–1 CNN)	0.868	0.091	1.051	0.133	1.747	0.494
4 (17 factors)	1.378	0.059	—	—	1.648	0.318
4 (41 factors)	1.126	0.061	—	—	1.965	0.371
5 (4 factors)	1.332	0.055	—	—	1.683	0.289

^a RMSE values are expressed in log units. ^b The RMSE is the average of ten scrambling experiments, and SD is the standard deviation of the ten experiments.

this type 3 model was somewhat suspect. This behavior was seen consistently for all type 3 models investigated. Again, the poor validation for the type 3 model may be due to the somewhat small body of available training data.

Two type 4 models were formed; one containing the number of factors deemed appropriate using leave-one-out cross-validation and one using all factors necessary to account for 95% of the data variance. Using leave-one-out cross-validation, 17 factors (76.9% variance) were selected. With this model, TSET RMSE = 0.741 log units ($R^2 = 0.76$) and PSET RMSE = 0.653 log units ($R^2 = 0.85$). To account for 95% of data variance, 41 factors are necessary. (This is the same number as for K_i predictions since PCR involves only the matrix of independent variables, and this matrix is the same for both properties). The 41-factor PCR model produced TSET RMSE = 0.543 log units ($R^2 = 0.87$) and PSET RMSE = 0.692 log units ($R^2 = 0.90$). Both models represent slight improvements relative to the type 1 model but do not provide the accuracy that can be achieved with CNN models.

The type 5 PLS model required only four factors, as determined via leave-one-out cross-validation of the TSET compounds. Table 6 shows that TSET RMSE = 0.718 log units ($R^2 = 0.78$) and PSET RMSE = 0.783 log units ($R^2 = 0.79$) for this model. Again, PLS requires far fewer factors than PCR. However, for this property, PLS does not perform as well on the external prediction set of compounds. The PLS model is roughly equivalent to the type 1 model in this case, offering no apparent advantage over linear regression.

Monte Carlo Experiments. Monte Carlo experiments were used to further validate the models obtained for prediction of selectivity. The dependent variable was randomly scrambled 10 times. Each time, the original independent variables were used to generate predictive models. Table 8 shows the Monte Carlo results for the five model types. For the previously developed models (Table 6), PSET RMSE varied from 0.513 to 0.783 log units. For the Monte Carlo models (Table 8), PSET RMSE varied from an average of 1.648 to 1.965 log units. The Monte Carlo results provide clear evidence that the original models developed were not modeling a chance relationship.

CONCLUSIONS

Models have been generated which allow accurate prediction of binding affinity for inhibitors of human type 1 5α -reductase based on structural descriptors. Additional models have been developed which allow prediction of selective

inhibition of human type 1 5α -reductase relative to 3-BHSD. These models would allow compound libraries to be screened in an effort to find new candidates for treatment of disorders related to increased production of DHT.

Five model types were generated for prediction of K_i and S_i values. Three of the model types were linear and two involved nonlinear CNNs. Of the linear models, the type 1 linear regression models are more suitable than the type 4 PCR or type 5 PLS models. Tables 3 and 6 show that type 1 RMSE values are generally somewhat larger than corresponding errors for type 4 and 5 models. However, once developed, type 1 models offer a larger (though still limited) degree of interpretability and are computationally less demanding. Type 4 and 5 models require calculation of all 196 descriptors, while only 8 or 10 descriptors are necessary for type 1 models.

Models developed with CNNs offer the advantage of reduced errors at the cost of increased training time. However, once trained, these networks provide predictions on a time scale similar to the other model types. The optimal model for both properties discussed has been provided by the type 2 CNN models. Figures 2 and 3 demonstrate that these models provide highly accurate predictions.

The models developed in this work compare favorably with existing models developed for various series of compounds with similar substitution patterns.¹⁵ Models developed in this work offer the advantage of requiring only a single model for all future predictions. However, developing models for compounds based on substitution patterns may allow some insight into important structural features of specific compound types as well as important structural features common to all inhibitors.

REFERENCES AND NOTES

- (1) Singh, S. M.; Gauthier, S.; Labrie, F. Androgen Receptor Antagonists (Antiandrogens): Structure–Activity Relationships. *Curr. Med. Chem.* **2000**, *7*, 211–247.
- (2) Clifford, G. M.; Farmer, R. D. T. Medical Therapy for Benign Prostatic Hyperplasia: A Review of the Literature. *Eur. Urol.* **2000**, *38*, 2–19.
- (3) Bartsch, G.; Rittmaster, R. S.; Klocker, H. Dihydrotestosterone and the Concept of 5α -Reductase Inhibition in Human Benign Prostatic Hyperplasia. *Eur. Urol.* **2000**, *37*, 367–380.
- (4) Weisser, H.; Tunn, S.; Debus, M.; Krieg, M. 5α -Reductase inhibition by finasteride (Proscar) in epithelium and stroma of human benign prostatic hyperplasia. *Steroids* **1994**, *59*, 616–620.
- (5) Tolman, R. L.; Sahoo, S. P.; Bakshi, R. K.; Gratale, D.; Patel, G.; Patel, S.; Toney, J.; Chang, B.; Harris, G. S. 4-Methyl-3-oxo-4-aza- 5α -androst-1-ene-17 β -N-aryl-carboxamides: an Approach to Combined Androgen Blockade [5α -Reductase Inhibition with Androgen Receptor Binding In Vitro]. *J. Steroid Biochem. Mol. Biol.* **1997**, *60*, 303–309.
- (6) Njar, V. C. O.; Kato, K.; Nnane, I. P.; Grigoryev, D. N.; Long, B. J.; Brodie, A. M. H. Novel 17-Azoly Steroids, Potent Inhibitors of Human Cytochrome 17α -Hydroxylase- $C_{17,20}$ -lyase (P450_{17 α}): Potential Agents for the Treatment of Prostate Cancer. *J. Med. Chem.* **1998**, *41*, 902–912.
- (7) Ling, Y.; Li, J.; Liu, Y.; Kato, K.; Klus, G. T.; Brodie, A. 17-Imidazolyl, Pyrazolyl, and Isoxazolyl Androstene Derivatives. Novel Steroidal Inhibitors of Human Cytochrome $C_{17,20}$ -Lyase (P450_{17 α}). *J. Med. Chem.* **1997**, *40*, 3297–3304.
- (8) Thiboutot, D.; Bayne, E.; Thorne, J.; Gilliland, K.; Flanagan, J.; Shao, Q.; Light, J.; Helm, K. Immunolocalization of 5α -Reductase Isozymes in Acne Lesions and Normal Skin. *Arch. Dermatol.* **2000**, *136*, 1125–1129.
- (9) Guarna, A.; Occhiato, E. G.; Banzo, G.; Conti, A.; Serio, M. 5α -reductase inhibitors, chemical and clinical models. *Steroids* **1998**, *63*, 355–361.
- (10) Guarna, A.; Machetti, F.; Occhiato, E. G.; Scarpi, D. Benzo[c]-quinolizin-3-ones: A Novel Class of Potent and Selective Nonsteroidal

- Inhibitors of Human Steroid 5 α -Reductase 1. *J. Med. Chem.* **2000**, *43*, 3718–3735.
- (11) Hogan, D. J.; Chamberlain, M. Male Pattern Baldness. *South. Med. J.* **2000**, *93*, 657–662.
 - (12) Russell, D. W.; Wilson, J. D. Steroid 5 α -reductase: two genes/two enzymes. *Annu. Rev. Biochem.* **1994**, *63*, 25–61.
 - (13) Frye, S. V.; Haffner, C. D.; Maloney, P. R.; Hiner, R. N.; Dorsey, G. F.; Noe, R. A.; Unwalla, R. J.; Batchelor, K. W.; Bramson, H. N.; Stuart, J. D.; Schweider, S. L.; van Arnold, J.; Bickett, D. M.; Moss, M. L.; Tian, G.; Lee, F. W.; Tippin, T. K.; James, M. K.; Grizzle, M. K.; Long, J. E.; Croom, D. K. Structure–Activity Relationships for Inhibition of Type 1 and 2 Human 5 α -Reductase and Human Adrenal 3 β -Hydroxy- Δ^5 -steroid Dehydrogenase/3-Keto- Δ^5 -steroid Isomerase by 6-Azaandrost-4-en-3-ones: Optimization of the C17 Substituent. *J. Med. Chem.* **1995**, *38*, 2621–2627.
 - (14) Frye, S. V.; Haffner, C. D.; Maloney, P. R.; Mook, R. A.; Dorsey, G. F.; Hiner, R. N.; Cribbs, C. M.; Wheeler, T. N.; Ray, J. A.; Andrews, R. C.; Batchelor, K. W.; Bramson, H. N.; Stuart, J. D.; Schweiker, S. L.; van Arnold, J.; Croom, S.; Bickett, D. M.; Moss, M. L.; Tian, G.; Unwalla, R. J.; Lee, F. W.; Tippin, T. K.; James, M. K.; Grizzle, M. K.; Long, J. E.; Schuster, S. V. 6-Azasteroids: Structure–Activity Relationships for Inhibition of Type 1 and 2 Human 5 α -Reductase and Human Adrenal 3 β -Hydroxy- Δ^5 -steroid Dehydrogenase/3-Keto- Δ^5 -steroid Isomerase. *J. Med. Chem.* **1994**, *37*, 2352–2360.
 - (15) Kurup, A.; Garg, R.; Hansch, C. Comparative QSAR Analysis of 5 α -Reductase Inhibitors. *Chem. Rev.* **2000**, *100*, 909–924.
 - (16) Rasmussen, G. H.; Reynolds, G. F.; Steinberg, N. G.; Walton, E.; Patel, G. F.; Liang, T.; Cascieri, M. A.; Cheung, A. H.; Brooks, J. R.; Berman, C. Azasteroids: Structure–Activity Relationships for Inhibition of 5 α -Reductase and of Androgen Receptor Binding. *J. Med. Chem.* **1986**, *29*, 2298–2315.
 - (17) Hartmann, R. W.; Reichert, M. New Nonsteroidal Steroid 5 α -Reductase Inhibitors. Syntheses and Structure–Activity Studies on Carboxamide Phenylalkyl-Substituted Pyridones and Piperidones. *Arch. Pharm. Pharm. Med. Chem.* **2000**, *333*, 145–153.
 - (18) Bakken, G. A.; Jurs, P. C. Prediction of Hydroxyl Radical Rate Constants from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1064–1075.
 - (19) Kauffman, G. W.; Jurs, P. C. Prediction of Inhibition of the Sodium Ion-Proton Antiporter by Benzoylguanidine Derivatives from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 753–761.
 - (20) Johnson, S. R.; Jurs, P. C. Prediction of the Clearing Temperature of a Series of Liquid Crystals from Molecular Structure. *Chem. Mater.* **1999**, *11*, 1007–1023.
 - (21) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal Adsorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726–735.
 - (22) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
 - (23) Jurs, P. C.; Chow, J. T.; Yuan, M. In *Computer Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; The American Chemical Society: Washington, DC, 1979.
 - (24) Stewart, J. P. P. *Mopac 6.0, Quantum Chemistry Program Exchange*; Indiana University: Bloomington, IN, Program 455.
 - (25) Stewart, J. P. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–105.
 - (26) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *4*, 162–175.
 - (27) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, λ . *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387–394.
 - (28) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Research Studies Press, Ltd.: Hertfordshire, England, 1986.
 - (29) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for All Self-Avoiding Paths for Molecular Graphs. *Comput. Chem.* **1979**, *3*, 5–13.
 - (30) Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1–7.
 - (31) Kier, L. B.; Hall, L. H. The E-State as an Extended Free Valence. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 548–552.
 - (32) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441–451.
 - (33) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4–12.
 - (34) Miller, K. J.; Savchik, J. A. A New Empirical Method to Calculate Average Molecular Polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206–7213.
 - (35) Abraham, R. J.; Smith, P. E. Charge Calculations in Molecular Mechanics IV: A General Method for Conjugated Systems. *J. Comput. Chem.* **1987**, *9*, 288–297.
 - (36) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure–Property Relationships. *J. Comput. Chem.* **1992**, *13*, 492–504.
 - (37) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Descriptors in Computer-Assisted Quantitative Structure–Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323–2329.
 - (38) Vinogradov, S. N.; Linnell, R. H. *Hydrogen Bonding*; van Nostrand Reinhold: New York, 1971.
 - (39) Pimentel, G. I.; McClellan, A. L. *The Hydrogen Bond*; Freeman: San Francisco, 1960.
 - (40) Stenberg, P.; Luthman, K.; Artursson, P. Virtual screening of intestinal drug permeability. *J. Contr. Relat.* **2000**, *65*, 231–243.
 - (41) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Beigi, F.; Lundahl, P.; Artursson, P. Evaluation of Dynamic Polar Molecular Surface Area as Predictor of Drug Absorption: Comparison with Other Computational and Experimental Predictors. *J. Med. Chem.* **1998**, *41*, 5382–5392.
 - (42) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Artursson, P. Correlation of Drug Absorption with Molecular Surface Properties. *J. Pharm. Sci.* **1996**, *85*, 32–39.
 - (43) Clark, D. E. Rapid Calculation of Polar Molecular Surface Area and Its Application to the Prediction of Transport Phenomena. 1. Prediction of Intestinal Absorption. *J. Pharm. Sci.* **1998**, *88*, 807–814.
 - (44) Anton, H. *Elementary Linear Algebra*, 6th ed.; John Wiley & Sons: New York, 1991.
 - (45) Kalivas, J. H.; Roberts, N.; Sutter, J. M. Global Optimization by Simulated Annealing with Wavelength Selection for Ultraviolet–Visible Spectrophotometry. *Anal. Chem.* **1989**, *61*, 2024–2030.
 - (46) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure–Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77–84.
 - (47) Kalivas, J. H. Generalized Simulated Annealing for Calibration Sample Selection from an Existing Set and Orthogonalization of Undesigned Experiments. *J. Chemom.* **1991**, *5*, 37–48.
 - (48) Belsley, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; John Wiley & Sons: New York, 1980.
 - (49) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure–Activity Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841–851.
 - (50) Wessel, M. D.; Jurs, P. C. Prediction of Reduced Ion Mobility Constants from Structural Information Using Multiple Linear Regression Analysis and Computational Neural Networks. *Anal. Chem.* **1994**, *66*, 2480–2487.
 - (51) Kimura, T.; Hasegawa, K.; Funatsu, K. GA Strategy for Variable Selection in QSAR Studies: GA-Based Region Selection for CoMFA Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 276–282.
 - (52) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure–Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279–1287.
 - (53) Kalivas, J. H.; Lang, P. M. *Mathematical Analysis of Spectral Orthogonality*; Marcel Dekker: New York, 1993.
 - (54) Malinowski, E. R. *Factor Analysis in Chemistry*, 2nd ed.; John Wiley & Sons: New York, 1991.

CI010036Q