

# QSAR study of selective ligands for the thyroid hormone receptor $\beta$

Huanxiang Liu and Paola Gramatica\*

*QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, via Dunant 3, 21100 Varese, Italy*

Received 1 December 2006; accepted 4 May 2007

Available online 10 May 2007

**Abstract**—In this paper, an accurate and reliable QSAR model of 87 selective ligands for the thyroid hormone receptor  $\beta$  1 (TR $\beta$ 1) was developed, based on theoretical molecular descriptors to predict the binding affinity of compounds with receptor. The structural characteristics of compounds were described wholly by a large amount of molecular structural descriptors calculated by DRAGON. Six most relevant structural descriptors to the studied activity were selected as the inputs of QSAR model by a robust optimization algorithm Genetic Algorithm. The built model was fully assessed by various validation methods, including internal and external validation, *Y*-randomization test, chemical applicability domain, and all the validations indicate that the QSAR model we proposed is robust and satisfactory. Thus, the built QSAR model can be used to fast and accurately predict the binding affinity of compounds (in the defined applicability domain) to TR $\beta$ 1. At the same time, the model proposed could also identify and provide some insight into what structural features are related to the biological activity of these compounds and provide some instruction for further designing the new selective ligands for TR $\beta$ 1 with high activity.

© 2007 Elsevier Ltd. All rights reserved.

## 1. Introduction

Nuclear receptors, comprised of a family of ligand-dependent transcription factors, can mediate the effects of hormones and other endogenous ligands to regulate the expression of specific genes. Among other members, this family includes receptors for the various steroid hormones (such as the estrogen, androgen, and progesterone receptor), thyroid hormones, retinoic acid, vitamin D, etc.<sup>1</sup> Unbalanced production or cell insensitivity to specific hormones may result in diseases associated with endocrine disfunction.<sup>2</sup> As the continuation of our research on steroid hormones, in particular estrogen receptor,<sup>3</sup> here, we focus on the thyroid hormones. Thyroid hormones exert a wide array of biologic activities and profound effects on growth, development, and homeostasis in mammals.<sup>4</sup> They regulate important genes in intestinal, skeletal, and cardiac muscles, liver, and the central nervous system, influence overall meta-

bolic rate, cholesterol and triglyceride levels, and heart rate, and affect mood and overall sense of well being. Some effects of thyroid hormones such as weight reduction for the treatment of obesity, cholesterol lowering to treat hyperlipidemia, amelioration of depression, and stimulation of bone formation in osteoporosis may be therapeutically useful in non-thyroid disorders.<sup>5</sup> Prior attempts to utilize thyroid hormones pharmacologically to treat these disorders have been limited by manifestations of hyperthyroidism and, in particular, by cardiovascular toxicity. If these adverse effects can be minimized or eliminated, thyroid hormones could potentially be very useful for treatment of various disorders.

The existence of two major subtypes of the thyroid hormone receptors,  $\alpha$  (TR $\alpha$ ) and  $\beta$  (TR $\beta$ ) expressed from two different genes, provides the possibility to design the therapeutically useful ligands without adverse effects. Differential RNA processing results in the formation of at least two isoforms from each gene. The TR $\alpha$ 1, TR $\beta$ 1, and TR $\beta$ 2 isoforms bind thyroid hormone and act as ligand-regulated transcription factors. The TR $\alpha$ 2 isoform is prevalent in the pituitary and other parts of the central nervous system, does not bind thyroid hormones, and acts in many contexts as

**Keywords:** QSAR; Thyroid hormone receptor  $\beta$ ; Selective ligands; Drug design; Theoretical molecular descriptors; Genetic Algorithm; Splitting methods; Model validation.

\* Corresponding author. Tel.: +39 0332 421573; fax: +39 0332 421554; e-mail: [paola.gramatica@uninsubria.it](mailto:paola.gramatica@uninsubria.it)

a transcriptional repressor. In adults, the TR $\beta$ 1 isoform is the most prevalent form in most tissues, especially in the liver. The TR $\alpha$ 1 isoform is also widely distributed although its levels are generally lower than those of the TR $\beta$ 1 isoform. Many or most effects of thyroid hormones on the heart, and in particular on the heart rate and rhythm, are mediated through the TR $\alpha$ 1 isoform.<sup>6</sup> The  $\beta$ -forms of the receptor mainly mediate in the liver and other tissues.<sup>7</sup> Consequently, development of thyroid hormone receptor agonists selective for the  $\beta$ 1 isoform could lead to specific therapies for these common disorders while avoiding cardiotoxicity.

To date there have been a limited number of reports on efforts to identify thyroid ligands that interact selectively with the various TR isoforms. Among the researches of selective thyroid ligands, Malm's group designed and synthesized new, highly active several series of ligands selective for TR $\beta$ .<sup>5,8–11</sup> In order to improve the efficiency and save money during the drug design process, quantitative structure–activity relationship (QSAR) studies on selective thyroid ligands were necessary, which could identify and provide some insight into what structural features are related to the biological activity of these compounds, predict the biological activity just from molecular structure before the compound is synthesized and tested by animal experiments, provide some instruction for further designing the new ligands, and narrow the search for future drug compounds. Based on the requirement and the advantages of QSAR method, Vedani et al. developed the satisfactory 4D–6D QSAR models to study the selective thyroid ligands by using their own software *Quasar and Raptor*.<sup>12,13</sup> Although their models provided good predictive results, it has a high level of complexity and limitation in the software availability and then is not easy to use for chemists and biologists.

For real life applications, here we developed an easy to use, fast-performing, and effective QSAR model with good predictive performance, which was certified by several validation paths, based on a relatively large data set. In order to interpret the structural characteristics of the compounds from all aspects, a large number of descriptors calculated by DRAGON were used to describe each compound's structure. At the same time, a robust optimization algorithm, Genetic Algorithm (GA), was used to select the related descriptors with the compound's activity and a simple correlation algorithm Multiple Linear Regression (MLR) was used to correlate the descriptors and  $-\log IC_{50}$  values of the compounds. In addition, the proposed models took into full account fundamental points required by the OECD principles for QSAR models,<sup>14</sup> namely their validation for predictivity (both by internal and external statistical validation) and the possibility of verifying their chemical applicability domain by the leverage approach. External validation was performed by splitting the original data set into training and prediction sets by two different methods: Kohonen Self Organizing Map (SOM) and random sampling by activity.

## 2. Results and discussion

### 2.1. The analysis of data set

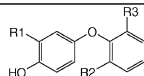
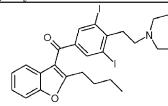
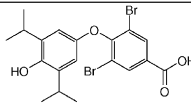
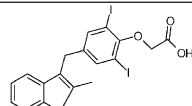
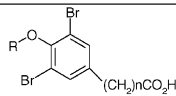
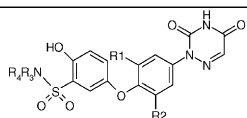


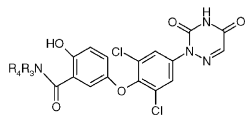


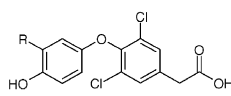

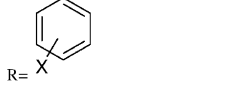





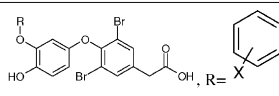

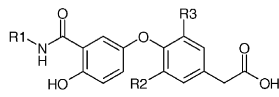
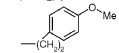
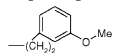


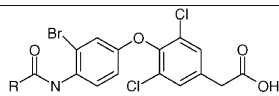
Since predictions from any QSAR models cannot be intrinsically better than the experimental data employed to develop the model, the quality of the input data will greatly influence the QSAR model performance. In order to build a QSAR model with good generalized performance, a preliminary analysis for the quality of the data set (mainly the detection of outliers) was performed by modeling the complete set of 87 chemicals.

A number of QSAR models for the whole data set were developed using Multiple Linear Regression by OLS based on the different sets of descriptors selected by Genetic Algorithms (GA). By comparing the Williams plot of every model to verify the presence of response outliers and structurally influential chemicals, most models showed two compounds numbered as 28 and 46 (Table 1) were response outliers. For the response outliers, it is difficult to find a reason why the models failed to predict them accurately; however, it must be kept in mind that the quality of the input experimental data for these chemicals could be questioned since the experimental values cannot be well predicted always by various descriptors. In order to build a reliable and general model, these chemicals were removed from the data set and the remaining 85 compounds were used in the following modeling.

### 2.2. The splitting of data set

Rational division of the experimental data set into training and prediction sets is a crucial part in the development and validation of reliable QSAR models.<sup>15,16</sup> Here, two criteria were applied to obtain appropriate external prediction sets for validation. In the first choice of training and prediction sets, the selection was made according to the distribution of the compounds in the space of molecular descriptors. The compounds were selected in this way to ensure that the training and prediction sets contained chemicals representative of the diversity of structures for which predictions were to be made. This complies with the hypothesis that in similarity-based QSAR approaches, the training set should be as structurally representative as possible and the prediction set as similar as possible to the training set to guarantee reliable prediction. Knowledge of the spatial distribution of the compounds in the descriptor space allowed the selection of a representative sample in both the training and prediction sets. The selection of a prediction set based on this principle was implemented by Kohonen Self Organizing Maps (SOM)<sup>17</sup> in the package *KOALA*.<sup>18</sup> SOM, that takes advantage of clustering capability, ensures that both sets are homogeneously distributed within the entire area of the descriptor space; in this case the chemicals in both sets, selected to maximize the coverage of the descriptor space (i.e., representativity), represent the depth of distribution of all existing chemicals.

Table 1. Structures of the studied compounds

 <p><b>1,2,4:</b> R<sub>1</sub>=R<sub>2</sub>=R<sub>3</sub>=I; <b>3,5-10:</b> R<sub>1</sub>=Isopropyl;  <b>3:</b> R<sub>2</sub>=R<sub>3</sub>=I; <b>5-7:</b> R<sub>2</sub>=R<sub>3</sub>=Br;  <b>8-10:</b> R<sub>2</sub>=R<sub>3</sub>=Cl.</p>	No.	R <sub>4</sub>	No.	R <sub>4</sub>
	1	-CH <sub>2</sub> CH(NH <sub>2</sub> )CO <sub>2</sub> H	2	-CH <sub>2</sub> CO <sub>2</sub> H
	3	-CH <sub>2</sub> CH(NH <sub>2</sub> )CO <sub>2</sub> H	4	-(CH <sub>2</sub> ) <sub>2</sub> CO <sub>2</sub> H
	5	-(CH <sub>2</sub> ) <sub>2</sub> CO <sub>2</sub> H	6	-CO <sub>2</sub> H
	7	-CH <sub>2</sub> CO <sub>2</sub> H	8	-CO <sub>2</sub> H
	9	-CH <sub>2</sub> CO <sub>2</sub> H	10	-(CH <sub>2</sub> ) <sub>2</sub> CO <sub>2</sub> H
 <p>11</p>	 <p>12</p>		 <p>13</p>	
	No.	R	No.	R
	14	Et	15	-CH <sub>2</sub> (CH <sub>2</sub> ) <sub>2</sub> CH <sub>3</sub>
	16	-CH <sub>2</sub> (CH <sub>2</sub> ) <sub>3</sub> CH <sub>3</sub>	17	-CH <sub>2</sub> CH <sub>2</sub> CH(CH <sub>3</sub> ) <sub>2</sub>
	18	-CH <sub>2</sub> (CH <sub>2</sub> ) <sub>4</sub> CH <sub>3</sub>	19	-CH <sub>2</sub> CH(CH <sub>2</sub> CH <sub>3</sub> ) <sub>2</sub>
	20	-CH <sub>2</sub> -Cyclohexyl	21	-CH <sub>2</sub> CH <sub>2</sub> CH(CH <sub>3</sub> ) <sub>2</sub>
 <p><b>14-20:</b> n=1;  <b>21-24:</b> n=2</p>	22	-CH <sub>2</sub> (CH <sub>2</sub> ) <sub>4</sub> CH <sub>3</sub>	23	-CH <sub>2</sub> CH(CH <sub>2</sub> CH <sub>3</sub> ) <sub>2</sub>
	24	-CH <sub>2</sub> -Cyclohexyl		
	No.	N-R <sub>3</sub> R <sub>4</sub>	No.	N-R <sub>3</sub> R <sub>4</sub>
	25	Piperidinyl	26	Piperidinyl
	27	Piperidinyl	28	4-Methylpiperidinyl
 <p><b>25:</b> R<sub>1</sub>=R<sub>2</sub>=Me;  <b>26, 28-29:</b> R<sub>1</sub>=Me, R<sub>2</sub>=Cl;  <b>27, 30-35:</b> R<sub>1</sub>=R<sub>2</sub>=Cl</p>	29	Morpholinyl	30	Cyclohexylamino
	31	Cyclobutylamino	32	Anilinyll
	33	Indolinyl		
	34		35	
	No.	N-R <sub>3</sub> R <sub>4</sub>	No.	N-R <sub>3</sub> R <sub>4</sub>
	36	Piperidinyl	37	Piperidinyl
	38	Cyclohexylamino	39	Cyclobutylamino
	40	Morpholinyl	41	(R)-(+)-Bornyl
	42		43	
	No.	X	No.	X
 <p>R = </p>	44	H <sub>2</sub>	45	-CF <sub>3</sub>
	47	4-CF <sub>3</sub>	48	3-Et
	50	3-Ph	51	3-OMe
	53	3-OCF <sub>3</sub>	54	2-OH
	56	4-OH	57	3-OH
 <p>R = </p>	58		59	
	60		61	
	62	n-C <sub>6</sub> H <sub>11</sub>		
	No.	X	No.	X
	63	H	64	2-CF <sub>3</sub>
 <p>R = </p>	65	3-CF <sub>3</sub>	66	4-CF <sub>3</sub>
	67	3-Ph	68	2-OH
	69	3-OH	70	4-OH
	No.	R <sub>1</sub>	No.	R <sub>1</sub>
 <p><b>71-80:</b> R<sub>2</sub>=R<sub>3</sub>=Br;  <b>81:</b> R<sub>2</sub>=R<sub>3</sub>=Cl.</p>	71	Ph	72	Bz
	73	(CH <sub>2</sub> ) <sub>2</sub> Ph	74	(CH <sub>2</sub> ) <sub>3</sub> Ph
	75	(CH <sub>2</sub> ) <sub>4</sub> Ph	76	CH <sub>2</sub> CHPh <sub>2</sub>
	77		78	
	79		80	
	81	(CH <sub>2</sub> ) <sub>2</sub> Ph		
	No.	R	No.	R
	82	CH <sub>2</sub> CH <sub>3</sub>	83	CH(CH <sub>3</sub> ) <sub>2</sub>
	84	CH(CH <sub>3</sub> )CH <sub>2</sub> CH <sub>3</sub>	85	CH(CH <sub>3</sub> )CH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub>
	86	CH(CH <sub>2</sub> CH <sub>3</sub> ) <sub>2</sub>	87	CH(CH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub> ) <sub>2</sub>

In *KOALA*, structural information was used as variables to build a Kohonen map (7 × 7 neurons, 400 epochs). At the end of 400 epochs of the net training, similar chemicals fell within the same neuron, that is, they carry the same information. To select the training set of chemicals it is assumed that the compound closest to each neuron centroid is the most representative of all the chemicals within the same neuron. Thus,

the selection of the training set chemicals was performed by the minimal distance from the centroid of each cell in the top map. The remaining objects, close to the training set chemicals, were used for the prediction set. As results, 21 compounds (about 25% of original data set (85 compounds)) fell into prediction set and 64 compounds were included in the training set.

In a second selection, and to ensure that the results were not conditioned by the data distribution in structural space, the splitting of the original data set was carried out by random selection through activity sampling. First, ordering the chemicals according to their descending experimental values, taking one chemical for every four chemicals from the original data set to put in prediction set (25% of the whole set, totally 21 compounds), the remaining chemicals were included in the training set for model development.

### 2.3. The construction and internal validation of QSAR models

As we cannot have a priori knowledge of which descriptors, and which particular combinations with others, are related to the studied response and are able to be used in models for prediction aims, we applied Genetic Algorithms as the variable selection procedure to select only the best combinations of those most relevant to obtaining models with the highest predictive power by using the training set obtained by SOM method.

The real usefulness of QSAR models is not just their ability to reproduce known data, verified by their fitting power ( $R^2$ ), but is mainly their possibility of predictive application. For this reason the model calculations were performed maximizing the explained variance in prediction, verified by the leave-one-out cross-validated correlation coefficient,  $Q_{\text{LOO}}^2$ . To avoid the danger of overfitting and the possibility of overestimating model predictivity by using only  $Q_{\text{LOO}}^2$ , the internal predictive ability of the models was also verified using the bootstrap ( $Q_{\text{BOOT}}^2$ ) procedure, as is strongly recommended for QSAR modeling.<sup>19</sup> The robustness of the proposed models and their predictive ability was guaranteed by the high  $Q_{\text{BOOT}}^2$  based on bootstrapping repeated 5000 times.

In addition, Y-randomization was applied to exclude the possibility of chance correlation, that is., fortuitous correlation without any predictive ability. It gave the following results: the random models, performed using a scrambled order of the experimental rate constant repeated 300 times, were found to have significantly lower  $R^2$  and  $Q^2$  than the original models, corroborating the statistical reliability of the actual models.

According to these validations, the best QSAR model based on the training set by SOM was selected among those with a smaller number of response outliers and structurally influential chemicals. The statistical analysis results of this model and the involved molecular descriptors as well as their full names in software DRAGON are summarized in Table 2. The detailed value of descriptors for every compound is given in Table 3.

From the parameters of the model in Table 2, it can be seen that the built model has satisfactory internal predictive ability and stability.

### 2.4. The external validation

For a QSAR model, internal validation, although important and necessary, does not sufficiently guarantee the predictive ability of a model. In fact, we, like other authors,<sup>20–22</sup> are strongly convinced from personal experience<sup>3,23–26</sup> that models with high apparent predictive ability, highlighted only by internal validation methods, can be unpredictable when verified on new chemicals not used in developing the model. Thus, for a stronger evaluation of model applicability for prediction on new chemicals, external validation of the models should always be performed.<sup>27</sup> In the present investigation, the built models were validated externally by the additional prediction set.

Twenty-one compounds in prediction data set (the compounds of which set value is 2 in Table 3) selected by SOM, not used during the development of model, were used to validate the built model externally. MDS (Multi-dimensional scaling) analysis was performed to check the distribution of compounds in prediction set in the training applicability domain. Figure 1 gives the MDS map of the first two dimensions. From this figure, the data set appears to be split into two representative sets thus confirming the efficiency of the applied SOM in the splitting: in fact the training set consists of representatives of the more dissimilar structures and the prediction set are well distributed in the training applicability domain. The predicted results for the training set and external prediction set can be seen in Figure 2. The correlation coefficient  $R_{\text{pred}}^2$ ,  $Q_{\text{EXT}}^2$ , and RMSE for the prediction set are 0.730, 0.711, and 0.702, respectively. The residual plot is shown in Figure 3. Residuals both

**Table 2.** The MLR model between the structural descriptor and the  $\text{pIC}_{50}$  of the compounds in the training set by SOM splitting

Variable	Meaning of variables	Regression coefficient	Error coefficient	Standardized coefficient
Intercept	Constant	9.996	1.424	0
GATS1e	Geary autocorrelation—lag 1 weighted by atomic Sanderson electronegativities	−7.683	1.621	−0.262
EEig08x	Eigenvalue 8 from edge adjacency matrix weighted edge degrees	−2.906	0.388	−1.399
EEig07d	Eigenvalue 7 from edge adjacency matrix weighted dipole moments	4.411	0.526	1.6649
GGI6	Topological charge index of order 6	1.994	0.633	0.3169
R6v+	R maximal autocorrelation of lag 6 index weighed by atomic van der waals volume	60.550	10.996	0.339
H-051	The number of the H atom attached to $\alpha$ C atom	−0.383	0.080	−0.287

Model parameters:  $n = 64$ ,  $R^2 = 0.836$ ,  $R_{\text{adj}}^2 = 0.819$ ,  $Q_{\text{LOO}}^2 = 0.793$ ,  $Q_{\text{BOOT}}^2 = 0.780$ , RMSE = 0.550,  $K_{\text{XX}} = 0.429$ ,  $K_{\text{YY}} = 0.466$ .

**Table 3.** The experimental and predicted  $-\log IC_{50}$  together with the selected descriptors

ID	GATS1e	EEig08x	EEig07d	GGI6	R6v+	H-051	Y Exp.	Y-Pred	Set <sup>a</sup>
1	0.722	2.372	1.846	0.771	0.042	0	9.59	9.84	1
2	0.71	1.869	1.846	0.567	0.046	2	10.32	10.43	1
3	0.709	2.372	2.145	0.771	0.031	0	9.96	10.62	1
4	0.705	1.984	1.846	0.69	0.049	2	10.72	10.47	1
5	0.732	2.334	2.145	0.69	0.033	2	10.6	9.56	1
6	0.747	2.334	1.873	0.446	0.034	0	8.68	8.69	1
7	0.739	2.334	2.144	0.567	0.033	2	10.02	9.28	1
8	0.743	2.334	1.873	0.446	0.02	0	7.68	7.91	1
9	0.734	2.334	2.145	0.567	0.019	2	8.96	8.48	1
10	0.727	2.334	2.145	0.69	0.018	2	9.82	8.64	1
11	0.833	2.874	1.952	0.884	0.02	2	6.22	6.03	1
12	0.726	2.72	1.904	0.609	0.012	0	6.04	7.03	1
13	0.801	2.83	1.961	0.488	0.023	3	5.29	5.51	1
14	0.864	0.941	0.745	0.202	0.022	2	4.49	4.88	2
15	0.828	1	0.775	0.304	0.022	2	5.72	5.24	1
16	0.814	1.095	0.989	0.304	0.02	2	5.24	6.11	1
17	0.814	1	0.986	0.467	0.022	2	5.74	6.88	1
18	0.802	1.345	1.155	0.304	0.02	2	5.82	6.11	1
19	0.802	1.393	0.8	0.406	0.029	2	6.1	4.86	1
20	0.754	2	1.415	0.325	0.025	2	6.11	6.02	1
21	0.802	1.421	1.087	0.59	0.023	2	5.74	6.4	1
22	0.792	1.666	1.171	0.427	0.02	2	6.44	5.53	2
23	0.792	1.618	1.071	0.529	0.027	2	6.09	5.82	1
24	0.747	2	1.418	0.448	0.026	2	6.72	6.38	1
25	0.711	3.218	2.454	0.909	0.013	0	8.89	8.61	2
26	0.735	3.218	2.527	0.909	0.014	0	9.44	8.75	1
27	0.763	3.218	2.623	0.909	0.014	0	10.25	9.01	2
29	0.837	3.218	2.555	0.909	0.014	0	7.12	8.14	2
30	0.734	3.232	2.623	1.013	0.013	0	8.89	9.39	1
31	0.753	3.232	2.623	0.912	0.017	0	8.95	9.24	2
32	0.734	3.232	2.623	1.013	0.015	0	9.55	9.46	2
33	0.719	3.232	2.623	1.103	0.013	0	9.19	9.64	2
34	0.85	3.387	2.773	1.278	0.011	0	9.06	9.07	1
35	0.77	3.387	2.77	1.278	0.011	0	9.11	9.78	1
36	0.803	3.218	2.502	0.828	0.018	0	8.28	8.25	1
37	0.822	3.218	2.623	0.828	0.017	0	8.76	8.56	1
38	0.792	3.232	2.623	0.891	0.015	0	9.71	8.77	2
39	0.81	3.232	2.623	0.83	0.017	0	8.75	8.64	2
40	0.92	3.218	2.623	0.828	0.019	0	8.7	7.65	1
41	0.747	3.273	2.623	1.207	0.015	0	8.99	9.73	1
42	0.752	3.232	2.623	1.133	0.012	0	9.78	9.33	1
43	0.752	3.232	2.623	1.133	0.011	0	9.95	9.24	1
44	0.688	2.962	2.266	0.625	0.025	2	8.54	8.09	2
45	0.693	3.232	2.423	0.851	0.022	2	8.47	8.21	1
46	0.693	3.232	2.402	0.933	0.024	2	9.03	8.35	1
47	0.693	3.232	2.397	0.869	0.025	2	7.17	8.33	2
49	0.675	3.226	2.375	0.811	0.018	2	8.4	7.78	1
50	0.646	3.232	2.376	0.829	0.02	2	7.68	8.32	1
51	0.788	3.201	2.372	0.729	0.024	2	7.96	7.21	1
52	0.763	3.227	2.569	0.85	0.02	2	9.11	8.2	1
53	0.768	3.232	2.601	0.971	0.023	2	8.47	8.75	1
54	0.699	2.966	2.359	0.665	0.024	2	7.3	8.43	2
55	0.699	3.179	2.367	0.688	0.026	2	7.12	8.01	2
56	0.699	3.213	2.371	0.746	0.026	2	7.06	8.04	2
57	0.736	2.962	2.266	0.625	0.023	2	6.86	7.64	1
58	0.736	2.962	2.266	0.625	0.023	2	7.39	7.61	1
59	0.736	2.962	2.266	0.625	0.026	2	6.91	7.82	1
60	0.788	2.962	2.266	0.625	0.027	2	7.17	7.44	2
61	0.767	2.849	2.277	0.565	0.026	2	7.55	7.8	2
62	0.713	2.697	1.944	0.668	0.021	2	8.11	7.02	1
63	0.807	2.965	2.356	0.667	0.032	2	8.22	8.06	1
64	0.761	3.232	2.573	0.975	0.022	2	8.12	8.64	1
65	0.761	3.232	2.572	0.871	0.031	2	8.96	8.95	1
66	0.761	3.232	2.567	0.788	0.027	2	8.02	8.54	1
67	0.756	3.232	2.498	0.808	0.022	2	7.22	8.01	1

(continued on next page)

Table 3 (continued)

ID	GATS1e	EEig08x	EEig07d	GGI6	R6v+	H-051	Y Exp.	Y-Pred	Set <sup>a</sup>
68	0.806	2.97	2.363	0.73	0.029	2	7.08	8.09	1
69	0.806	3.232	2.378	0.749	0.032	2	7.33	7.59	1
70	0.806	3.232	2.418	0.747	0.025	2	8.22	7.26	1
71	0.735	3.219	2.495	0.731	0.029	2	7.46	8.5	1
72	0.73	3.223	2.545	0.75	0.025	2	8.38	8.5	1
73	0.724	3.224	2.549	0.769	0.028	2	9.21	8.75	1
74	0.719	3.224	2.55	0.728	0.025	2	8.17	8.55	2
75	0.715	3.225	2.55	0.687	0.025	2	7.74	8.5	2
76	0.814	3.232	2.571	0.81	0.025	2	7.43	8.09	1
77	0.814	3.232	2.551	0.769	0.027	2	8.32	7.98	1
78	0.814	3.224	2.55	0.81	0.026	2	8.07	8.04	1
79	0.683	3.232	2.549	1.03	0.019	2	9.33	9.04	2
80	0.769	3.224	2.549	0.769	0.029	4	7.17	7.8	1
81	0.72	3.224	2.549	0.769	0.02	2	8.46	8.31	1
82	0.802	2.82	2.369	0.708	0.021	4	6.85	7.31	1
83	0.791	2.834	2.37	0.728	0.02	3	7.74	7.64	1
84	0.782	2.842	2.37	0.769	0.023	3	8.37	7.93	1
85	0.773	2.848	2.37	0.769	0.014	3	7.33	7.48	1
86	0.773	2.854	2.37	0.81	0.017	3	7.8	7.7	1
87	0.758	2.874	2.37	0.81	0.012	3	7.35	7.47	2

<sup>a</sup> The training set and prediction set obtained by SOM method, '1' accounts for the compound in the training set, '2' accounts for the compound in the prediction set.

for training and prediction set are distributed normally around zero (the mean value), therefore the linear correlation between response and selected variables is reliable. The plot of predicted versus experimental activity (Fig. 2) tells the same story, adding the information that visually the predicted values seem to capture the actual values very well.

The chemical applicability domain of the models and the reliability of the predictions are also verified by the leverage approach. In Williams plot,<sup>28</sup> chemicals influential on the structural domain of the model, characterized by a hat value exceeding the cut off one (vertical dashed line in Fig. 4), can be explained as compounds with peculiar features poorly represented in the training set, which could affect the variables, selection for a better modeling of those chemicals. Outliers, of which the standardized residual values exceed the cut off value (here,  $\pm 2.5 \sigma$ , horizontal dashed line in Fig. 4), could be associated with errors in the experimental values. On analyzing the AD of built model in the Williams plot (Fig. 4), there is no response outlier and structure influential compound both for training set and prediction set, which indicated further the reliability of the predictions from another aspect.

From the above results, both for the training set and external prediction set, it can be seen the model we suggest matches the high quality parameters not only with good fitting power, but mainly with high capability of assessing external data.

## 2.5. The further validation by the training and prediction set obtained from random splitting

In order to ensure that the results were not conditioned by the data distribution in descriptor space, the built model was also validated by the training and prediction set selected from random method and activity sampling,

which has been described in the Section 2.2. Figure 5 is the MDS map of the first two dimensions for the new training and prediction set. Compared with Figure 1, where the compounds in prediction set were included in the applicability domain of training set, it was certified further that the training set obtained by SOM was more representative than that by randomness. In fact, in Figure 5, several compounds in the prediction set are out of the structural domain of training set. However, we can also obtain the very satisfactory results when we apply the selected descriptors to this new training and prediction set. The detailed comparison of results between the models obtained by two different splitting methods is given in Table 4. The satisfactory results proved the built model was not conditioned by the data distribution in structural space as well as the robustness and reliability of the built model further. In addition,

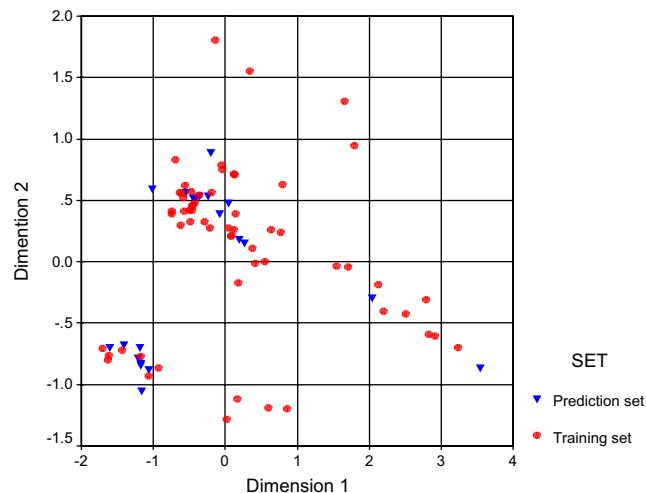
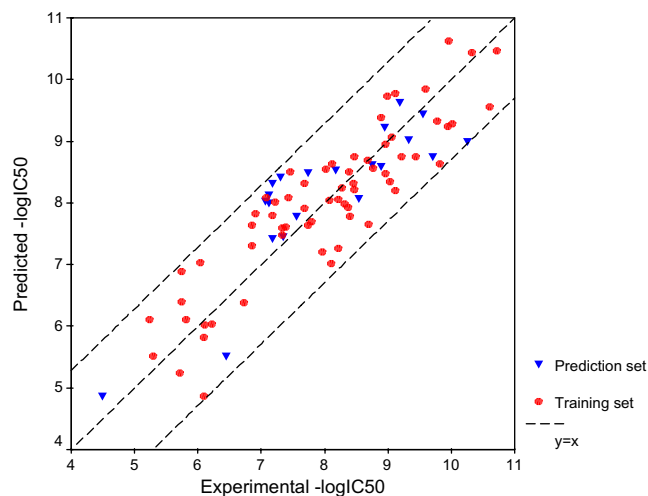
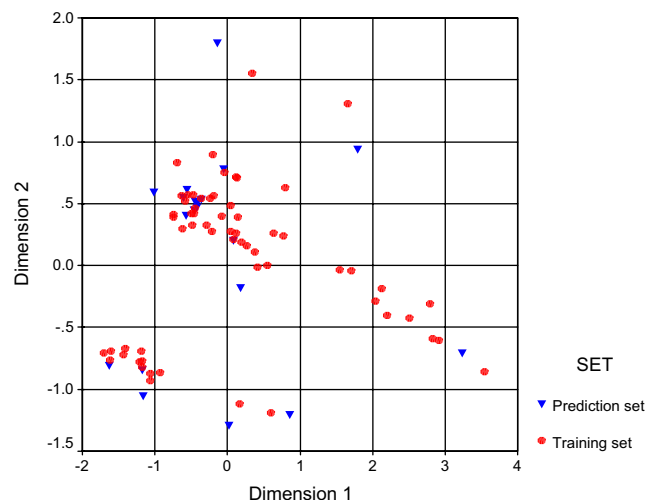


Figure 1. MDS map for the training set and prediction set split by SOM neural network.

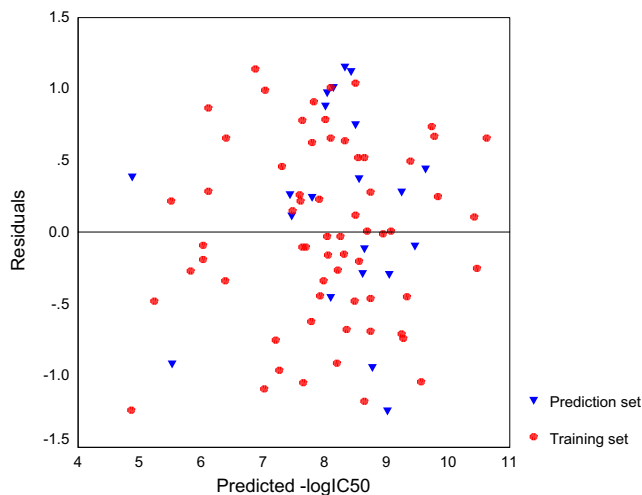




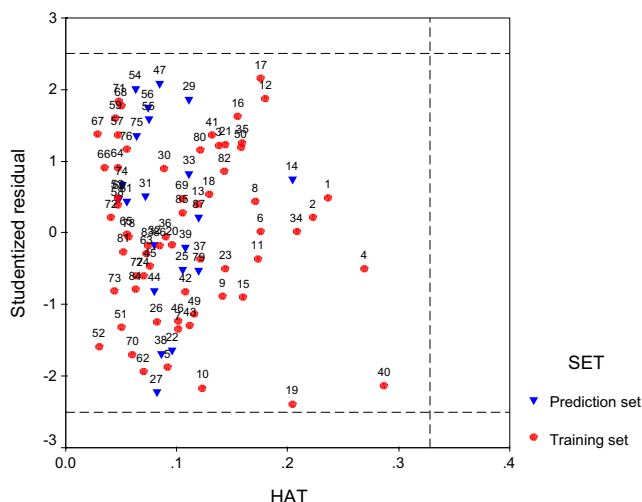
**Figure 2.** Predicted  $-\log IC_{50}$  values versus experimental values for training set and prediction set (two side lines express the confidence interval of 95%).



**Figure 5.** MDS map for the training set and prediction set split by randomness.



**Figure 3.** Residuals versus predicted  $-\log IC_{50}$  values for training set and prediction set.



**Figure 4.** Williams plot of standardized residuals versus hat values.

although it is impossible to have an absolute measure of comparison between our model and Vedani's model because we used a larger data set and different end-point ( $IC_{50}$  in our study,  $K_i$  in their study), a rough comparison indicated our results can be comparable to their model as well as our model is much simpler and easier to use.

## 2.6. Structural features responsible for activity

By interpreting the molecular descriptors in the regression model, it is possible to gain some insight into structural features that are likely to govern the affinity of chemicals with the thyroid hormone receptor  $\beta$ .

In the built models, the relative importance of the variables was determined by their standardized regression coefficients. In fact, since molecular descriptors do not have equal variance (i.e., they are not autoscaled), their relative importance in the model is measured better by standardized regression coefficients (i.e., the coefficients multiplied by the standard deviation of the corresponding predictor). From the standardized coefficient of descriptors in Table 2, the most important descriptor is an edge adjacency index EEig07d (Eigenvalue 7 from edge adjacency matrix weighted dipole moments), which was developed by Estrada et al.,<sup>29,30</sup> is mainly correlated with molecular polarity. The second important descriptor EEig08x (Eigenvalue 8 from edge adjacency matrix weighted edge degrees) also belongs to the edge adjacency index. This descriptor can reflect molecular branching and complexity, and then accounts for molecular steric feature. GGI6 (topological charge index of order 6)<sup>31</sup> can reflect the charge distribution of a molecule and could be related to the electrostatic interaction between ligand and receptor. Another descriptor related with molecular electrostatic features is the descriptor GAT1e, which responds to Geary autocorrelation—lag 1 weighted by atomic Sanderson electronegativities.

**Table 4.** The statistical parameters of the models based on two kinds of splitting methods

Splitting method	Training set (64 chemicals)			Prediction set (21 chemicals)		
	$R^2$	$Q^2$	RMSE	$R^2_{\text{pred}}$	$Q^2_{\text{EXT}}$	RMSE
SOM	0.836	0.793	0.550	0.730	0.711	0.702
Random by activity	0.813	0.766	0.600	0.808	0.790	0.554

R6v+ ( $R$  maximal autocorrelation of lag 6 index weighed by atomic van der Waals volumes) belongs to the GETAWAY descriptors developed by Consonni et al.,<sup>32</sup> and is defined based on the influence/distance matrix  $R$  weighting the atoms by van der Waals volumes. Thus, it can reflect molecular size in addition to the distribution of influential atoms (for instance, heteroatoms).

The meaning of the descriptor H-051 is the number of H atom attached to  $\alpha$  C atom, which is a very active atom and easy to lose acidic proton and then affects the reactivity of molecule and their nucleophilicity.

From the above discussion, we can conclude that the activity of the studied compounds mainly depends on molecular polarity, size, shape, and nucleophilic reactivity.

However, it must be taken into account that, in multivariate models such as Multi-Linear Regression models, even though the interpretation of the singular molecular descriptor can be useful, only the combination of the selected set of descriptors is able to model the studied end-point, and this combination is not always easy to be interpreted in a very reduced way for such a complex biological effect based on a lot of interactions between ligands and receptor.

### 3. Conclusion

In this article, QSAR study of 85 selective ligands for the thyroid hormone receptor  $\beta$  was performed based on theoretical molecular descriptors calculated by DRAGON software and selected by Genetic Algorithm. The built model was assessed comprehensively (internal and external validation) and all the validations indicate that the QSAR model we built is robust and satisfactory, and that the selected descriptors can account for the structural features responsible for the binding affinity of compounds to TR $\beta$ 1 as well as GA is an effective method to select descriptors. By interpreting the molecular descriptors in the regression model, we can conclude that the activity of the studied compounds mainly depends on molecular polarity, size, shape, and nucleophilicity. The QSAR model developed in this study can provide a useful tool to predict the activity of the new compounds and also to design new compounds with high activity.

## 4. Methodology

### 4.1. Experimental data

The affinity data of 87 ligands to  $\beta$ <sub>1</sub> isoform of the human thyroid hormone receptor (TR $\beta$ <sub>1</sub>) were taken from

five references in terms of  $\text{IC}_{50}$ <sup>5,9–11</sup> or  $K_i$ <sup>8</sup> values and were converted to  $\text{pIC}_{50}$  (that is  $-\log \text{IC}_{50}$ ). The structures of compounds used in the study are given in Table 1 and the experimental affinity activities ( $\text{pIC}_{50}$ ) are shown in Table 3.

### 4.2. Calculation of molecular descriptors

To obtain a QSAR model, compounds are represented by theoretical molecular descriptors. The calculation process of the molecular descriptors was described as below. All structures were drawn and preoptimized using MM+ molecular mechanics method within the framework of the HYPERCHEM program.<sup>33</sup> The final geometries of the minimum energy conformation were obtained by more precise optimization with the AM1 method. The molecular descriptors for the given compounds were calculated using the software DRAGON 5.4<sup>34</sup> on the minimal energy conformations.

By DRAGON, a total of 1354 different type molecular descriptors were calculated to describe the compounds' structural diversity. They include (a) 0D-constitutional (atom and group counts); (b) 1D-functional groups, 1D-atom centered fragments; (c) 2D-topological, 2D-BCUTs, 2D-walk and path counts, 2D-autocorrelations, 2D-connectivity indices, 2D-information indices, 2D-topological charge indices, and 2D-eigenvalue-based indices; and (d) 3D-Randic molecular profiles from the geometry matrix, 3D-geometrical, 3D-WHIM,<sup>35</sup> and 3D-GETAWAY<sup>32</sup> descriptors. In order to reduce redundant and non-useful information, constant or near constant values and descriptors found to be highly correlated pairwise (one of any two descriptors with a  $K$  correlation greater than 0.99) were excluded in a pre-reduction step, thus 680 molecular descriptors underwent subsequent variable selection.

### 4.3. Variable selection based on Genetic Algorithm

After calculating the molecular descriptors, the next step is to reduce descriptor space by selecting only pertinent descriptors that faithfully describe the activity of interest. Choosing appropriate descriptors for QSAR studies is difficult as there are no absolute rules that govern this choice. However, it is well known, in both chemical and statistical fields, that the accuracy of classification and regression techniques is not monotonic with respect to the number of features employed by the model. Thus, depending on the nature of the regression technique, the presence of irrelevant or redundant features can cause the system to focus attention on the idiosyncrasies of the individual samples and lose sight of the broad picture that is essential for generalization beyond the training set. For this reason descriptor selection is very important. Recently, some published papers suggested



that Genetic Algorithms (GA) might be useful in data analysis, especially in the task of reducing the number of features for regression models.<sup>36–39</sup> GA is a novel optimization technique that mimics selection in nature. The essence of ‘selection in nature’ is that, under certain environmental conditions, species of high fitness can prevail in the next generation, and the best species may be reproduced by crossover together with random mutations of chromosomes in surviving species. Rogers and Hopfinger<sup>40</sup> first applied this method in QSAR analysis, and proved GA a very effective tool with many merits, compared to other methods. Much of the research carried out by our group also verifies the success of GA in the selection of descriptors.<sup>3,23–26</sup> Thus, in this paper, Genetic Algorithm, as a powerful optimization method, was used for variable selection. To make the chemists’ job easier, the models were built using the simple Multiple Linear Regression method. Genetic Algorithm and Multiple Linear Regression analysis were performed by the software MOBY DIGS<sup>19</sup> using the Ordinary Least Square regression (OLS) method and GA-VSS (Genetic Algorithm-Variable Subset Selection).

First of all, models with 1–2 variables were developed by the all-subset-method procedure to explore all the low dimension combinations. The number of descriptors was subsequently increased one by one, and new models were formed. The outcome of the Genetic Algorithms in MOBY DIGS software is a population of 100 regression models, ordered according to their decreasing internal predictive performance, verified by the leave-one-out cross-validated correlation coefficient  $Q^2$ . The GA was stopped when increasing the model size did not increase the  $Q^2$  value to any significant degree. Particular attention was paid to the collinearity of the selected molecular descriptors: in fact, to avoid multicollinearity without, or with, ‘apparent’ prediction power (due to chance correlation), regression was calculated only for variable subsets with an acceptable multivariate correlation with response, by applying the *QUICK* rule (*Q* Under Influence of *K*).<sup>41</sup> According to this rule, acceptable models are only those with a global correlation of  $[X + Y]$  block ( $K_{XY}$ ) greater than the global correlation of the *X* block ( $K_{XX}$ ) variable, *X* being the molecular descriptors and *Y* the response variable. The collinearity in the original set of molecular descriptors results in many similar models of different dimensionality that more or less yield the same predictive power. Therefore, when there were models of similar performance, those with higher  $\Delta K$  ( $K_{XY} - K_{XX}$ ) were selected and further verified.

#### 4.4. Internal and external validation of models

The robustness of the models and their internal predictive ability were evaluated by both  $Q^2$  based on leave-one-out cross-validation and bootstrap. The leave-one-out (LOO) procedure consists of removing one sample from the training set, constructing the model only on the basis of the remaining training data and then testing it on the removed sample. In this fashion all the training data samples were tested and  $Q^2$  was calculated. In the Bootstrap procedure, *K* *n*-dimensional groups are generated

by a randomly repeated selection of *n*-objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then  $Q^2$  is calculated for each model. The bootstrapping was repeated 5000 times for each validated model.

The proposed models were also checked for reliability and robustness by permutation testing: new models were recalculated for randomly reordered response (*Y* scrambling). The resulting models obtained on the data set with randomized response should have significantly lower  $Q^2$  values than the proposed ones because the relationship between the structure and response is broken. This is proof of the proposed model’s validity as it can be reasonably excluded that the originally proposed model was obtained by chance correlation.<sup>42</sup> *Y* scrambling was performed by response scrambling with maximum iterations of 300.

It is worthwhile to point out that cross-validation and bootstrap methods only assess the internal predictive ability of built models.<sup>42,27,43</sup> Compared with cross-validation and bootstrap, the external validation can provide a more rigorous evaluation of the model’s predictive capability for untested chemicals. When a sufficiently large number of new (i.e., obtained after the model development) and reliable experimental data are available, the best proof of already developed model accuracy is to test model performance on these additional data. However, in the absence of available additional data (in useful quantity and quality), statistical external validation can be done by adequately splitting the available input data set, before model development, into training set (for model development) and prediction set (for model predictive assessment) by different procedures. In this investigation, the built models were validated externally by splitting the original data set into training set and prediction set. The prediction set does not take part in the selection of descriptor and the construction of model. Thus, the chemicals in the prediction set are completely new for the developed QSAR model.

In the splitting of the original data set into training and prediction set, the composition of the two sets is of crucial importance. The best splitting must guarantee that the training and prediction sets are scattered over the whole area occupied by representative points in the descriptor space (representativity) and that the training set is distributed over the area occupied by representative points for the whole data set (diversity). In order to select such training and prediction set, Kohonen Self Organizing Map (SOM)<sup>17</sup> was applied.

The splitting of the data set based on SOM takes advantage of the clustering capabilities of Kohonen artificial neural network allowing the selection of a meaningful training set and a representative prediction set, which was implemented by using the package *KOALA*.<sup>18</sup>

The above splitting methodology is based on similarity analysis, therefore the external prediction set of chemicals is, by definition, as structurally similar as possible

to the training set chemicals or is included in the training structural space: this allows the same chemical domain to be maintained. However, in this situation, there is a reasonable doubt that the developed models could, obviously, be predictive for chemicals which, even if not included in the training set, are in some way structurally very similar to these compounds. To eliminate this doubt from our model, and to verify if it is applicable to 'completely external' chemicals that do not participate not even in the similarity-based splitting, we also split the experimental data into training set and prediction set randomly according to the ranking of activity.

In order to check the distribution of compounds of the training and prediction sets in the descriptor space, multidimensional scaling (MDS) analysis<sup>44</sup> was performed. Multidimensional scaling (MDS) is a set of data analysis techniques that display the structure of distance-like data as a geometrical picture. MDS pictures the structure of the objects from data that approximate the distances between pairs of the objects. Each object is represented by a point in a multidimensional space. The points are arranged in this space so that the distances between pairs of points have the strongest possible relation to the similarities among the pairs of objects. That is, two similar objects are represented by two points that are close together, and two dissimilar objects are represented by two points that are far apart. The space is usually a two- or three-dimensional Euclidean space, but may be non-Euclidean and may have more dimensions. Here, two-dimensional Euclidean space was used.

A final validation procedure was performed by evaluating the model's external predictive power on the selected prediction set by every method. The reliability of the built models was assessed by the coefficient of determination  $R^2_{\text{pred}}$  for prediction set and the external validation parameter ( $Q^2_{\text{EXT}}$ ), which was calculated by  $Q^2_{\text{EXT}} = 1 - \text{PRESS}/\text{SD}$ , where PRESS is the sum of the squared differences between the measured response and the predicted values for each molecule in the prediction set, and SD is the sum of the squared deviations between the measured response for each molecule in the prediction set and the mean measured value of the training set. Root mean square of errors (RMSE), calculated separately for the training and the prediction sets, are reported to indicate the predicted accuracy of models, which is calculated by the root square of the sum of squared errors in prediction divided by their total number.

#### 4.5. Applicability domain of models

As even a robust, significant, and validated QSAR model cannot be expected to reliably predict the studied property for the entire universe of chemicals, its domain of application must be defined, and the predictions for only those chemicals that fall in this domain can be considered reliable. The applicability domain (AD) is a theoretical region in the space defined by the descriptors of the model and the modeled response, for which a given QSAR should make reliable predictions. This region is

defined by the nature of the chemicals in the training set, and can be characterized in various ways.

In this investigation, the chemical domain of the studied chemicals in the models was verified by the leverage approach to verify prediction reliability.<sup>28</sup> To visualize the AD of a QSAR model, the plot of standardized cross-validated residuals versus leverage (Hat diagonal) values (HAT) (the Williams plot) can be used to have an immediate and simple graphical detection of both the response outliers ( $Y$  outliers) and the structurally influential chemicals ( $X$  outliers) in a model. In this plot the horizontal and vertical straight lines indicate the limits of normal values: the first for the  $Y$  outliers (i.e., compounds with cross-validated standardized residuals greater than 2.5 standard deviation units,  $\pm 2.5\sigma$ ) and the second for  $X$  outliers. The limit of normal values for  $X$  outliers ( $h^*$ ) was calculated by  $3p'/n$ , where  $p'$  is the number of model variables plus one, and  $n$  is the number of the objects used to calculate the model.<sup>28</sup>

In fact, leverage can be used as a quantitative measure of the model applicability domain suitable for evaluating the degree of extrapolation: it represents a sort of compound 'distance' from the model experimental space. Prediction must be considered unreliable for compounds with a high leverage value ( $h > h^*$ ). Conversely, when the leverage value of a compound is lower than the critical value, the probability of accordance between predicted and actual values is as high as that for the training set chemicals.

In addition, it is important to note that the outliers for the response can be highlighted only for chemicals with known responses and the possibility of a chemical to be out of the structural applicability domain of a model can be verified for every new chemical, the only knowledge needed being the molecular structure information represented by the molecular descriptors selected in the model.

#### Acknowledgment

We thank the University of Insubria for providing a post doc fellowship to Dr. Huanxiang Liu.

#### References and notes

1. Lazar, M. A. *J. Invest. Med.* **1990**, *47*, 364.
2. Zubay, G. L.; Parson, W. W.; Vance, D. E. *Principles of Biochemistry*; Wm. C. Brown Communications, Inc.: Dubuque, USA, 1995.
3. Liu, H.; Papa, E.; Gramatica, P. *Chem. Res. Toxicol.* **2006**, *19*, 1540–1548.
4. Lazar, M. A. *Endocr. Rev.* **1993**, *14*, 184–193.
5. Ye, L.; Li, Y.-L.; Mellström, K.; Mellin, C.; Bladh, L.-G.; Koehler, K.; Garg, N.; Collazo, A. M. G.; Litten, C.; Husman, B.; Persson, K.; Ljunggren, J.; Grover, G.; Sleph, P. G.; George, R.; Malm, J. *J. Med. Chem.* **2003**, *46*, 1580–1588.
6. Forrest, D.; Vennström, B. *Thyroid* **2000**, *10*, 41–52.
7. Takeda, K.; Sakurai, A.; DeGroot, L. J.; Refetoff, S. *J. Clin. Endocrin., Metab.* **1992**, *74*, 49–55.

8. Dow, R. L.; Schneider, S. R.; Paight, E. S.; Hank, R. F.; Chiang, P.; Cornelius, P.; Lee, E.; Newsome, W. P.; Swick, A. G.; Spitzer, J.; Hargrove, D. M.; Patterson, T. A.; Pandit, J.; Chrunyk, B. A.; LeMotte, P. K.; Danley, D. E.; Rosner, M. H.; Ammirati, M. J.; Simons, S. P.; Schulte, G. K.; Tate, B. F.; DaSilva-Jardine, P.. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 379.
9. Hangeland, J. J.; Doweiko, A. M.; Dejneka, T.; Friends, T. J.; Devasthale, P.; Mellström, K.; Sandberg, J.; Grynfar, M.; Sack, J. S.; Einspahr, H.; Färnegårdh, M.; Husman, B.; Ljunggren, J.; Koehler, K.; Sheppard, C.; Malm, J.; Ryono, D. E. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 3549.
10. Li, Y.-L.; Litten, C.; Koehler, K. F.; Mellström, K.; Garg, N.; Collazo, A. M. G.; Färnegård, M.; Grynfarb, M.; Husman, B.; Sandberg, J.; Malm, J. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 884–886.
11. Collazo, A. M. G.; Koehler, K. F.; Garg, N.; Färnegård, M.; Husman, B.; Ye, L.; Ljunggren, J.; Mellström, K.; Sandberg, J.; Grynfarb, M.; Ahola, H.; Malm, J. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1240–1244.
12. Vedani, A.; Dobler, M.; Lill, M. A. *Basic Clin. Pharmacol. Toxicol.* **2006**, *99*, 187.
13. Vedani, A.; Zumstein, M.; Lill, M. A.; Ernst, B. *ChemMedChem* **2007**, *2*, 78–87.
14. [http://www.oecd.org/document/23/0,2340,en\\_2649\\_201185\\_33957015\\_1\\_1\\_1\\_1,00.html](http://www.oecd.org/document/23/0,2340,en_2649_201185_33957015_1_1_1_1,00.html).
15. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. *J. Comput. Aid. Mol. Des.* **2003**, *17*, 241–253.
16. Leonard, J. T.; Roy, K. *QSAR Comb. Sci.* **2006**, *25*, 235–251.
17. Zupan, J.; Novic, M.; Ruisánchez, I. *Chemom. Int. Lab. Syst.* **1997**, *38*, 1–23.
18. Todeschini, R.; Consonni, V. *KOALA-Software for Kohonen Artificial Neural Networks*. Rel. 1.0 for Windows, Talet srl, Milan, Italy, **2001**.
19. Todeschini, R.; Consonni, V.; Pavan, M. *MOBY DIGS—software for multilinear regression analysis and variable subset selection by genetic algorithm*. Version 1.2 for Windows, Talet srl, Milan, Italy, **2002**.
20. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269–276.
21. Cash, G. G.; Anderson, B.; Mayo, K.; Bogaczyk, S.; Tunkel, J. *Mutat. Res.* **2005**, *585*, 170–183.
22. Oberg, T. *Chem. Res. Toxicol.* **2004**, *17*, 1630–1637.
23. Papa, E.; Villa, F.; Gramatica, P. *J. Chem. Inf. Model.* **2005**, *45*, 1256–1266.
24. Gramatica, P.; Pilutti, P.; Papa, E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1794–1802.
25. Gramatica, P.; Giani E.; Papa, E. *J. Mol. Graph. Model.*, in press.
26. Gramatica, P.; Papa, E. *QSAR Comb. Sci.* **2005**, *24*, 953–960.
27. Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.* **2003**, *22*, 69–76.
28. Atkinson, A. C. *Plots, Transformations and Regression*; Clarendon Press: Oxford (UK), 1985.
29. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31–33.
30. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 701–707.
31. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
32. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693–705.
33. *HyperChem*. Release 7.03 for Windows, Molecular Modeling System. Hypercube, Inc., Gainesville, FL, USA **2002**.
34. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *DRAGON—software for the calculation of molecular descriptors*. Version 5.4 for Windows, Talet srl, Milan, Italy, **2006**.
35. Todeschini, R.; Gramatica, P. *SAR QSAR Environ. Res.* **1997**, *7*, 89–115.
36. Yasri, A.; Hartsough, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218–1227.
37. Hou, T. J.; Wang, J. M.; Liao, N.; Xu, X. J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 775–781.
38. Hasegawa, K. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 112–120.
39. Leardi, R.; Boggia, R.; Terrile, M. *J. of Chemom.* **1992**, *6*, 267–281.
40. Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
41. Todeschini, R.; Maiocchi, A.; Consonni, V. *Chemom. Int. Lab. Syst.* **1999**, *46*, 13–29.
42. Eriksson, L.; Jaworska, J.; Worth, A.; Cronin, M.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111*, 1361–1375.
43. Baumann, K. *Trends Anal. Chem.* **2003**, *22*, 395–406.
44. Borg, I.; Groenen, P. J. F. *Modern Multidimensional Scaling: Theory and Applications*; Springer: New York, 1997.