

# QSAR models for predicting the activity of non-peptide luteinizing hormone-releasing hormone (LHRH) antagonists derived from erythromycin A using quantum chemical properties

Michael Fernández · Julio Caballero

Received: 24 May 2006 / Accepted: 17 October 2006 / Published online: 10 January 2007  
© Springer-Verlag 2007

**Abstract** Multiple linear regression (MLR) combined with genetic algorithm (GA) and Bayesian-regularized Genetic Neural Networks (BRGNNs) were used to model the binding affinity ( $pK_i$ ) of 38 11,12-cyclic carbamate derivatives of 6-*O*-methylerythromycin A for the Human Luteinizing Hormone-Releasing Hormone (LHRH) receptor using quantum chemical descriptors. A multiparametric MLR equation with good statistical quality was obtained that describes the features relevant for antagonistic activity when the substituent at the position 3 of the erythronolide core was varied. In addition, four-descriptor linear and nonlinear models were established for the whole dataset. Such models showed high statistical quality. However, the BRGNN model was better than the linear model according to the external validation process. In general, our linear and nonlinear models reveal that the binding affinity of the compounds studied for the LHRH receptor is modulated by electron-related terms.

**Keywords** QSAR analysis · Bayesian-regularized Genetic Neural Network · Quantum chemical descriptors · Macrolide LHRH antagonists

## Introduction

Luteinizing Hormone-Releasing Hormone (LHRH), which is secreted from the hypothalamus, acts on the pituitary gland to stimulate the secretion of both luteinizing hormone and follicle-stimulating hormone [1]. These gonadotropins, in turn, act on the reproductive organs, where they participate in the regulation of gonadal steroid production, spermatogenesis in male and follicular development in female. Antagonists of LHRH bind to its receptor in the pituitary gonadotrophs causing inhibition of gonadotropin release, which subsequently causes the suppression of sex steroids in mammals [2]. This property of suppressing sex hormones renders the LHRH antagonists potentially useful in the treatment of endocrine-based diseases, such as prostate cancer, breast cancer, endometriosis, uterine leiomyoma and precocious puberty [2]. Intensive research has been focused on the development of potent and safe antagonists [3]. The relatively low potency and adverse effects due to histamine release have been the main obstacles to their acceptance and clinical use. In this sense, peptide antagonists with low histamine-release properties have been reported [4]. However, peptide antagonists still have their limitations, such as poor oral bioavailability. Recently, by conducting synthetic structure-activity relationship (SAR) studies, researchers have identified several non-peptide antagonists of the LHRH [5–8].

Non-peptide LHRH receptor antagonists were discovered by Cho et al [5]. These compounds are based on the introduction of crucial functional groups for receptor binding into a bicyclic scaffold that mimics a type II  $\beta$ -turn involving residues 5–8 (Tyr-Gly-Leu-Arg) of LHRH, which has been proposed as the bioactive conformation. Non-peptide antagonists resemble active peptides by keeping their main characteristics: a hydrophobic scaffold

**Electronic supplementary material** Supplementary material is available in the online version of this article at <http://dx.doi.org/10.1007/s00894-006-0163-6> and is accessible to authorized users.

M. Fernández · J. Caballero (✉)  
Molecular Modeling Group, Center for Biotechnological Studies,  
University of Matanzas,  
Matanzas C.P. 44740, Cuba  
e-mail: julio@fq.uh.cu

that interacts with a hydrophobic ligand-binding pocket of the receptor, and a positively charged residue which interacts with Asp302 in the seventh transmembrane domain of the human receptor [9]. The LHRH receptor belongs to the G-protein-coupled receptor family and consists of seven transmembrane segments that presumably adopt an  $\alpha$ -helical conformation. The binding mode of non-peptide antagonists to the LHRH receptor has been studied by docking active compounds into the LHRH receptor. Since X-ray diffraction data of LHRH receptor are not available, a model of the human LHRH receptor was built by homology modeling [5].

The correlation of biological data with various molecular descriptors constitutes an important and widely used field of the application of Quantitative Structure-Activity relationships (QSARs) [10]. A QSAR model proposes a mathematically quantified and computerized form from the chemical structure. In this sense, the QSAR conserves resources and accelerates the process of development of new molecules for use as drugs. A crucial step in constructing the QSAR model is to find a set of molecular descriptors that represents variation in the structural characteristics of the molecules tested. A wide variety of descriptors for use in QSAR analysis has been reported [11]. Many descriptors reflect simple molecular properties, and can thus provide insight into the physicochemical nature of the activity under consideration. Quantum chemical calculations are a reliable source of molecular descriptors, which can, in principle, express all of the electronic and geometric properties of molecules and their interactions [12]. In the performance of a QSAR analysis, quantum chemistry provides an accurate and detailed description of electronic effects.

In a recent paper, we conducted the first QSAR analysis on non-peptide antagonists of LHRH including thieno[2,3]pyridine-4-ones, thieno[2,3]pyrimidine-2,4-diones, imidazo[1,2]pyrimidin-5-ones and benzimidazole derivatives [13]. We first applied the multiple linear regression (MLR) analysis method, and then Bayesian-regularized genetic neural networks (BRGNNs) were used for the QSAR study with 2D-autocorrelation descriptors. The BRGNN approach overcame the limitations of linear methods. In the current contribution, we have performed the same analysis on recently reported macrolide LHRH antagonists: 11,12-cyclic carbamate derivatives of 6-*O*-methylerythromycin A (Fig. 1) [14, 15]. Macrolide LHRH antagonists differ considerably from previous non-peptide antagonists; their structures are based on cyclic decapeptides [14]. There are no previous reports of theoretical models to account for the physico-chemical interactions responsible for the activity of these compounds. With this in mind, we explored a pool of quantum chemical descriptors for inspecting the electronic features of these

molecules relevant for antagonistic activity. Optimum variable subsets of descriptors were selected using linear and nonlinear Genetic Algorithm (GA) searches. Both MLR and BRGNN techniques were used for modeling the observed activity of the training set (32 compounds). The adequacy of the models was examined by means of their statistical significance, leave-one-out (LOO) cross-validation and the quality of prediction of a test set (6 compounds).

## Methods

### Biological activity data set

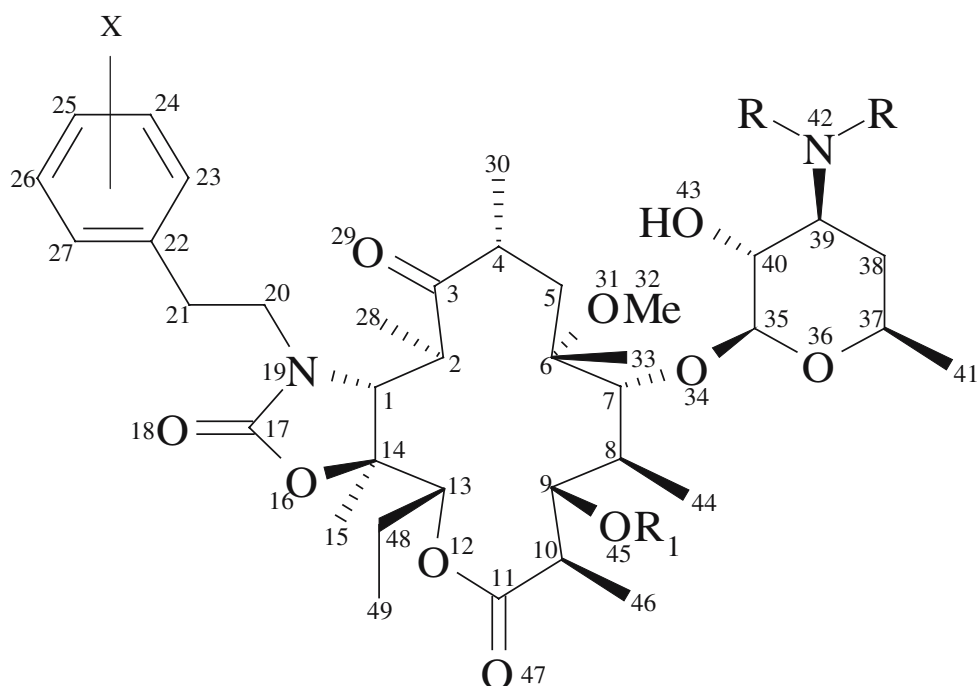
A data set of 38 macrolide LHRH antagonists was collected from the literature [14, 15]. Their *in vitro* binding affinities for human LHRH receptors cloned in CHO cells were expressed as  $pK_i$  values. The structural features and the biological activities of the compounds used in this study are shown in Fig. 1 and Table 1.

### Quantum chemical descriptors

The molecular structures of all the macrolide LHRH antagonists were built using the Hyperchem software [16], and geometrically optimized using the semiempirical quantum-chemical method PM3 [17] implemented in Gaussian98 [18]. The calculated descriptors for each molecule are summarized in Table 2. We selected the common structure between all the compounds studied, and calculated the net atomic Coulson charges ( $Q_A$ ) and electrostatic potentials ( $P_A$ ) at the core of each atom A without considering the H atoms. In addition, we calculated the net charges of the most negative and most positive atoms ( $Q_{\min}$ ,  $Q_{\max}$ ), sum of the absolute charges on all atoms ( $\Sigma Q$ ), average of the absolute values of the charges on all atoms ( $Q_m$ ), sum of squares of charges on all atoms ( $\Sigma Q^2$ ), sum of squares of positive and negative charges [ $\Sigma Q^2(+)$ ,  $\Sigma Q^2(-)$ ], average of square of charges on all atoms ( $Q_m^2$ ), most negative and least negative electrostatic potentials ( $P_{\min}$ ,  $P_{\max}$ ), average of electrostatic potentials ( $P_m$ ), molecular dipole moment ( $\mu$ ), energies of the highest occupied (HOMO), and lowest unoccupied (LUMO) molecular orbitals ( $\epsilon_{\text{HOMO}}$ ,  $\epsilon_{\text{LUMO}}$ ). Quantum chemical indices of electronegativity ( $\chi$ ), hardness ( $\eta$ ), softness ( $S$ ), and electrophilicity ( $\omega$ ) were calculated according to the methods given in Table 2 [19].

In all, 115 descriptors were calculated. Descriptors that stayed constant or almost constant were eliminated, and pairs of variables with a correlation coefficient greater than 0.7 were classified as intercorrelated; only one of these was included in the models.

**Fig. 1** General structure and numbering of the 11,12-cyclic carbamate derivatives of 6-*O*-methylerythromycin A used in this study



### Modeling procedure

The MLR analysis was used to derive a QSAR model for studying the effect of minor structural changes separately. Before this MLR analysis, the correlation between the selected descriptors was examined, and those descriptors with low colinearity were considered for the QSAR study. The GA was used to select the most relevant set of descriptors. The resulting model was evaluated by LOO cross-validation.

In the models that encompassed all the compounds, the dataset was divided into training and test sets. Six compounds were chosen randomly as a prediction set and were used for external validation of the MLR and BRGNN models. The compounds in the external prediction set were reserved for validating potential models. For the development of the MLR and BRGNN models, the training sets included all the remaining 32 compounds.

Since many descriptors are available for QSAR analysis and only a subset of them is statistically significant in terms of correlation with biological activities, deriving an optimal QSAR model through variable selection needs to be addressed. Following Occam's Razor [20], we selected just the variables that contain the information necessary for the modeling, but nothing more. In this sense, linear and nonlinear GA searches were carried out in order to build the linear and nonlinear models. The quality of each model was demonstrated by the square multiple correlation coefficient ( $R^2$ ) and the standard deviation (s). The models with  $R$ -values above 0.8 were selected and tested in cross-validation experiments.

To demonstrate the absence of chance correlations further, we generated analogous models by GA search using random numbers instead of the pool of descriptors, and correlations between selected random numbers and real biological activities were established [21, 22]. This process was repeated 20 times, and the  $R^2$  of training and LOO cross-validation ( $R^2_{CV}$ ) of the resulting models were averaged and compared to the real model. The rationale behind this test is that the significance of the real QSAR model decreases if there is a significant chance correlation between the selected random numbers and the response variable.

### Artificial neural network regression procedure

BRGNN is a framework that combines Bayesian-regularized artificial neural networks (BRANNs) with GA feature selection [23]. Our BRGNN approach is a version of the So and Karplus GA feature selection method [24] incorporating Bayesian regularization.

Bayesian networks are optimal devices for solving learning problems. They diminish the inherent complexity of artificial neural networks (ANNs) because they are governed by Occam's Razor, as complex models are automatically self-penalizing under Bayes' rule. The Bayesian approach to ANN modeling considers all possible values of network parameters weighted by the probability of each set of weights. The BRANN method was designed by Mackay [25, 26] for overcoming the deficiencies of ANNs. The Bayesian approach yields an a posteriori distribution of network parameters  $P(w|D,H)$  from a prior

**Table 1** Experimental and predicted activities of 11,12-cyclic carbamate derivatives of 6-*O*-methylerythromycin A from the training and test sets

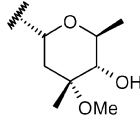
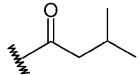
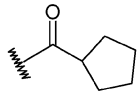
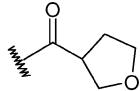
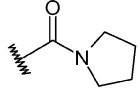
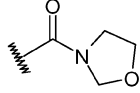
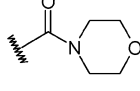
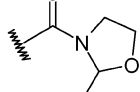
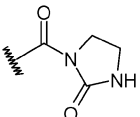
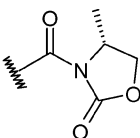
	X	R	R <sub>1</sub>	pK <sub>i</sub>		
				Exp	MLR (Eq. 2)	BRGNN
<b>Training set</b>						
1 <sup>a</sup>	4-Cl	Me, iPr		7.65	7.32	7.52
2 <sup>a</sup>	4-Cl	Me, iPr	H	5.00	5.04	5.48
3 <sup>a</sup>	4-Cl	Me, iPr		5.69	6.22	5.54
4 <sup>a</sup>	4-Cl	Me, iPr		6.20	6.42	6.27
5 <sup>a</sup>	4-Cl	Me, iPr		6.67	6.71	6.05
6 <sup>a</sup>	4-Cl	Me, iPr		5.79	6.03	6.11
7 <sup>a</sup>	4-Cl	Me, iPr		6.70	6.28	6.03
8 <sup>a</sup>	4-Cl	Me, iPr		5.25	5.78	5.78
9 <sup>a</sup>	4-Cl	Me, iPr		6.10	5.49	6.11
10 <sup>a</sup>	4-Cl	Me, iPr		5.69	6.73	6.40
11 <sup>a</sup>	4-Cl	Me, iPr		7.31	6.77	6.93

Table 1 (continued)

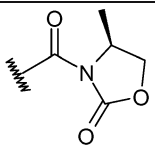
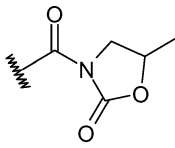
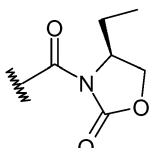
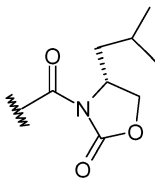
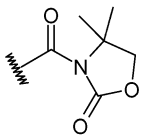
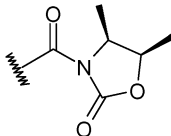
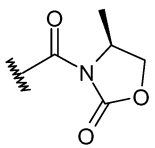
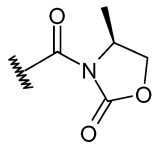
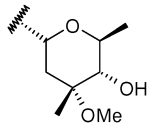
12 <sup>a</sup>	4-Cl	Me, iPr		7.73	7.42	6.91
13 <sup>a</sup>	4-Cl	Me, iPr		7.44	6.56	6.98
14 <sup>a</sup>	4-Cl	Me, iPr		7.18	7.48	6.93
15 <sup>a</sup>	4-Cl	Me, iPr		6.56	6.73	6.89
16 <sup>a</sup>	4-Cl	Me, iPr		7.10	7.08	7.18
17 <sup>a</sup>	4-Cl	Me, iPr		7.00	7.92	7.16
18	3,4-diCl	Me, cPent		8.24	8.61	8.40
19	3-Cl, 4-F	Me, cPrMe		9.48	9.22	9.37
20	4-Cl	Me, Me		6.72	7.20	7.47
21	4-Cl	Me, Me	H	5.00	4.99	5.36

Table 1 (continued)

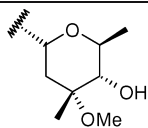
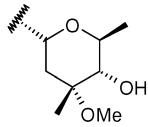
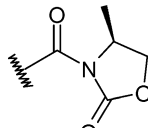
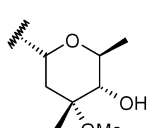
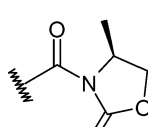
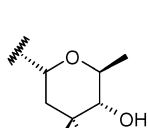
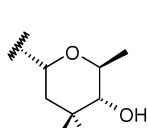
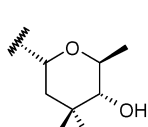
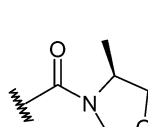
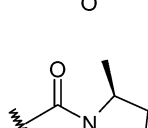
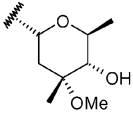
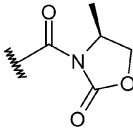
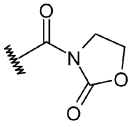
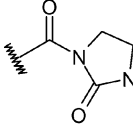
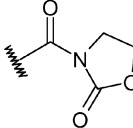
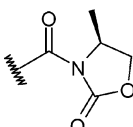
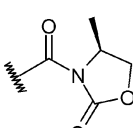
22	4-Cl	H, Me		7.10	7.05	7.14
23	4-Cl	Me, cPrMe		7.88	7.47	7.68
24	4-Cl	Me, cPrMe		7.73	7.52	7.78
25	4-Cl	Me, cPent		8.12	7.31	7.62
26	4-Cl	Me, cPent		8.41	7.53	7.59
27	4-Cl	Me, cHex		6.69	7.51	7.70
28	4-Cl	iBu, cPrMe		7.56	7.78	7.85
29	3,4-diCl	Me, cPent		8.82	8.57	8.43
30	3,4-diCl	Me, cHex		8.14	8.44	8.69
31	3,4-diCl	cPrMe, cPrMe		8.89	8.60	8.46

Table 1 (continued)

32	3-Cl, 4-F	cPrMe, cPrMe		9.07	9.11	9.09
<b>Test set</b>						
33	3,4-diCl	Me, cPrMe		8.73	8.90	8.59
34 <sup>a</sup>	4-Cl	Me, iPr		6.94	6.90	6.61
35 <sup>a</sup>	4-Cl	Me, iPr		5.33	6.27	5.12
36 <sup>a</sup>	4-Cl	Me, iPr		6.25	6.85	6.35
37	4-Cl	Me, cPen		7.42	7.77	7.40
38	3,4-diCl	Me, iPr		7.82	8.30	7.99

<sup>a</sup>Compounds included in model Eq. (1).

probability distribution  $P(w|H)$ , according to updates provided by the training set  $D$  using the BRANN model  $H$ . Predictions are expressed in terms of expectations with respect to this posterior distribution. Bayesian methods can simultaneously optimize the regularization constants in ANNs, a process that is very laborious using cross-validation. Instead of trying to find the global minimum, the Bayesian approach finds the (locally) most probable parameters (for in more detail, see Reference [13]).

The Bayesian approach produces predictors that are robust and well matched to the data. In BRANNs, these predictors are well suited for QSAR analysis [27, 28]. They give models that are relatively independent of the ANN architecture, above a minimum architecture, since the Bayesian regularization method estimates the number of effective parameters. Concerns about overfitting and over-

training are also eliminated by this method so that a definitive and reproducible model is produced. Joining a BRANN and GA feature selection (BRGNN) increases the possibilities of BRANNs for modelling, as we have indicated in previous work [13, 23, 29, 30]. This method is relatively fast and considers the whole dataset in the training process. For other hybrids of ANN and GA, the use of the mean square error (MSE) as the fitness function could lead to undesirably well fitted but poorly generalized networks as algorithmic solutions. In this connection, BRGNN avoids such results in two ways: (1) by keeping network architectures as simple as possible inside the GA framework, and (2) by implementing Bayesian regularization in the network training function.

Fully connected, three-layer BRANNs with back-propagation training were implemented in the MATLAB envi-

**Table 2** Symbols of the calculated quantum chemical descriptors used in this study and their definitions

Descriptor	Definition
$Q_A$	Net atomic Coulson charge at each atom A (heavy atoms) of the common structure between all the studied compounds.
$Q_{\min}, Q_{\max}$	Net charges of the most negative and most positive atoms.
$\Sigma Q$	Sum of absolute of charges on all atoms.
$Q_m$	Average of the absolute values of the charges on all atoms.
$\Sigma Q^2$	Sum of squares of charges on all atoms.
$\Sigma Q^2(+)$	Sum of squares of positive charges.
$\Sigma Q^2(-)$	Sum of squares of negative charges.
$Q_m^2$	Average of square of charges on all atoms.
$P_A$	The electrostatic potential at each atom A (heavy atoms) of the common structure between all the studied compounds.
$P_{\min}, P_{\max}$	Most negative and least negative electrostatic potentials.
$P_m$	Average of electrostatic potentials.
$\mu$	Molecular dipole moment.
$\epsilon_{\text{HOMO}}, \epsilon_{\text{LUMO}}$	Energies of the highest occupied (HOMO) and lowest unoccupied (LUMO) molecular orbitals.
$\chi$	Electronegativity: $-0.5 (\epsilon_{\text{HOMO}} - \epsilon_{\text{LUMO}})$
$\eta$	Hardness: $0.5 (\epsilon_{\text{HOMO}} + \epsilon_{\text{LUMO}})$
$S$	Softness: $1/\eta$
$\omega$	Electrophilicity: $\chi^2/2\eta$

ronment [31]. In these nets, the transfer functions of input and output layers were linear, and the hidden layer had neurons with a hyperbolic tangent transfer function. Inputs and targets took the values from independent variables selected by the GA and  $pK_I$  values, respectively; both were normalized prior to network training. BRANN training was carried out according to the Levenberg-Marquardt optimization [32]. The initial value for  $\mu$  was 0.005, with decrease and increase factors of 0.1 and 10, respectively. The training was stopped when  $\mu$  became larger than  $10^{10}$ .

The GA implemented in this paper retains the characteristics of that reported previously [23]. Initially, a set of 50 chromosomes was generated randomly. The population fitness was then calculated and the members were rank ordered according to fitness. The two best-scoring models were automatically retained as members for the next round of evolution. More progeny models were then created for the next generation by preferentially mating parent models with higher scores. Crossover operator and single-point mutations were used in the evolution process until the best MSE scoring model remains constant for at least ten generations. Our GA was programmed within the MATLAB environment using the genetic algorithm and neural networks toolboxes [31]. The predictors are BRANNs with

a simple architecture (two or three neurons in a single hidden layer). We tested the MSE of data fitting for BRANN models in some cases as the individual fitness function. The best models were selected according to their  $R$  value ( $R > 0.8$ ) and the results of cross-validation experiments (higher  $R^2_{\text{CV}}$ ).

## Results and discussion

### Multiple linear regression analysis

As shown in Table 1, there are 20 compounds with 4-Cl substituent on the phenethyl 11,12-cyclic carbamate and methyl and isopropyl substituents at the 3'-amino group of desosamine (Table 1). This set can be used for studying the properties that are relevant for the binding affinity when substituents at position 3 of the erythronolide core ( $R_1$ ) are varied. Equation (1) shows the best QSAR model selected by the GA (the search included 33 descriptors when constant and correlated descriptors were eliminated) for 20 derivatives:

$$pK_I = -0.564(\pm 0.250) \times P_{\max} + 449.761(\pm 135.080) \\ \times Q_{O16} - 203.358(\pm 46.102) \times Q_{O31} + 40.211(\pm 26.908) \quad (1)$$

$N=20$ ,  $R^2=0.723$ ,  $s=0.475$ ,  $F=13.915$ ,  $p=10^{-4}$ ;  $t(P_{\max} \text{ coeff})=-2.26$ ;  $t(Q_{O16} \text{ coeff})=3.33$ ;  $t(Q_{O31} \text{ coeff})=-4.41$ ;  $t(\text{intercept})=1.49$ ;  $R^2_{\text{CV}}=0.611$ ,  $s_{\text{CV}}=0.562$ , where  $n$  is the number of compounds included in the model,  $R^2$  is the square correlation coefficient,  $s$  is the standard deviation of the regression,  $F$  is the Fischer ratio and  $p$  is the significance of the variables in the model. The  $t$ -test values for the coefficients are also included.  $R^2_{\text{CV}}$  and  $s_{\text{CV}}$  are the correlation coefficient and standard deviation of the LOO cross-validation, respectively.

There is no significant intercorrelation between the selected descriptors (correlation matrixes are provided in the Supplementary material 1). The correlation between the calculated and experimental values of  $pK_I$  (from training and LOO cross-validation) is shown in Fig. 2a. The above equation shows a negative influence of the electrostatic term  $P_{\max}$  related to the least negative electrostatic potential on any atom in the whole molecule. This descriptor suggests that the change in the electrostatic surroundings due to the substituent  $R_1$  influences the binding affinity. The presence of local charge terms  $Q_{O16}$  and  $Q_{O31}$  suggests that the substituent  $R_1$  modulates the antagonist-receptor electrostatic interactions due to the O16 (which belongs to the cyclic carbamate) and O31 atoms (substituent at position 6 of the erythronolide core). Depending on the strength of these interactions, molecules increase or



decrease their affinities toward the human LHRH receptor. According to Eq. (1), binding is favored by the least negative charge on O16 and the most negative charge on O31.

As an additional validation, the GA search was performed using 33 random variables and real biological activities to evaluate the likelihood of chance correlations in this model. The averaged  $R^2$  and  $R^2_{CV}$  values when the random variables were changed 20 times and MLR analysis combined with GA search were applied to each generated data set, were  $R^2=0.565$  and  $R^2_{CV}=0.340$ . From this, we concluded that chance correlation had little effect in driving model development.

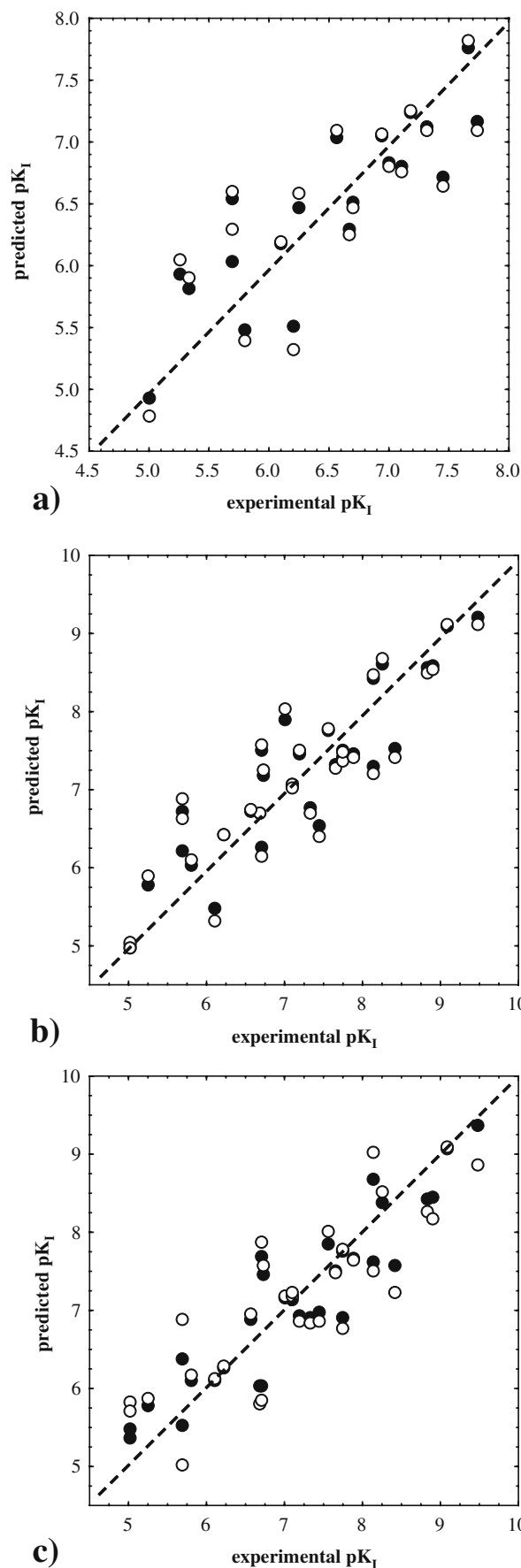
The best set of descriptors for the whole dataset was selected. MLR analysis combined with a linear GA feature selection (the search included 30 descriptors when constant and correlated descriptors were eliminated) was performed on the training set (Table 1). The following equation resulted:

$$pK_I = 245.208(\pm 42.010) \times \omega + 42.377(\pm 5.351) \times Q_{O45} + 207.729(\pm 745.042) \times Q_{C15} + 19.939(\pm 5.810) \times Q_{\min} + 124.536(\pm 27.581) \quad (2)$$

$N=32$ ,  $R^2=0.826$ ,  $s=0.530$ ,  $F=32.018$ ,  $p < 10^{-5}$ ;  $t(\omega \text{ coeff})=5.84$ ;  $t(Q_{O45} \text{ coeff})=7.92$ ;  $t(Q_{C15} \text{ coeff})=2.77$ ;  $t(Q_{\min} \text{ coeff})=3.43$ ;  $t(\text{intercept})=4.52$ ;  $R^2_{CV}=0.756$ ,  $s_{CV}=0.628$ .

The predictions of  $pK_I$  values for the 32 non-peptide antagonists using this equation are shown in Table 1. Equation (2) shows a four-descriptor model including the electrophilicity ( $\omega$ ), the local charges at atoms O45 and C15 and the charge of the most negative atom. It is noteworthy that there is no significant intercorrelation between these descriptors (see [Supplementary material](#)). The correlation between the calculated and experimental values of  $pK_I$  (from training and LOO-cross-validation) is shown in Fig. 2b. The above equation shows that when substituents on the phenethyl 11,12-cyclic carbamate, on the 3'-amino group of desosamine and the position 3 of the erythronolide core, are varied, the binding affinity for human LHRH receptor is modulated by electron-related terms. All the terms in this equation have positive coefficients. However, their effects are negative since such terms have negative values. According to this model, binding is favored by less negative charge on O45 (substituent at position 3 of the erythronolide core, which supports  $R_1$  substituent), less negative charge on C15

**Fig. 2** Scatter plot of the experimental activities versus predicted activities: (●) training predictions and (○) LOO cross-validated predictions for **a** MLR model Eq. (1), **b** MLR model Eq. (2) and **c** BRGNN model



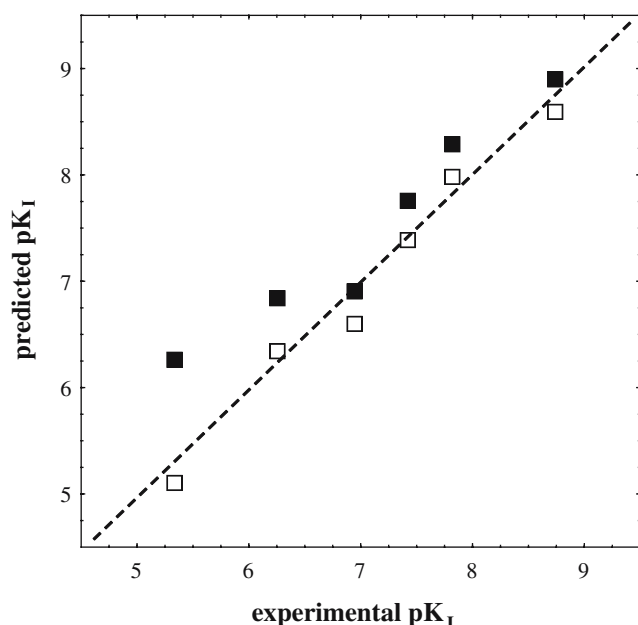
(substituent at position 12 of the erythronolide core), a low electrophilicity, and a low value of most negative charge in the whole molecule.

To investigate the performance of the MLR model, it was used to predict the activity of the six molecules of the test set. The results of the predictions are shown in Table 1 and in Fig. 3 as a scatter plot of the predicted versus experimental values. This analysis reveals that the proposed MLR model fails in the prediction of compound 35 [ $pK_I=6.27$  instead of 5.33; percent relative error (PRE)=17.6)] However, the remaining compounds were predicted adequately.

Ultimately, the likelihood of chance correlations in this model was evaluated by performing a GA search using a pool of 30 random variables and real biological activities. The averaged  $R^2$  and  $R^2_{CV}$  values when the random variables were changed 20 times and MLR analysis combined with GA search were applied to each generated data set, were  $R^2=0.478$  and  $R^2_{CV}=0.272$ . From this, we concluded that chance correlation had little effect in driving model development.

#### Bayesian-regularized genetic neural networks

The BRGNN technique was used in order to discover the possible existence of nonlinear relationships between binding affinity and molecular descriptors. The best model selected by the BRGNN method includes four descriptors and is shown in Table 3. The predictions of  $pK_I$  values for the 32 non-peptide antagonists using this model are shown in Table 1 and scatter plots of the predicted (from training



**Fig. 3** Scatter plot of the experimental activities versus predicted activities for the test set: (■) MLR and (□) BRGNN models

and LOO cross-validation) versus the experimental values in Fig. 2c. The BRGNN approach yielded an optimum variable subset that was similar to the descriptor subset selected by the linear GA. The nonlinear model keeps the electrophilicity ( $\omega$ ) and the local charge at C15 ( $Q_{C15}$ ) as relevant properties. The new properties included are the local charges at O31 and O43 ( $Q_{O31}$  and  $Q_{O43}$ ). Similarly to the variables selected by MLR, there is no significant intercorrelation between these descriptors (see Supplementary material).

The nonlinear model improved on the MLR one by fitting the training set with a higher  $R^2$  of 0.848 in comparison with 0.826 for the linear model. When we compared the predictivity of linear and nonlinear models according to LOO cross-validation experiments, we found that MLR ( $R^2_{CV}=0.756$ ; Eq. 2) is superior to BRGNN ( $R^2_{CV}=0.700$ ). However, some authors have cast doubt on the results of internal validation [33–35]. Golbraikh and Tropsha [33] found that there is no relationship between  $R^2_{CV}$  and the model's ability to predict an external test set. Aptula et al. [34] demonstrated that  $R^2_{CV}$  is not a good criterion for evaluating model predictivity; instead of this, they consider that root mean square error of test set predictions is a better fitness criterion. Doweiko [35] considers that a higher  $R^2_{CV}$  reflects that the model identified the redundancy in the training set and this has nothing to do with predictivity.

Predicting the activity of some components in a test set is the most reliable way to establish the predictivity of a QSAR model. To investigate the performance of the nonlinear model, it was used to predict the activity of the six molecules of the test set. The results of the predictions are shown in Table 1 and in Fig. 3 as scatter plot of the experimental activities versus predicted activities. This analysis reveals that the proposed model has a high prediction ability with low PREs between  $-4.8$  and  $2.1$  for the test set. In this sense, the BRGNN procedure improves the linear results.

As in linear model, the GA search was performed using 30 random variables and real biological activities to evaluate the likelihood of chance correlations in this model. The averaged  $R^2$  and  $R^2_{CV}$  values when the random variables were changed 20 times and BRGNN analysis was applied to each generated data set, were  $R^2=0.503$  and  $R^2_{CV}=0.164$ . From this we concluded that chance correlation had little effect in driving nonlinear model development. Furthermore, we examined if the choice of test set influences our results. We divided the data set on several randomly constructed training/test set partitions of 32 and 6 compounds, respectively, and computed the averaged PREs and  $R^2$  values for test sets using MLR and BRGNN approaches. The linear approach gives an averaged PRE=7.90, while BRGNN approach gives a lower value PRE=

**Table 3** Statistic parameters of the best BRGNN model for prediction of  $pK_I$  of the macrolide LHRH antagonists

Model	Descriptors	Hidden neurons	$n$	$R^2$	$s$	$R^2_{CV}$	$s_{CV}$
BRGNN	$\omega$ , $Q_{O31}$ , $Q_{O43}$ , $Q_{C15}$	2	32	0.848	0.456	0.700	0.641

7.44. Otherwise, the averaged value of  $R^2$  of test sets is higher for BRGNN approach:  $R^2=0.807$  in comparison to  $R^2=0.777$  of the linear approach. This result clearly demonstrates that our results do not depend on the training/test set partition selected.

In order to gain a deeper insight into the relative effects of each quantum chemical descriptor in our model, a recently reported weight-based input-ranking scheme was carried out. The black-box nature of three layer ANNs has been “deciphered” in a recent report of Guha et al. [36] Their method allows an understanding of how an input descriptor is correlated to the predicted output by the network and consists of two parts. First, the nonlinear transform for a given neuron is linearized. Afterwards, the magnitude in which a given neuron affects the downstream output is determined. Next, a ranking scheme for neurons in the hidden layer is developed. The ranking scheme is carried out by determining the Square Contribution Values (SCV) for each hidden neuron (see Reference [36] for details). This method for ANN model interpretation is similar to the partial least squares interpretation method for linear models described by Stanton [37].

The results of the ANN deciphering study are shown in Table 4. The reported effective weight matrix for our model shows that the first hidden neuron has the major contribution to the model with a SCV value 6-fold higher than the second hidden neuron. On this neuron,  $Q_{O31}$  has the highest impact equal to  $-2.435$ . From this analysis, we can also derive the approximate effect of the selected descriptors. The sign of the weights indicates the trend of the output value. According to the sign of the effective weight,  $Q_{O31}$  has a negative influence in  $pK_I$ , which signifies that binding is favored by a high negative charge on O31 (substituent at position 6 of the erythronolide core). Following the same procedure, we conclude that a lower negative charge on O43 (hydroxyl group of desosamine residue) is required for high  $pK_I$  values. As in the linear model, BRGNN model showed that a less negative electrophilicity increases the antagonistic activity. However,  $Q_{C15}$  showed opposing effects in first and second neurons, which suggests a complex nonlinear effect not perceived by the linear model. The electronic features described by this model suggest that the potency of 11,12-cyclic carbamate derivatives of 6-*O*-methylerythromycin A as LHRH antagonists is related to some negatively charged substituents. In general, the structural pattern repeated in all compounds of the dataset modeled (Fig. 1) contains ten oxygen atoms that contribute

with electron density. Inside this active scaffold, our model details that increase in electron density at some positions of the scaffold reinforces the interactions with LHRH receptor.

## Conclusions

MLR and BRGNN were used to model the biological activity of recently reported macrolide LHRH antagonists. In the application of the MLR procedure, we found a predictive equation for studying the effects of varying the substituent at position 3 of the erythronolide core. In addition, a four-parameter equation resulted for the entire set of compounds.

The BRGNN model also contains four descriptors. The results of the BRGNN approach seem to be more reliable according to the external validation. The MLR model fails in the prediction of compound 35 (PRE=17.6) in the test set. However, the BRGNN model predicts the activities of all compounds of the test set with a relative error lower than 4.8%. The better results achieved by the BRGNN approach suggest that structure-LHRH antagonistic activity relationship is a complex phenomenon that can be described appropriately by nonlinear analysis and the chemical characteristics it captures.

## Supporting information available

Descriptors selected by GA for MLR Eq. (1), MLR Eq. (2) and BRGNN model and correlation matrix of these descriptors in such models.

**Table 4** Effective weight matrix for the optimum BRGNN model

Network	Hidden neurons	
	1	2
Inputs		
$\omega$	0.228	0.597
$Q_{O31}$	<b>-2.435</b>	<b>0.798</b>
$Q_{O43}$	1.620	-0.002
$Q_{C15}$	-0.691	0.307
SCV <sup>a</sup>	0.862	0.138

The most relevant descriptors appear in bold

<sup>a</sup> The columns are ordered by the SCVs for the hidden neurons, shown in the last row.

## References

- Huirne JAF, Lambalk CB (2001) *Lancet* 358:1793–1803
- Filicori M, Flamigni C (1988) *Drugs* 35:63–82
- Kutscher B, Bernd M, Beckers T, Polymeropoulos EE, Engel J (1997) *Angew Chem Int Ed* 36:2148–2161
- Karten MJ (1992) An overview of GnRH antagonist development: Two decades of progress. In: Crowley Jr WF, Conn PM (eds) *Modes of Action of GnRH and GnRH Analog*. Elsevier, New York, pp 277–297
- Cho N, Harada M, Imaeda T, Imada T, Matsumoto H, Hayase Y, Sasaki S, Furuya S, Suzuki N, Okubo S, Ogi K, Endo S, Onda H, Fujino M (1998) *J Med Chem* 41:4190–4195
- De Vita RJ, Hollings DD, Goulet MT, Wyvratt MJ, Fisher MH, Lo JL, Yang YT, Cheng K, Smith RG (1999) *Bioorg Med Chem Lett* 9:2615–2620
- Chu L, Hutchins JE, Weber AE, Lo JL, Yang YT, Cheng K, Smith RG, Fisher MH, Wyvratt MJ, Goulet MT (2001) *Bioorg Med Chem Lett* 11:509–513
- Zhu YF, Struthers RS, Connors Jr PJ, Gao Y, Gross TD, Saunders J, Wilcoxon K, Reinhart GJ, Ling N, Chen C (2002) *Bioorg Med Chem Lett* 12:399–402
- Flanagan CA, Becker II, Davidson JS, Wakefield IK, Zhou W, Sealfon SC, Millar RP (1994) *J Biol Chem* 269:22636–22641
- Gasteiger J (2006) *Anal Bioanal Chem* 384:57–64
- Todeschini R, Consonni V (2000) *Handbook of molecular descriptors*. Wiley, Weinheim
- Karelson M, Lobanov VS, Katritzky AR (1996) *Chem Rev* 96:1027–1043
- Fernández M, Caballero J (2006) *J Mol Graph Modell* 25:410–422
- Randolph JT, Waid P, Nichols C, Sauer D, Haviv F, Diaz G, Bammert G (2004) *J Med Chem* 47:1085–1097
- Randolph JT, Sauer DR, Haviv F, Nilius AM, Greer J (2004) *Bioorg Med Chem Lett* 14:1599–1602
- HyperChem 7.0 (2002) Hypercube, Gainesville
- Stewart JJP (1989) *J Comput Chem* 10:210–220
- Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zarkzewski VG, Montgomery JA, Stratmann RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Baone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Peterson GA, Ayala PY, Cui Q, Morokuma K, Makick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Baboul AG, Stefanov BB, Liu G, Kiashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-laham MA, Peng CY, Nanayakkara A, González C, Challacombe M, Gill PMW, Johnson BG, Chen W, Wong MW, Andres JL, Head-Gordon M, Repogle ES, Pople JA (1998) *Gaussian 98, Revision A1*. Gaussian, Pittsburgh, PA
- Thanikaivelan P, Subramanian V, Rao JR, Nair BU (2000) *Chem Phys Lett* 323:59–70
- Hawkins DM (2004) *J Chem Inf Comput Sci* 44:1–12
- Topliss JG, Costello RJ (1972) *J Med Chem* 15:1066–1068
- Topliss JG, Edwards RP (1979) *J Med Chem* 22:1238–1244
- Caballero J, Fernández M (2006) *J Mol Model* 12:168–181
- So SS, Karplus M (1996) *J Med Chem* 39:1521–1530
- Mackay DJC (1992) *Neural Comput* 4:415–447
- Mackay DJC (1992) *Neural Comput* 4:448–472
- Burden FR, Winkler DA (1999) *J Med Chem* 42:3183–3187
- Winkler DA, Burden FR (2004) *Biosilico* 2:104–111
- Fernández M, Tundidor-Camba A, Caballero J (2005) *J Chem Inf Model* 45:1884–1895
- Caballero J, Garriga M, Fernández M (2005) *J Comput-Aided Mol Des* 19:771–789
- MATLAB 7.0 (2004) The Mathworks Inc, 3 Apple Hill Drive, Natick, MA 01760–2098, USA
- Foresee FD, Hagan MT (1997) Gauss-Newton approximation to Bayesian learning. *Proceedings of the 1997 International Joint Conference on Neural Networks*. IEEE, Houston, pp 1930–1935
- Golbraikh A, Tropsha A (2002) *J Mol Graph Modell* 20:269–276
- Aptula AO, Jeliaskova NG, Schultz TW, Cronin MTD (2005) *QSAR Comb Sci* 24:385–396
- Doweyko AM (2004) *J Comput-Aided Mol Des* 18:587–596
- Guha R, Stanton DT, Jurs PC (2005) *J Chem Inf Model* 45:1109–1121
- Stanton DT (2003) *J Chem Inf Comput Sci* 43:1423–1433