

# 2D and 3D Quantitative Structure–Activity Relationship Study of Hepatitis C Virus NS5B Polymerase Inhibitors by Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis Methods

Eslam Pourbasheer,<sup>\*,†</sup> Reza Aalizadeh,<sup>‡</sup> Samira Shokouhi Tabar,<sup>†</sup> Mohammad Reza Ganjali,<sup>§,||</sup> Parviz Norouzi,<sup>§,||</sup> and Javad Shadmanesh<sup>⊥</sup>

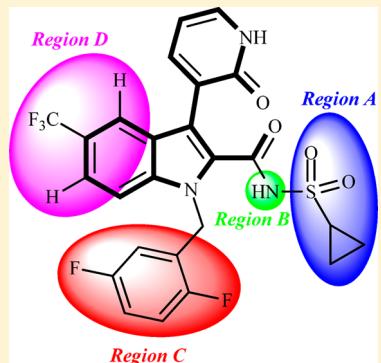
<sup>†</sup>Department of Chemistry, Payame Noor University (PNU), P. O. Box 19395-3697, Tehran, Iran

<sup>‡</sup>Laboratory of Analytical Chemistry and <sup>⊥</sup>Laboratory of Inorganic Chemistry and Technology, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece

<sup>§</sup>Center of Excellence in Electrochemistry, Faculty of Chemistry, University of Tehran, P.O. Box 143981-7435, Tehran, Iran

<sup>||</sup>Biosensor Research Center, Endocrinology and Metabolism Molecular-Cellular Sciences Institute, Tehran University of Medical Sciences, P. O. Box, 14114-13137, Tehran, Iran

**ABSTRACT:** In this present study, three-dimensional quantitative structure–activity relationship (3D-QSAR) and 2D-QSAR analyses were performed on the series of compounds Hepatitis C Virus NS5B polymerase inhibitors using comparative molecular field analysis (CoMFA), comparative molecular similarity indices analysis (CoMSIA), and stepwise multiple linear regression (SW-MLR) approaches. A CoMFA model with good predictive ability was generated based on training set of 39 compounds and showed satisfactory statistical results ( $q^2 = 0.600$ ,  $r^2 = 0.871$ ). To improve the contribution of points for next analyses, CoMFA (after region focusing) was employed in biases of similar alignment and showed appropriate predictive results ( $q^2 = 0.691$ ,  $r^2 = 0.889$ ). A reliable CoMSIA model out of 31 different combinations with the higher “leave-one-out” cross-validation correlation coefficients ( $q^2$ ) were obtained and indicated suitable statistical results ( $q^2 = 0.664$ ,  $r^2 = 0.911$ ). An external test set of nine compounds were used to evaluate the predictive ability of generated models. The 2D-QSAR model was built with the four descriptors selected by stepwise technique and presented high predictive ability ( $R_{\text{train}}^2 = 0.833$ ,  $R_{\text{test}}^2 = 0.773$ ,  $Q_{\text{LOO}}^2 = 0.758$ ,  $Q_{\text{BOOT}}^2 = 0.736$ ). The derived contour maps from each model were assessed to identify the significant structural features required for improving biological activity so as to design potent HCV NS5B polymerase inhibitors.



## 1. INTRODUCTION

The Hepatitis C virus (HCV) was first identified in 1989 as a member of Flaviviridae family and known to be the result of infectious agent of non-A, non-B viral hepatitis.<sup>1,2</sup> HCV infection causes chronic liver diseases, cirrhosis, and hepatocellular carcinoma.<sup>3</sup> HCV contains a positive single-stranded RNA genome that encodes a large poly protein of 3010 amino acids.<sup>4</sup> This poly protein subsequently encodes the proteases for manufacture of proteins with at least 10 functional viral proteins: core, envelope 1, envelope 2, p7, nonstructural protein 2 (NS2), NS3, NS4A, NS4B, NSSA, and NSSB.<sup>5,6</sup> Viral genome replication occurring in HCV infection with high prevalence of the disease<sup>3</sup> made the treatment difficult to produce vaccine, and therefore, some novel effective therapies are urgently required.<sup>7,8</sup> Among the nonstructured proteins, the NSSB, which is an RNA-dependent RNA polymerase (RdRP),<sup>9,10</sup> a core enzyme required for HCV replication,<sup>10</sup> has attracted much attentions due to its role in synthesis of the negative strand copy of the RNA genome, and also the positive strand RNA copy that is included in the new progeny virus.<sup>10–14</sup> In this contribution, several approaches have

been developed to target the NSSB enzyme so as to diminish and inhibit the activity of virus.<sup>15–23</sup>

Various novel therapies have been developed to target viral translation and replication in HCV.<sup>3,7,24–26</sup> Recently, Chen and his co-workers<sup>27</sup> suggested some novel compounds as HCV NSSB polymerase inhibitors. The proposed inhibitors were based on idole moiety and could reveal potent biological activity. Consequently, they introduce some substituents as the leading effects so as to increase the potency, and any significant enhancement observed by replacing different groups was taken to be constant to obtain potent inhibitors. However, further assessment could be done to understand the significance of different varieties of groups on inhibitory activity. Considering compounds 13–24 in the literature,<sup>27</sup> it can be seen that there are more comparisons required to understand the effect of substituents R<sub>4</sub> and R<sub>6</sub> on inhibitory activity while they remained constant. Also there was a lack of information about the groups

Received: April 8, 2014

**Table 1.** Chemical Structures of HCV NS5B Polymerase Inhibitors with Their Experimental and Predicted Activities

N.o.	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	X	Exp. (pIC <sub>50</sub> )	CoMFA-1	CoMFA-2	CoMSIA	Predicted SW-MLR	
1 <sup>b</sup>	H	Cl	H	H	3-NH <sub>2</sub>	N	7.28	7.25	7.32	7.38	7.20
2 <sup>b</sup>	Cl	Cl	H	3-NH <sub>2</sub>	N	6.12	6.98	7.06	6.83	6.71	
3 <sup>a</sup>	Br	H	H	3-NH <sub>2</sub>	N	5.40	6.98	7.06	6.33	5.72	
4	H	Cl	Cl	3-NH <sub>2</sub>	N	7.33	7.25	7.33	7.33	7.23	
5 <sup>a</sup>	H	Br	H	3-NH <sub>2</sub>	N	7.36	7.26	7.33	7.59	7.45	
6	H			H	3-NH <sub>2</sub>	N	5.15	4.89	4.89	5.10	5.78
7	H			H	3-NH <sub>2</sub>	N	7.19	7.19	7.14	7.16	6.88
8	H			H	3-NH <sub>2</sub>	N	5.85	6.46	6.26	5.81	6.11
9 <sup>a</sup>	H			H	3-NH <sub>2</sub>	N	6.48	6.82	6.63	6.75	6.75
10	H			H	3-NH <sub>2</sub>	N	7.32	7.08	7.15	7.04	6.86
11	H	OCF <sub>3</sub>	H	3-NH <sub>2</sub>	N	7.80	7.40	7.49	7.61	7.15	
12	H	CF <sub>3</sub>	H	3-NH <sub>2</sub>	N	7.77	7.51	7.58	7.63	7.65	
13	H	CF <sub>3</sub>	H	4-OCH <sub>3</sub>	C	7.60	7.40	7.44	7.31	7.63	
14	H	CF <sub>3</sub>	H	4-OCHF <sub>2</sub>	C	7.17	7.20	7.27	6.83	7.53	
15	H	CF <sub>3</sub>	H	2-CH <sub>3</sub>	C	8.05	8.06	8.04	7.97	8.19	
16	H	CF <sub>3</sub>	H	2-F	C	8.05	8.06	8.04	8.09	8.00	
17	H	CF <sub>3</sub>	H	2,4-CH <sub>3</sub>	C	7.80	7.97	7.95	7.86	8.01	
18	H	CF <sub>3</sub>	H	2,4-F	C	7.60	8.01	7.98	7.78	8.03	
19 <sup>b</sup>	H	CF <sub>3</sub>	H	2-F, 3-CH <sub>3</sub>	C	8.30	8.15	8.10	8.15	8.78	
20 <sup>b</sup>	H	CF <sub>3</sub>	H	2,5-F	C	8.40	8.29	8.29	8.32	8.10	
21	H	CF <sub>3</sub>	H	2-F, 5-CONH <sub>2</sub>	C	8.15	8.42	8.34	8.44	7.85	
22 <sup>a,b</sup>	H	OCH <sub>3</sub>	H	2,5-F	C	7.62	8.17	8.24	8.29	7.58	
23 <sup>a</sup>	H	OCH <sub>3</sub>	H	2,5-F	C	7.82	8.41	8.51	7.59	7.56	
24 <sup>b</sup>	H	OEt	H	2,5-F	C	7.10	7.59	7.41	7.52	7.47	
25	H			H	2,5-F	C	6.82	6.73	6.67	7.36	7.37
26	H	CH <sub>3</sub>	H	2,5-F	C	8.10	8.03	8.08	7.89	7.95	
27 <sup>a</sup>	H	Et	H	2,5-F	C	8.40	8.23	8.30	8.06	7.91	
28	H	Br	H	2,5-F	C	8.52	8.02	8.07	8.27	8.84	
29	H	CH <sub>2</sub> CF <sub>3</sub>	H	2,5-F	C	8.10	7.93	7.94	8.12	7.95	
30 <sup>a</sup>	H	tert-Bu	H	2,5-F	C	8.10	7.79	7.78	8.25	7.52	
31 <sup>b</sup>	H			H	2,5-F	C	8.05	7.68	7.65	7.97	8.07
32	H	CH <sub>3</sub>	CH <sub>3</sub>	2,5-F	C	8.15	8.14	8.12	8.09	8.17	
33	H	Et	CH <sub>3</sub>	2,5-F	C	8.00	8.16	8.17	8.09	7.90	
34	H	iso-Propyl	CH <sub>3</sub>	2,5-F	C	8.15	8.22	8.18	8.09	8.10	
35	H	Cyclopropyl	CH <sub>3</sub>	2,5-F	C	8.22	8.18	8.19	8.10	8.53	
36 <sup>b</sup>	H	CH <sub>3</sub>	Et	2,5-F	C	8.22	8.33	8.35	8.28	7.88	
37	H	Et	Et	2,5-F	C	8.22	8.29	8.30	8.30	7.78	
38	H	iso-Propyl	Et	2,5-F	C	8.30	8.32	8.34	8.31	7.83	
39 <sup>a</sup>	H	Cyclopropyl	Et	2,5-F	C	8.52	8.34	8.38	8.29	8.47	
40	H	Cyclopropyl	CF <sub>3</sub>	2,5-F	C	8.52	8.39	8.44	8.49	8.25	
41	H	Cyclopropyl	tert-Bu	2,5-F	C	8.22	7.97	7.93	8.48	8.02	
42 <sup>a</sup>	H	Cyclopropyl			2,5-F	C	8.22	7.96	7.94	8.44	8.67
43 <sup>b</sup>	H	Cyclopropyl	CH <sub>2</sub> CF <sub>3</sub>	2,5-F	C	8.15	7.98	7.94	8.28	8.38	
44 <sup>b</sup>	F	Cyclopropyl	CH <sub>3</sub>	2,5-F	C	8.10	8.14	8.13	7.93	8.65	
45	Cl	Cyclopropyl	CH <sub>3</sub>	2,5-F	C	8.30	8.17	8.20	8.04	8.72	
46	CF <sub>3</sub>	Cyclopropyl	CH <sub>3</sub>	2,5-F	C	8.15	8.24	8.26	8.13	7.98	
47	F	Cyclopropyl	CF <sub>3</sub>	2,5-F	C	8.15	8.31	8.30	8.29	8.52	
48	F	Cyclopropyl	Et	2,5-F	C	8.30	8.40	8.45	8.12	8.53	

<sup>a</sup>Test set for 3D-QSAR. <sup>b</sup>Test set for 2D-QSAR.

containing nitrogen in some compounds given in series of 34–43.<sup>27</sup> Therefore, further studies should be performed to understand the combination effect of different groups in different substituents.

The novel designed compounds are needed to be tested in biological materials such as human or calf origin in order to identify their biological activities. However, performing these experiments is costly and time-consuming and may result in producing of toxic side effects. Regarding this issue, computer-aided drug design approaches can be employed to obtain the biological activity of newly designed compounds. One of the well-known computational methods used widely in drug design is quantitative structure–activity relationships (QSARs).<sup>28–32</sup> In this method, the biological activities of compounds can be predicted based on the molecular properties.<sup>29,33</sup> The QSAR method can lead to generation of a prediction model to identify the key chemical features associated with enhancing the inhibitory activities, but it considers neither the 3D structures of the compounds nor their chirality. Attempts to consider these properties in computational drug design methodology resulted in the generation of a new approach termed the three-dimensional quantitative structure–activity relationship (3D-QSAR) method.<sup>34,35</sup> Employing the 3D-QSAR method would result in identification of regions in space where the interactive fields may affect the biological activities.<sup>36–39</sup> The predictive 3D-QSAR models can be derived using comparative molecular field analysis (CoMFA)<sup>40</sup> and comparative molecular similarity indices analysis (CoMSIA)<sup>41</sup> methods. Performing these methods would lead to graphical visualization of key chemical structural features attributed to enhance the biological activities. The obtained graphical results are defined as the contour maps. The steric and electrostatic characteristics in 3D Cartesian space around the aligned molecules are calculated using the CoMFA method,<sup>40</sup> whereas the CoMSIA method represents the similarity indices which results in additional contour maps including hydrophobic and hydrogen-bonding donor and acceptor.<sup>41</sup> These additional fields can provide extra information for understanding the effect of chemical structure features and make the interpretation of results easier.

HCV NS5B polymerase is a very important biological target for the prevention of viral genome replication.<sup>15–23</sup> To the best of our knowledge, there has not been very much data reported concerning the 3D-QSAR study on HCV NS5B polymerase. In this work, the 3D-QSAR models have been established for a series of compounds as HCV NS5B polymerase inhibitors to derive the contour maps, in order to explain the impact of physicochemical and structural parameters of compounds on HCV NS5B polymerase inhibition activity. In addition to the 3D-QSAR study, 2D-QSAR study has been derived to compare and interpret the attribution of chemical structures on biological activity.

## 2. MATERIALS AND METHODS

**2.1. Data Set.** Molecular structures and biological activities of 48 compounds as HCV NS5B polymerase inhibitors were taken from the literature.<sup>27</sup> The biological activities as  $IC_{50}$  ( $\mu M$ ) values were converted to logarithmic scale  $pIC_{50}$  (M) to give numerically larger data values. The  $pIC_{50}$  values were then used as the dependent variables in subsequent CoMFA and CoMSIA analyses. The data set was split into the training and test

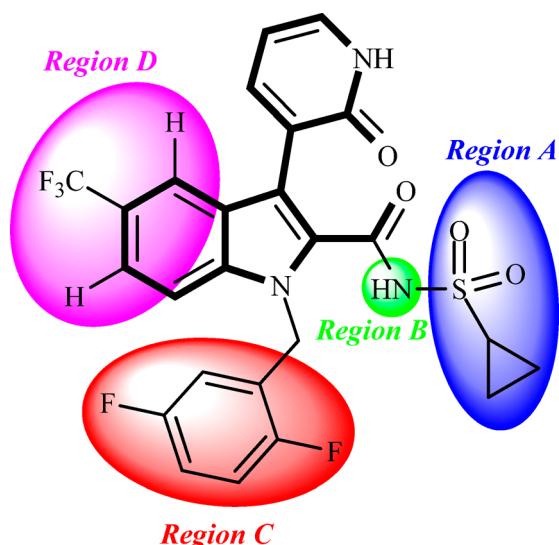
sets considering the principle of distribution in the range of the biological data for the both sets and their chemical structures diversity. A training set of 39 compounds (80%) was taken to construct the 3D-QSAR models, and the test set consisting of 9 compounds (20%) was employed to evaluate the external prediction ability of built models. The chemical structures for whole compounds along with their biological activities are given in Table 1. The test set compounds are marked and shown in Table 1.

**2.2. 2D-QSAR Method.** The 2D structure of whole compounds in the taken data set was drawn in Hyperchem v7.3<sup>42</sup> and then preoptimized using the molecular mechanic force field (MM+). The final optimizations of compounds were carried out by the semiempirical method (AM1) using the Polak–Ribiere algorithm until the root mean-square gradient was 0.01 kcal mol<sup>-1</sup>. Then, the optimized compounds were used to calculate the theoretical descriptors by employing the DRAGON software.<sup>43</sup> The constant and near constant descriptors were removed, and then, the correlations between remaining descriptors and inhibitory activities were calculated. The collinear descriptors ( $r > 0.9$ ) were then distinguished, and among the collinear descriptors, the one presenting the highest correlation with the inhibitory activity was retained and the others removed from the data set. The total numbers of 401 descriptors were remained for subsequent analyses.

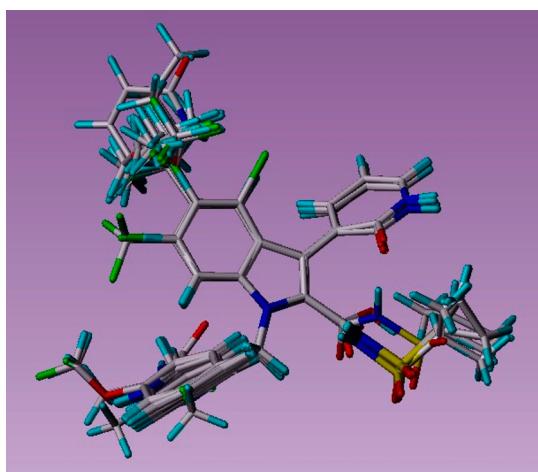
**2.2.1. Stepwise Multiple Linear Regression (MLR) Modeling.** The most relevant descriptors were selected using a stepwise variable selection tool. This variable selection method was first introduced by Hocking.<sup>44</sup> It selects the variables which are statistically meaningful descriptors. The selection process is an automatic procedure, and generally, the technique of F-test was taken into consideration for selecting variables.<sup>45</sup> Stepwise regression analysis generally follows two procedures including forward selection and backward elimination. The forward procedure starts by selecting a descriptor which has the highest correlation with biological activity, and then, the other variables are added to the model. The significance and improvement of the model for added descriptors were then evaluated using the statistical F-test method. In backward elimination methodology, all descriptors were selected, and then, by using the F-test method, the improvement on the model by eliminating the certain subset of descriptors was inspected. The stepwise program was written in MATLAB 6.5. MLR analysis combined with a stepwise feature selection was applied on the training set, and the best set of descriptors was selected.

**2.3. 3D-QSAR Method.** **2.3.1. Molecular Modeling and Alignment.** The three-dimensional chemical structures of the molecules were drawn by molecular modeling package SYBYL 7.3.<sup>46</sup> Energy minimization was performed using Tripos force field with a dependent dielectric and Powell conjugate gradient algorithm convergence criterion of 0.01 kcal/mol·Å.<sup>47</sup> To calculate the partial atomic charges, the Gasteiger–Hückel method<sup>48</sup> was employed. Since the predictive capability and strength of the built models are directly dependent on alignment of molecules with a reference compound, selecting the template molecule is one of the important steps in performing 3D-QSAR analyses. Generally, the lowest energy conformer of the most active compound is selected as a template molecule and remaining compounds are aligned on the basis of the common structure. In this present works, molecule **40** was used as a

reference molecule due to its highest activity, and all of the remaining compounds were aligned on it to derive the subsequent 3D-QSAR analyses. The structure of the template molecule with common substructure and aligned molecules is demonstrated in Figures 1 and 2, respectively.



**Figure 1.** Structure of template compound (molecule 40). The four regions A, B, C, and D are depicted, and the common substructure is in bold.



**Figure 2.** Alignment of training and test set compounds on compound 40.

**2.3.2. CoMFA Procedure.** The CoMFA method was employed to construct the predictive 3D-QSAR model on the biases of aligned compounds. Performing CoMFA model results in calculation of steric (Lennard-Jones potentials) and electrostatic (Coulombic potentials) fields in 3D grid with a spacing of 2.0 Å, and extended at least 4.0 Å beyond the van der Waals around aligned compounds in all directions. An  $sp^3$  hybridized carbon was used as a probe atom to generate the steric and electrostatic fields with a charge of +1.0 and a van der Waals radius of 1.52 Å. The steric and electrostatic contributions were set as a default cutoff energy value of 30 kcal/mol so as to decrease the electrostatic energies and set the domination of large steric to a minimum. Column filtering was set to 1.0 kcal/mol to enhance the efficiency of the predictive models and

decrease the noise so that the columns possessing less energy variance in contrast to this value were excluded from analysis. CoMFA region focusing, which is a method of application of weights to the lattice points in a CoMFA region, is applied to improve the contribution of points for subsequent analyses. To derive appropriate results, the “St Dev coefficients” values as different weighing factors in addition to the grid spacing were performed.

**2.3.3. CoMSIA Procedure.** The CoMSIA method was performed with the same lattice box used in CoMFA analysis so as to evaluate the similarity indices between the compounds. In addition to steric and electrostatic fields, the CoMSIA model could result in some extra information including hydrophobic and hydrogen-bond donor and acceptor descriptors. The CoMSIA descriptors were obtained using the settings of an  $sp^3$  carbon atom as a probe atom<sup>41</sup> with a +1 charge, +1 hydrophobicity, +1 hydrogen bond donor, and +1 hydrogen bond acceptor at each lattice and grid spacing 2.0 Å.<sup>49</sup> Among the CoMSIA descriptors, the steric indices are related to the third power of atomic radii, the electrostatic descriptors are comprised from the atomic partial charges, hydrophobic fields are calculated from atom-based parameters,<sup>50</sup> and hydrogen bond donor and acceptor indices are calculated by the rule-based method derived from experimental data.<sup>49</sup> The similarity indices between the molecules and the probe atom were calculated using a Gaussian function. Column filtering was set to 1.0 kcal/mol in order to improve the predictive ability of the model.

**2.3.4. Partial Least-Squares Analysis and Validations.** In this present work, a partial least-squares (PLS) method,<sup>51</sup> which is an extension of multiple regression analysis, was used to calculate the minimal set of grid points and then linearly correlate the CoMFA and CoMSIA fields to the  $pIC_{50}$  values in order to generate the 3D-QSAR models.<sup>52</sup> To assess the predictive ability of the constructed PLS model, cross-validation analysis was employed using the leave-one-out (LOO) method<sup>51,52</sup> in which one of the molecules was excluded from the data set and its activity was predicted by the model derived from the rest of the data set. This procedure is continuing until every molecule available in the data set has been excluded once, and then, its biological activity was predicted by the constructed model. The result of this procedure is defined as cross-validation correlation coefficient ( $q^2$ ) and is a good statistical parameter to show the predictive capability of the model. The column filtering value ( $\sigma$ ) was set to 1.0 kcal/mol so as to enhance the efficiency of the built 3D-QSAR model and minimize the noise (here noise is the parts of the molecule which are not relevant for the biological activity).<sup>53</sup> PLS was conjunct with the cross-validation option to calculate the optimum number of components with the lowest standard error of predictions (SEP). This cross-validated correlation coefficient leads to estimate of the variability of the parameters in a final model. The final models are obtained based on the optimal number of compounds without employing cross-validation analysis. The cross-validated correlation coefficient ( $q^2$ ) and the squared correlation coefficient ( $r^2$ ) were utilized to investigate the robustness and statistical validity of the derived models. In addition, the stability and strength of the constructed models were evaluated further using bootstrapping and the leave-group-out cross validation procedures. For leave-group-out, 20% of the data points are removed from the data set, and the model was refitted; the predicted values for those points were then compared with the experimental values. Again, this is repeated until each data point has been omitted once. The leave-group-out cross validation value was validated using the average value for

**Table 2.** Correlation Coefficient of Selected Descriptors and Corresponding VIF and MF Values Based on SW-MLR

	BELv1	GATS6v	GATS1e	RDF090p	MF <sup>a</sup>	VIF <sup>b</sup>
BELv1	1	0	0	0	0.7161	1.903
GATS6v	-0.255	1	0	0	0.1613	1.143
GATS1e	-0.362	0.302	1	0	0.1044	1.347
RDF090p	0.686	-0.275	-0.428	1	0.0182	1.945

<sup>a</sup>Mean effect. <sup>b</sup>Variation inflation factor.

100 runs from each cross-validation. By the bootstrap technique, validation is performed by randomly generating training sets with sample repetitions, and then evaluating the predicted responses of the samples not included in the training set. The bootstrapping was repeated 5000 times for each validated model. Based on the calculated statistical results, the models presented lowest standard error of predictions (SEP), higher squared correlation coefficient ( $r^2$ ), and F values were selected. The external predictive ability of the built model was evaluated by some series of compounds as a test set (9 compounds) so that these compounds were excluded from the data scale in building the predictive models, and then, their activities were determined by constructed models. Therefore, the predictive correlation coefficient ( $r_{\text{pred}}^2$ ) based on the compounds included in test set was calculated. To validate the outcome of generated model further, the concordance correlation coefficients for each model in both sets were calculated. The concordance correlation coefficient was first introduced by Lin,<sup>54</sup> and this parameter evaluates the degree to which pairs of observations fall on the 45° line through the origin. The calculation of concordance correlation coefficient was performed using Medcalc program (<http://www.medcalc.org>).

### 3. RESULTS AND DISCUSSION

**3.1. 2D-QSAR Analysis.** **3.1.1. Stepwise Multiple Linear Regression Analysis.** Performing the stepwise technique resulted in selection of four descriptors, and then the 2D-QSAR model was developed based on the 38 compounds as a training set by stepwise multiple linear regression analysis. The external ability of the constructed model was then evaluated by using 10 compounds as a test set. The variation inflation factor (VIF) and mean effect (MF)<sup>55</sup> for each descriptor along with the correlation between the selected descriptors were calculated and listed in Table 2. The obtained 2D-QSAR model was given as follows:

$$\begin{aligned} \text{pIC}_{50} = & 48.42(\pm 9.803) - 14.90(\pm 4.947)\text{BELv1} \\ & - 3.254(\pm 0.7455)\text{GATS6v} - 3.345(\pm 0.3103) \\ & \text{GATS1e} - 0.1035(\pm 0.03131)\text{RDF090p} \end{aligned} \quad (1)$$

$$\begin{aligned} N_{\text{train}} &= 38, \quad R_{\text{train}}^2 = 0.833, \quad R_{\text{test}}^2 = 0.773, \\ R_{\text{adj}}^2 &= 0.813, \quad F_{\text{train}} = 41.12, \quad F_{\text{test}} = 3.974, \\ Q_{\text{LOO}}^2 &= 0.758, \quad Q_{\text{LGO}}^2 = 0.539, \quad Q_{\text{BOOT}}^2 = 0.736 \end{aligned}$$

where  $N$  is the number of compounds included in the training set,  $Q_{\text{LOO}}^2$ ,  $Q_{\text{LGO}}^2$ , and  $Q_{\text{BOOT}}^2$  are the squared cross-validation coefficients for leave-one-out, leave-group-out, and bootstrapping, respectively. The obtained high values for  $Q_{\text{LOO}}^2$ ,  $Q_{\text{LGO}}^2$ , and  $Q_{\text{BOOT}}^2$  indicate that the model is well-established without any overfitting and overestimating effect. The  $Q_{\text{BOOT}}^2$  was also a good indication robustness of the model.  $Q_{\text{BOOT}}^2$  was based on bootstrapping repeated 5000 times.  $R^2$  is the squared correlation

coefficient,  $R_{\text{adj}}^2$  is adjusted  $R^2$ , and  $F$  is the Fisher F statistic. The calculated  $R_{\text{test}}^2$  value (0.773) indicates the reliability and external predictive ability of the built model. The statistical parameters derived by SW-MLR demonstrate the lower root-mean-square error (RMSE) values (RMSE train = 0.331 and RMSE test = 0.357), and the higher value of  $Q_{\text{LOO}}^2$  and  $F$ , which confirms the reliability of the built model. Moreover, a Y-randomization test<sup>56</sup> was used to evaluate the robustness and the predictive ability of the proposed model. In this procedure, the  $\text{pIC}_{50}$  values were shuffled randomly and a new model was constructed. To accept the reliability of the model, the obtained new  $R^2$  and  $Q_{\text{LOO}}^2$  values for new proposed models should be less than that of obtained for the main initial model. The results of the Y-randomization test were summarized in Table 3. As can be seen, the all new  $R^2$  and

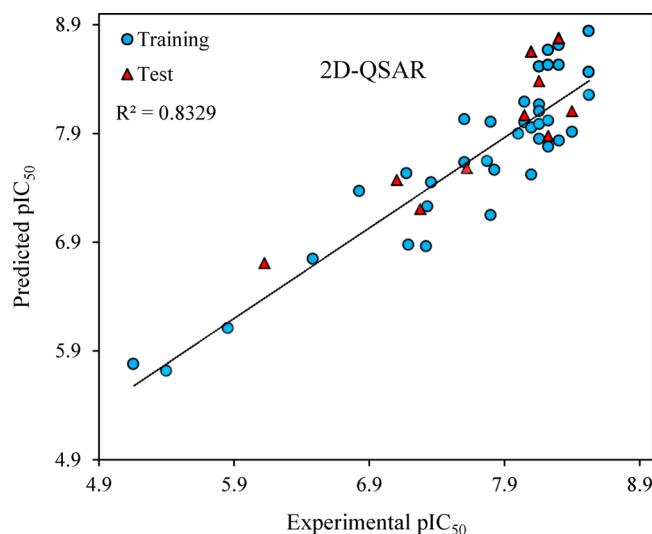
**Table 3.**  $Q_{\text{LOO}}^2$  and  $R_{\text{train}}^2$  Values after Several Y-Randomization Tests

no.	$Q^2$	$R^2$
1	0.1414	0.0254
2	0.0213	0.1700
3	0.0027	0.0816
4	0.0005	0.0803
5	0.0131	0.0731
6	0.1230	0.0255
7	0.1360	0.0133
8	0.0499	0.2280
9	0.0680	0.0556
10	0.2150	0.0112

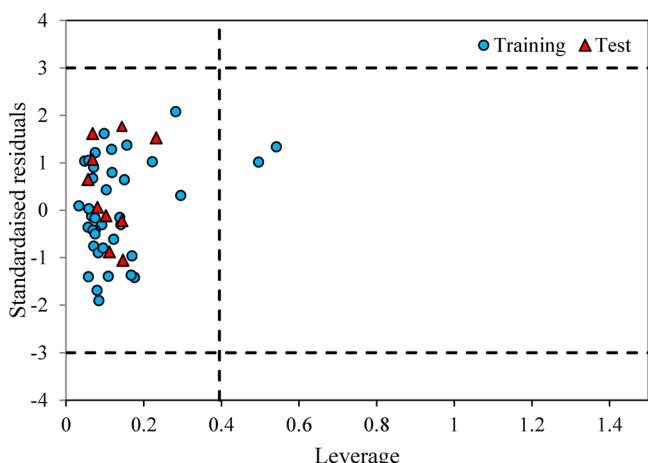
$Q_{\text{LOO}}^2$  values were less than that of obtained for the main model. The predicted values for inhibition activity of Hepatitis C Virus NSSB polymerase inhibitors by SW-MLR model are listed in Table 1. The plot of predicted  $\text{pIC}_{50}$  values versus experimental  $\text{pIC}_{50}$  values is presented in Figure 3.

**3.1.2. Applicability Domain.** The whole data set was investigated for the possible presence of outliers using the William plot (Figure 4).<sup>57</sup> This plot demonstrates the standardized residuals versus the leverage. The high leverage ( $h$ ) indicates the warning leverage ( $h^*$ ), where the presence of compounds in this area suggesting that the compounds were very influential on the predicted model.<sup>57</sup> The cutoff line (high leverage value) for this study was calculated as  $h^* = 0.395$ . In addition, a value of  $\pm 3\delta$  for standardized residual is widely used value as a cut off value for accepting the predicted value, since the points lie into the mean cover of 99% of normally distributed data. As it can be seen from Figure 4, two compounds (m3 and m8) were out of range of warning leverage; however, they were not considered as outliers due to the standard residuals less than  $\pm 3\delta$ . Therefore, the built model is reliable and can be used to predict the activity of new Hepatitis C Virus NSSB polymerase inhibitors.

**3.2. 3D-QSAR Analysis.** **3.2.1. CoMFA Analysis.** The CoMFA model was developed based on the set of 39



**Figure 3.** Plot of predicted pIC<sub>50</sub> against the experimental pIC<sub>50</sub> by SW-MLR.



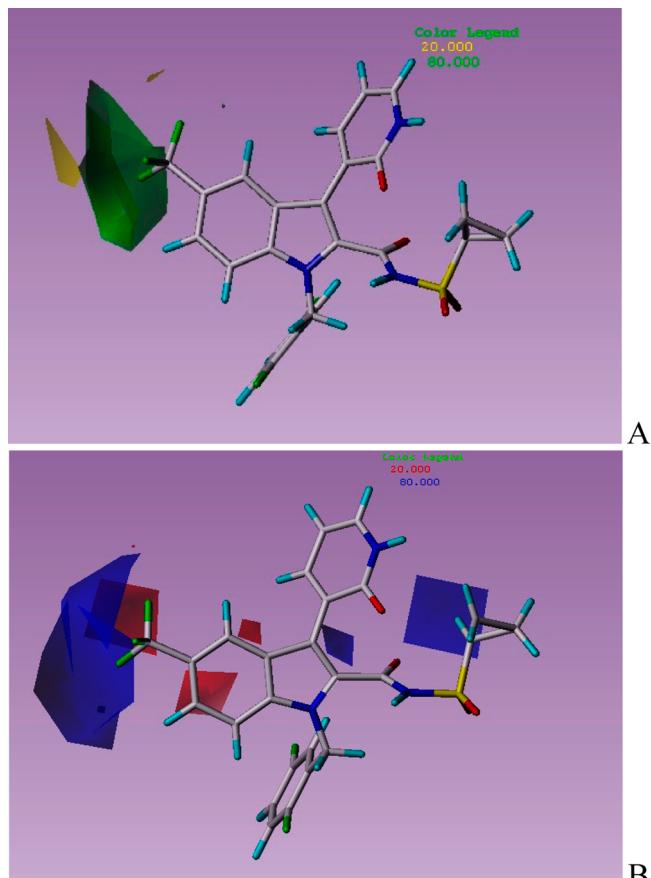
**Figure 4.** Williams plot of GA-MLR model for the training and test sets.

compounds, and then, an external test set including 9 compounds was employed to evaluate the reliability and applicability of the built model. PLS method was used to linearly correlate the chemical structures of compounds to inhibitory activities. Statistical parameters derived for CoMFA model were summarized in Table 4. The derived CoMFA-1 model illustrates a cross-validated  $q^2$  of 0.600 by four components. The optimal number of components was calculated by selecting the highest  $q^2$  value. The non-cross-validated PLS analysis with the four components resulted in a conventional  $r^2$  of 0.871,  $F = 57.203$ , and a standard error of estimation value (SEE) of 0.203. According to Table 4, the contributions of steric and electrostatic fields calculated by the CoMFA-1 model are 50.6% and 49.4% of variance, respectively. The steric and electrostatic fields obtained by the CoMFA-1 model are shown in Figure 5. The obtained high  $F$ , bootstrapping,  $q^2$ , and  $r^2$  values along with the lower standard error of estimation value (SEE) indicate the satisfactory predictive ability of the derived model (Table 4). The predicted pIC<sub>50</sub> values by CoMFA-1 model are listed in Table 1. Figure 6 demonstrates the correlation between experimental and predicted pIC<sub>50</sub> values by CoMFA-1 model. The calculated concordance correlation coefficient (CCC test = 0.7005, CCC training = 0.9308) indicate that the built CoMFA-1 model is

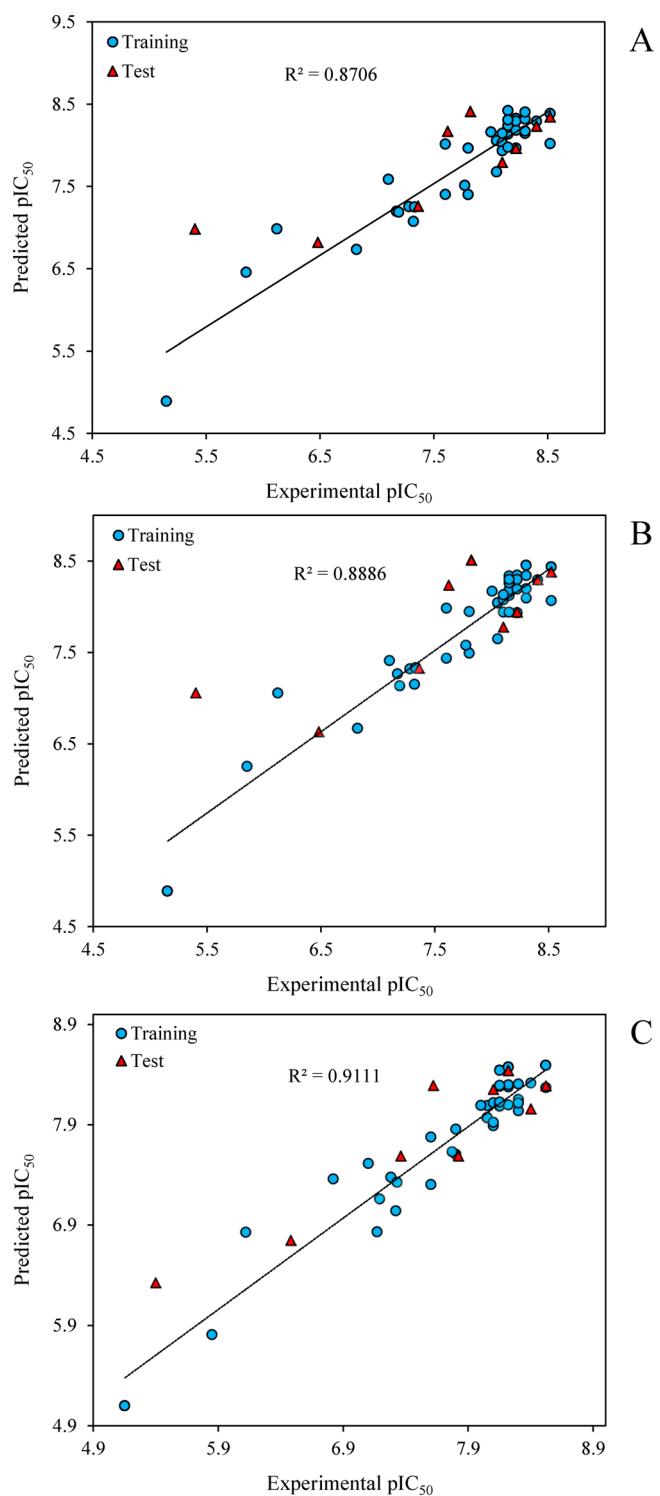
**Table 4. Statistical Results of CoMFA and CoMSIA Models**

	CoMFA-1	CoMFA-2 (after region focusing)	CoMSIA (best model)
PLS Statistics			
LOO cross $q^2$ /SEP <sup>a</sup>	0.600/0.497	0.691/0.436	0.664/0.455
group cross $q^2$ /SEP	0.604/0.494	0.717/0.418	0.678/0.446
nonvalidated $r^2$ /SEE <sup>b</sup>	0.871/0.282	0.889/0.262	0.911/0.234
$F$	57.203	67.832	87.130
$r_{\text{bootstrap}}^2$	0.910 $\pm$ 0.033	0.939 $\pm$ 0.034	0.916 $\pm$ 0.022
$S_{\text{bootstrap}}$	0.217 $\pm$ 0.133	0.179 $\pm$ 0.106	0.216 $\pm$ 0.126
CCC <sup>c</sup> for training set	0.9308	0.9410	0.9535
optimal compounds	4	4	4
$r_{\text{test}}^2$	0.689	0.626	0.867
CCC <sup>c</sup> for test set	0.7005	0.6875	0.8680
Field Distribution Percent			
steric	50.6	55.5	10.1
electrostatic	49.4	44.5	51.8
hydrophobic			38.1
H-bond donor			
H-bond acceptor			

<sup>a</sup>Standard error of prediction. <sup>b</sup>Standard errors of estimate.  
<sup>c</sup>Concordance correlation coefficient.



**Figure 5.** Contour maps of CoMFA: (A) steric field based on compound 40, (B) electrostatic field based on compound 40.

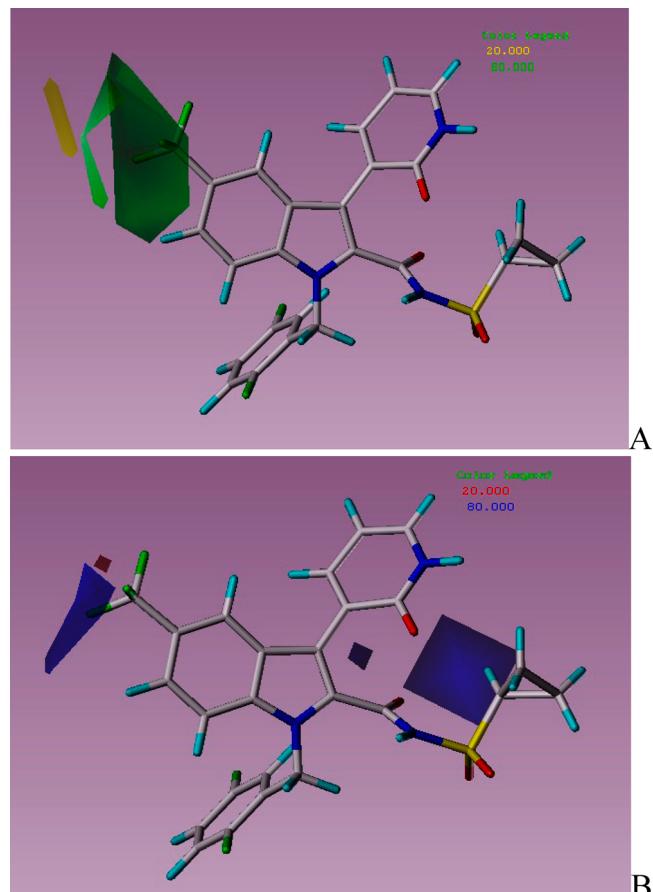


**Figure 6.** Plot of experimental against predicted pIC<sub>50</sub> by (A) CoMFA-1, (B) CoMFA-2, and (C) CoMSIA.

acceptable and can be used to predict the activity of Hepatitis C Virus NS5B polymerase inhibitors.

The constructed CoMFA-1 model was further developed to derive CoMFA region focusing. This method is based on the weights to the lattice points in a CoMFA region and can be used to improve the contribution of points for the analyses and then provide appropriate results with higher  $q^2$  values. Performing the region focusing represents a new model (CoMFA-2) with satisfactory statistical results. According to Table 4, the  $q^2$  value

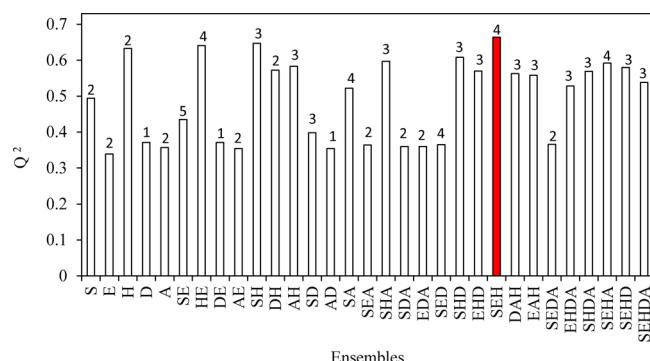
was increased from 0.600 to 0.691, and the  $r^2$  value (0.889) did not show a significant changes. Therefore, the obtained value indicates the suitable statistical outcomes for the same number of components. As it can be seen from Table 4, the new constructed model (CoMFA-2) indicates the appropriate results for bootstrapping (0.939),  $F$  (67.832) and SEE (0.262). The contributions of steric and electrostatic fields derived by CoMFA-2 model are 55.5% and 44.5% of variance, respectively. The steric and electrostatic fields after applying the region focusing (CoMFA-2) are shown in Figure 7. The predicted pIC<sub>50</sub>



**Figure 7.** Contour maps of CoMFA-2: (A) steric field based on compound 40., (B) electrostatic field based on compound 40.

values calculated by new constructed CoMFA-2 model are given in Table 1. The correlation between predicted versus experimental pIC<sub>50</sub> values obtained by CoMFA-2 model was demonstrated in Figure 6B. The calculated concordance correlation coefficient (CCC test = 0.6875, CCC training = 0.9410) indicate that the built CoMFA-2 model has medium strength for predicting of test set compounds, however the obtained CoMFA models, due to similar descriptors, could not present more appropriate results for test sets than the CoMSIA model.

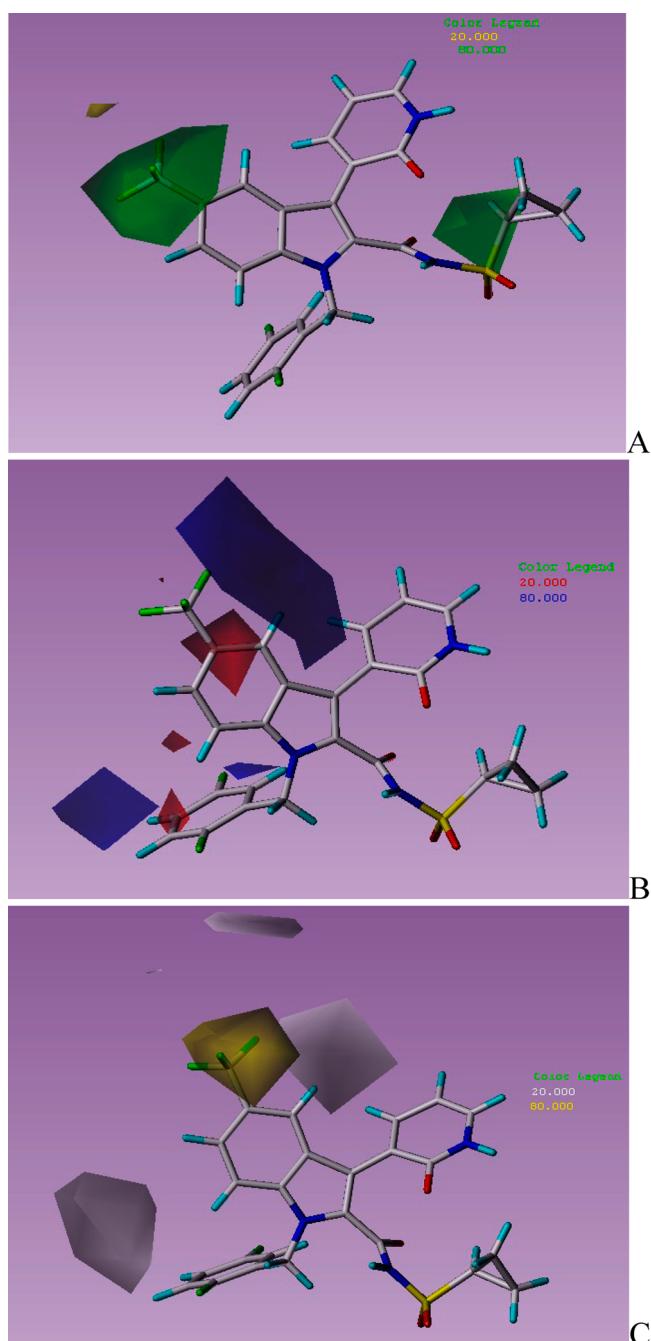
**3.2.2. CoMSIA Analysis.** In addition to steric and electrostatic fields calculated by the CoMFA approach, hydrophobic and hydrogen-bond donor and acceptor fields were derived using the CoMSIA method. In the CoMSIA model, each field is combined with others to determine their effects. All of the 31 possible combinations were calculated using their relevant  $q^2$  values and are shown in Figure 8. As can be seen from Figure 8, the higher  $q^2$  value (0.664) can be calculated where the steric, electrostatic, and



**Figure 8.** Graph of the 31 possible CoMSIA descriptor combinations (S = steric, E = electrostatic, H = hydrophobic, D/A = H-bond donor/acceptor) with their respective  $q^2$  values.

hydrophobic fields are presented. The best CoMSIA model among these 31 combinations was selected, and the statistical results are shown in Table 4. According to Table 4, the constructed CoMSIA model could provide high statistical values for  $q^2$  (0.664) and  $r^2$  (0.911) values with the four optimum components. The statistical parameters calculated by CoMSIA model indicates the satisfactory bootstrap of 0.916, group cross  $q^2$  value of 0.678 and  $F = 87.130$  with the less standard error of estimation (SEE) value of 0.234. Comparing the obtained statistical results shows that all constructed models could provide the appropriate results; however, the results for CoMSIA model ( $r_{\text{test}}^2 = 0.867$ ,  $r^2 = 0.911$ ,  $r_{\text{bootstrap}}^2 = 0.916$ , and  $F = 87.130$ ) are higher than those ones obtained by CoMFA-1 ( $r_{\text{test}}^2 = 0.689$ ,  $r^2 = 0.871$ ,  $r_{\text{bootstrap}}^2 = 0.910$ , and  $F = 57.203$ ) and CoMFA-2 ( $r_{\text{test}}^2 = 0.626$ ,  $r^2 = 0.889$ ,  $r_{\text{bootstrap}}^2 = 0.939$ , and  $F = 67.832$ ) models. The highest predicted  $r^2$  value for the test set and lowest SEE value with the highest  $q^2$  value calculated for CoMSIA model among the derived models indicates the superiority of CoMSIA model over the other constructed models. The model provided contributions of steric, electrostatic, and hydrophobic fields of 10.1%, 51.8%, and 38.1%, respectively. As can be seen, the electrostatic field represented by the CoMSIA model has higher contribution to the biological activity. Figure 9 demonstrates the steric, electrostatic, and hydrophobic fields derived by the CoMSIA model. The predicted  $\text{pIC}_{50}$  values are listed in Table 1, and the correlation between predicted and observed  $\text{pIC}_{50}$  values are shown in Figure 6C. The calculated concordance correlation coefficient (CCC test = 0.8680, CCC training = 0.9535) indicate that the built model is well-established.

**3.2.3. CoMFA-1 Contour Maps.** The steric and electrostatic contour maps resulted by CoMFA-1 model based on the reference molecule (molecule **40**) are shown in Figure 5. In this figure, the green contours indicate the regions where presence of bulkier groups (80% contribution) would contribute to increase of biological activity while the yellow contours (20% contribution) show the regions where such bulkier groups result in decrease of biological activities. Figure 5 demonstrates the steric field effects. As can be seen, Figure 5 introduced some regions where using bulky groups would be beneficial to increasing biological activity and also displayed some regions where the presence of bulky groups does not contribute to improvement of biological activities. The CoMFA electrostatic contour maps (Figure 5B) are represented by red and blue contours. The red contours (20% contribution) demonstrate the regions where electron-rich (electronegative) groups result in increase of biological activity whereas the blue contours (80%



**Figure 9.** Contour maps of CoMSIA: (A) steric field based on compound **40**, (B) electrostatic field based on compound **40**, (C) hydrophobic field based on compound **40**.

contribution) indicate the regions where presence of positive charge (or decreased negative charge) group would result in an increase of biological activity. To aid the visualizations, the selected template molecule was divided into four regions as shown in Figure 1.

The steric contours of the CoMFA-1 display a large green contour near positions 2 and 3 of region D ( $R_1$  and  $R_2$  substituents in the series of compounds **1–31** and  $R_1$  and  $R_3$  substituents in the series of compounds **32–48**). This green contours in region D at position 2 shows the favorable effect of bulky groups in increasing the biological activities of molecules. According to Table 1, this can be explained by comparing the activities of molecules **11** and **12** where using bulk groups

( $-\text{OCF}_3 > -\text{CF}_3$ ) at  $\text{R}_2$  substituent would result in higher  $\text{pIC}_{50}$  values ( $7.80 > 7.77$ ), respectively. This is also observed between molecules **1** and **5** where compound **5** possessed Br group and displayed higher inhibitory activity ( $\text{pIC}_{50} = 7.36$ ) and bulky characteristics than compound **1** which contained Cl substituents in region D at the  $\text{R}_2$  substituent ( $\text{pIC}_{50} = 7.28$ ). As can be seen from Table 1, using the bulky substituents ( $-\text{CH}_2\text{CH}_3 > -\text{CH}_3$ ) in position 2 of region D would lead to increase of  $\text{pIC}_{50}$  values in the molecules: **27** (having  $-\text{CH}_2\text{CH}_3$  group with  $\text{pIC}_{50} = 8.40$ ) and **26** (having  $-\text{CH}_3$  group with  $\text{pIC}_{50} = 8.10$ ). This effect can also be observed between molecules **33** and **37** where their higher activities can be related to the presence of  $-\text{CH}_3$  and  $-\text{CH}_2\text{CH}_3$  substituents, namely in which by increasing the bulky behavior of substituent ( $-\text{CH}_3 < -\text{CH}_2\text{CH}_3$ ) the inhibitory activities would be increased ( $8.00 < 8.22$ ). The green contours near region D in the  $\text{R}_1$  substituents can be explained by comparing molecules **45** (having  $-\text{Cl}$  group at the  $\text{R}_1$  substituent in region D) and **35** (having  $-\text{H}$  group at the  $\text{R}_1$  substituent in region D) where using a bulky group influenced the outcome of  $\text{pIC}_{50}$  values ( $8.30 > 8.22$ ). As it can be seen from Figure 5, the steric contours map shows a small yellow contour at the  $\text{R}_1$  substituent in the 2-position of region D. The presence of yellow contours near region D illustrates the unfavorable effect of bulky groups in some series of compounds. As shown by the  $\text{pIC}_{50}$  values, using bulky groups such as  $-\text{Cl}$  (molecule **2** with  $\text{pIC}_{50} = 6.12$ ) in contrast to  $-\text{H}$  (molecule **1** with  $\text{pIC}_{50} = 7.28$ ) in the same series at the 3-position of region D ( $-\text{Cl} > -\text{H}$ ) would result in decrease of activities. These results are in line with those derived from the developed 2D-QSAR model where the lowest eigenvalue  $n$ . One of Burden matrix/weighted by atomic van der Waals volumes (BELv1) displayed the positive mean effect. This descriptor is associated with van der Waals volumes characteristic, and since its MF value was positive, increasing its value would lead to increase of inhibitory activities. If we consider the certain cores in the data set, it can be seen that the van der Waals volumes feature in compound **27** (having  $-\text{CH}_2\text{CH}_3$  group with) increased if compare to the compound **26** (having  $-\text{CH}_3$  group) and subsequently lead to increase of  $\text{pIC}_{50}$  ( $8.40 > 8.10$ ), respectively.

Figure 5B indicates that increased positive charge (or decrease negative charge) would lead to enhance the inhibitory activity in blue regions (80% contribution) while the presence of negative charge (or decrease positive charge) is favored and can be resulted in higher activity in red regions (20% contribution). As can be seen from Figure 5B, there are two medium blue contours near region A and D and a small blue contour in the bottom of region B. The small blue contour at the bottom of region B can be explained by comparing molecules **32** and **26** where replacing a group containing nitrogen in place of a group containing oxygen would increase the positive charge behavior of this substituent, and consequently, the  $\text{pIC}_{50}$  values would be increased ( $8.15 > 8.10$ ). This blue contour map clearly explains that why compounds in the series of **32–48** showed higher inhibitory activities than those in the series of **1–31**. The medium blue contours at the bottom of  $\text{R}_1$  and  $\text{R}_2$  substituent in region D can be understood by comparing molecules **22** (having  $-\text{OCF}_3$  group with  $\text{pIC}_{50} = 7.62$ ) and **23** (having  $-\text{OCH}_3$  group with  $\text{pIC}_{50} = 7.82$ ) where using the negative charge group at this position would lead to decreasing the inhibitory activities. This fact is also observed between molecules **8**, **9**, and **7** where increasing the negative charge properties ( $-\text{SO}_2\text{CH}_3 > -\text{COCH}_3 >$  cyclopropyl) would result in decrease of inhibitory activities ( $5.85 < 6.48 < 7.19$ ), respectively. The blue contours

observed near region A can be described considering compounds **38** and **37**; whereas, positive charge features increased (*iso*-propyl  $> -\text{CH}_2\text{CH}_3$ ), the inhibitory activities ( $8.30 > 8.22$ ) are also increased. According to Table 1, the presence of the blue contours in this position can be explained further by considering molecules **35** (having cyclopropyl group with  $\text{pIC}_{50} = 8.22$ ) and **33** (having  $-\text{CH}_2\text{CH}_3$  group with  $\text{pIC}_{50} = 8.00$ ) where using the positive charge group at this position would lead to increase the  $\text{pIC}_{50}$  values. Figure 5B demonstrated two medium red contours near the 3-position of region D and the 2-position of region C and small red contours near the 5-position of substituent  $\text{R}_4$  in region C. The medium red contours near the 3-position of region D can be explained by comparing molecules **46** and **44** where using the negative charge groups ( $-\text{CF}_3 > \text{F}$ ) would result in higher biological activities ( $8.15 > 8.10$ ), respectively. The two red contours near the 2- and 5-position of the  $\text{R}_4$  substituent in region C can be understood by comparing molecules **20**, **21**, and **16** where the presence of negative charge groups ( $2,\text{S}-\text{F} > -\text{CONH}_2 > 2-\text{F}$ ) would result in higher biological activities ( $8.40 > 8.15 > 8.05$ ), namely. These results are in line with results of the developed 2D-QSAR model. The second and third descriptors provided by the 2D-QSAR model are GATSSv and GATS1e which belong to 2D-autocorrelation descriptors. In these classes of descriptors, the Geary coefficient is a distance type function, and function is any physicochemical property calculated for each atom of the compound, such as atomic mass, polarizability, electronegativity, and atomic van der Waals volumes. Hence, the atoms of molecules represent the set of discrete points in space, and the atomic property, the function evaluated at those points. These two descriptors displayed positive mean effect for their contribution to the biological activities. The effect of van der Waals volumes were discussed above, and therefore, the effects of GATS1e which is weighted by electronegativity were explained. The mean effect value for this descriptor is less than those obtained for BELv1 and GATS6v; therefore, it indicates that some special compounds in some exclusive cores follow this rule in which increase of electronegativity would increase the biological activity. As it was described, the electronegativity in molecule **20** ( $2,\text{S}-\text{F}$ ) is increased if compared to molecule **16** ( $2-\text{F}$ ), and it resulted in higher biological activities ( $8.40 > 8.05$ ).

**3.2.4. CoMFA-2 Contour Maps.** The steric and electrostatic contour maps were shown by CoMFA-2 model (after region focusing) based on the template molecule (molecule **40**) in Figure 7. The concept of contributions of each field is the same as the details illustrated in the CoMFA-1 contours map.

The steric contours of the CoMFA (after region focusing) displays (Figure 7A) a large green contour near the 2- and 3-positions in region D. According to Table 1, this can be explained by comparing the order of activities in molecules **45** and **44** where compound **45** possessed a Cl substituent in contrast to compound **44** (having Fluorine substituent at the same position) and showed higher inhibitory activity ( $8.30 > 8.10$ ), respectively. This is also observed between molecules **39** and **35** where compound **39** possessed an ethyl group and displayed higher inhibitory activity ( $\text{pIC}_{50} = 8.52$ ) than compound **35** which contained a methyl substituent in region D ( $\text{pIC}_{50} = 8.22$ ). As it can be seen from Figure 7, the steric contour maps show small yellow contours near region D at the  $\text{R}_1$  substituent. The presence of yellow contours near region D beside the presence of green contours indicates the unfavorable effect of bulky groups in some special series of compounds. This fact can be described by comparing molecules **3** (having a Br group in the 3-position of region D) with **5** (having a Br group in the 2-position of region

**Table 5.** Chemical Structures of Newly Designed HCV NS5B Polymerase Inhibitors with Respective Predicted Activities

N.o.	Substituents					R <sub>6</sub>	Predicted		
	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>		CoMFA-1	CoMFA-2	CoMSIA
<b>M40</b>	H	-CF <sub>3</sub>	H	H	H	<b>5-F</b>	<b>8.39</b>	<b>8.44</b>	<b>8.49</b>
N1	I	-CF <sub>3</sub>	-NH <sub>2</sub>	H	H	5-F	8.51	8.56	8.80
N2	H	-CF <sub>3</sub>	-NH <sub>2</sub>	H	H	5-F	8.49	8.50	8.79
N3	Br	-CF <sub>3</sub>	-NH <sub>2</sub>	H	H	5-F	8.52	8.55	8.78
N4	H	-CF <sub>3</sub>	-NH <sub>2</sub>	H	2-CH <sub>3</sub>	5-F	8.46	8.45	8.77
N5	Br	-CF <sub>3</sub>	-NH <sub>2</sub>	H	2-CH <sub>3</sub>	5-F	8.46	8.49	8.75
N6	Br	-CF <sub>3</sub>	-NH <sub>2</sub>	H	H	4-NH <sub>2</sub> , 5-F	8.51	8.54	8.74
N7	Cl	-CF <sub>3</sub>	-NH <sub>2</sub>	H	H	5-F	8.49	8.52	8.71
N8	Cl	-CF <sub>3</sub>	-NH <sub>2</sub>	H	2-CH <sub>3</sub>	5-F	8.43	8.45	8.69
N9	Br	-CF <sub>3</sub>	- CONH <sub>2</sub>	H	H	4-CONH <sub>2</sub> , 5-F	8.46	8.45	8.61
N10	Br	-CF <sub>3</sub>	-NH <sub>2</sub>	H	H	4-CONH <sub>2</sub> , 5-F	8.67	8.60	8.59
N11	Br	-CF <sub>3</sub>	-NH <sub>2</sub>	H	H	4-OH, 5-F	8.45	8.46	8.54
N12	Br	-CF <sub>3</sub>	-NH <sub>2</sub>	H	H	4-OH, 5-H	8.46	8.48	8.52
N13	Br	-CH <sub>2</sub> CF <sub>3</sub>	-NH <sub>2</sub>	-CH <sub>2</sub> CH <sub>3</sub>	H	4-CONH <sub>2</sub>	8.44	8.46	8.51

D) where replacing the bulkier group close to the R<sub>1</sub> substituent in this series would result in decreasing of the pIC<sub>50</sub> values (5.40 < 7.36).

Figure 7B demonstrates the electrostatic contours map. As it can be seen from Figure 7B, there are two medium blue contours near region D and A. The large blue contours in region A can be explained by comparing molecules 39 (cyclopropyl group with pIC<sub>50</sub> of 8.52), 38 (iso-propyl group with pIC<sub>50</sub> of 8.30), and 37 (ethyl group with pIC<sub>50</sub> of 8.22) where the positively charged effects are increasing and would result in an increase of inhibitory activity. This also can be seen in molecules 35 (cyclopropyl group with pIC<sub>50</sub> of 8.22), 34 (iso-propyl group with pIC<sub>50</sub> of 8.15), and 33 (ethyl group with pIC<sub>50</sub> of 8.00) where the positively charged groups led to an increase of inhibitory activities. The electrostatic contour of CoMFA (after region focusing) (Figure 5B) shows the blue contour enclosing the 3-position of region D. This can be understood by comparing molecules 2 and 1 where using the negative charged group (Cl > H) in special positions of this core would result in decrease of inhibitory activities (6.12 < 7.28), respectively.

**3.2.5. CoMSIA Contour Maps.** Figure 9 displays the steric, electrostatic, and hydrophobic fields derived by the CoMSIA model. As shown in Figure 9, the steric field displayed a large green contour in the bottom of region D (2- and 3-position of region D) and a medium green contour in region A. The medium green contour located at region A can be understood by comparing the molecules 39 (having cyclopropyl group with pIC<sub>50</sub> = 8.52) and 38 (having iso-propyl group with pIC<sub>50</sub> = 8.30) where the presence of bulkier size group would result in increase of pIC<sub>50</sub> values. As can be seen from Table 1, this effect can be explained further by considering compounds 35 (having cyclopropyl group with pIC<sub>50</sub> = 8.22) and 33 (having ethyl group with pIC<sub>50</sub> = 8.00) where cyclopropyl group has bulkier size in contrast to ethyl group and, therefore, showed higher inhibitory activity. The large green contours in the bottom of

region D can be explained by comparing molecules 26 (having -CH<sub>3</sub> group with pIC<sub>50</sub> = 8.10) and 27 (having -CH<sub>2</sub>CH<sub>3</sub> group with pIC<sub>50</sub> = 8.40) where the bulky groups in this position would lead to better inhibitory activities.

Figure 9B shows the electrostatic contours map. As can be seen, the electrostatic field displayed a large blue contour in the bottom of region D. This can be understood by comparing the molecules 20 (having -CF<sub>3</sub> group with pIC<sub>50</sub> = 8.40) and 22 (having -OCF<sub>3</sub> group with pIC<sub>50</sub> = 7.62) where the presence of more negative charged behavior group in this position would result in decrease of pIC<sub>50</sub> values, respectively. According to Table 1, this effect can further be explained considering molecules 29 (having -CH<sub>2</sub>CF<sub>3</sub> group with pIC<sub>50</sub> = 8.10) and 27 (having -CH<sub>2</sub>CH<sub>3</sub> group with pIC<sub>50</sub> = 8.40) where using additional fluorine group would increase the negative charge characteristic of molecule 29 and, consequently, decrease its inhibitory activity. Figure 9B shows another blue contour enclosing region C which can be described by comparing compounds 17 and 18 where adding a positively charged group (2,4-CH<sub>3</sub> > 2,4-F) led to increase the inhibitory activities (7.80 > 7.60). The presence of these contours can be further explained by comparing compounds 14 and 13 where increasing the level of negatively charged atom (4-OCH<sub>2</sub> > 4-OCH<sub>3</sub>) would result in decrease of the inhibitory activity (7.17 < 7.60), respectively. As it can be seen from Figure 9B, the CoMSIA model demonstrated medium and small red contours near the 2- and 5-positions of region C. This can be described by considering molecules 20 and 16 where replacing the hydrogen with a negatively charged group (-F) in the 5-position of substituent R<sub>4</sub> in region C would result in increase of negative charge behavior of substituent (F > H) and therefore an increase in pIC<sub>50</sub> values (8.40 > 8.05). Additionally, the obtained results are in accordance with the derived result of 2D-QSAR model. Referring to the last selected descriptor (RDF090p, radial distribution function 090/weighted by polarizability), the effect of polarizability of groups can be

better understood. The polarizability was increased in molecules in the series of 32–48 by using  $-\text{NH}-\text{SO}_2-$  in place of OH which lead to higher inhibitory activity in this series when compared to compounds in the series of 1–31.

Figure 9C demonstrated the hydrophobic contour maps. In a hydrophobic field, yellow and white contours are highlighted areas where the yellow contours represents the presence of hydrophobic substituents and using them would be favorable for better activity (favored level 80%), while white contours indicate regions where hydrophobic groups decrease the biological activity (unfavored level 20%). Figure 9C displays the large yellow contours near the 3-position in region A, and several white contours near region A and region C. The large white contours enclosing region C can be explained by comparing molecules 17 and 15 where increasing the hydrophobicity by adding more alkyl groups would result in decreasing of biological activities ( $7.80 < 8.05$ ). The medium white contour map near region D can be described by comparing molecules 23 and 24 where increasing the hydrophobic characteristics of group ( $-\text{OCH}_2\text{CH}_3 > -\text{OCH}_3$ ) would result in a decrease of the biological activities ( $7.10 < 7.82$ ), respectively. Considering the interaction of electrostatic and hydrophobic fields, it can be understood that substituents in 4-position of region C should be from a positively charged group and also demand hydrophilic effects.

**3.2.6. Some Insights for Design of New HCV NS5B Polymerase Inhibitors.** According to the information derived by contours maps generated by 3D-QSAR models, some important information about the chemical structures requirement were presented to investigate the effect of each kind of group as the substituent for region A, B, C, and D, on inhibitory activities. Based on this derived information, new potent compounds can be designed as HCV NSSB polymerase inhibitors. According to the analysis of derived contours map by each model, the better biological activities can be obtained considering the bulky group at region A and D. The obtained results explain that the positively charged groups with higher hydrophilic features in region C at the 4-position can lead to higher inhibitory activity. The better biological activities can also be achieved where the positively charged groups are used in region B. Taking the above results into consideration, some new compounds as potent HCV NSSB polymerase inhibitors were designed and listed in Table 5. To investigate the results of each substituent on activity results, CoMSIA results as the best modeling tool to use. The CoMSIA contours maps showed a large blue contour enclosing region D near the substituent ( $R_3$  in Table 5). The previous compounds did not indicate any functional groups for the  $R_3$  substituent in compounds 32–48; however the effect of different groups can be studied in this position. Therefore, the first attempt to design new compounds was to locate the positively charged groups (i.e.: donor groups) so as to increase the contribution of the positively charged groups in the molecular graph. In place of the H atom in this series of compounds, two donor groups were used and showed remarkable improvements from the previous most potent compound. The second effort to improve the results even more was to evaluate the effect of the  $R_1$  position (Table 5) in region D on inhibition activities. Since the contribution of electrostatic field is more than two other selected fields, the combination of effects should be considered for positioning new functional groups in this location. The new designed compounds showed that at the first place (while the donor groups are presented in the  $R_3$  position (Table 5)) the steric effects in the  $R_1$  position have positive effects, but employing the functional

groups with lower electronegativity is of interest. This can be understood by comparing compounds N7 and N3 where the Cl atom possessed higher electronegativity and showed lower inhibition activities. The next attempt was to discover and improve the effects of functional groups in region C where the hydrophobic and electrostatic effects are presented. It was observed that addition of donor groups ( $-\text{NH}_2$ ) in  $R_6$  at the fourth position (Table 5) can lead to better inhibition activities (N6 and M40 with  $\text{pIC}_{50}$  of 8.74 and 8.49, respectively). Among the designed compounds, N1 was presented the highest activities considering the CoMSIA model. To understand the origin of this increase in activity, compounds M40 and N1 can be compared. The dipole moment is a good tool to measure the intermolecular forces and molecular polarity that encodes the charge separation in molecule. Therefore, the larger the electronegativity of bonded atom is, the larger the dipole moment would be, and it results in generation of slight positive and negative charge in a bond. The presence of high electronegative compounds in region D at  $R_1$  would result in partial positive charge in the left side of region D where it is not of interest due to driven contours, and therefore, the compounds with larger electronegativity can lead to lower activity. Moreover, the steric effects enhance the activity at the same position, and therefore, it can be concluded that slight increase in dipole moment in bound  $R_1$  along with increasing the steric effect improves the biological activities. The variation in electron density in some moiety of compounds can benefit the inhibition activities if we consider the electrostatic fields. Based on the obtained results, the effects of increasing the steric field in region B was better. An ethyl group was replaced at the  $R_4$  position; however, it presented less appropriate results than other designed compounds.

#### 4. CONCLUSION

The present work was employed to perform the CoMFA and CoMSIA, 3D-QSAR models to investigate the effects of each substituent in each region for the series of compounds as HCV NSSB polymerase inhibitors. Additionally the 2D-QSAR was constructed with appropriate results ( $R_{\text{train}}^2 = 0.833$ ,  $R_{\text{test}}^2 = 0.773$ ,  $F_{\text{train}} = 41.12$ ,  $F_{\text{test}} = 3.974$ ,  $Q_{\text{LOO}}^2 = 0.758$ ) and compared with the results of derived 3D-QSAR models. The PLS analysis was carried out to linearly correlate the CoMFA and CoMSIA descriptors with inhibitory activities. The calculated statistical parameters (high  $q^2$ ,  $r_{\text{test}}^2$ ,  $F$ ,  $r^2$ , bootstrap, and less SEE values) illustrated the appropriate predictive ability of the constructed 3D-QSAR models. The external predictive capability of each developed model is further validated by using nine compounds as test set. The obtained information by CoMFA and CoMSIA were described the contribution of steric, electrostatic and hydrophobic fields based on the aligned molecules. Consequently, the presence of bulky group in region A and D can cause better inhibitory activities. In addition, strong positive charge groups with higher hydrophilic features in region C at the 4-position can result in higher biological activity. The positively charged group with higher bulky features in region B is favorable and can also lead to better  $\text{pIC}_{50}$  values. In this work, the proposed models could identify and provide better insight in some regions where further studies should have been done to understand the effects of different functional groups on inhibition activities so as to design new inhibitors with higher  $\text{pIC}_{50}$  values.

## AUTHOR INFORMATION

### Corresponding Author

\*E-mail: pourbasheer@ut.ac.ir. Tel.: +98-21-61114714. Fax: +98-21-88632976.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The authors would like to thank the State Scholarships Foundation of Greece (I.K.Y.) for financial support.

## REFERENCES

- (1) Rosenberg, S. Recent advances in the molecular biology of hepatitis C virus. *J. Mol. Biol.* **2001**, *313*, 451–464.
- (2) Tan, S.-L.; Pause, A.; Shi, Y.; Sonenberg, N. Hepatitis C therapeutics: current status and emerging strategies. *Nat. Rev. Drug Discovery* **2002**, *1*, 867–881.
- (3) Kronenberger, B.; Zeuzem, S. Antiviral Targets in HCV. In *Chronic Hepatitis C Virus*; Schiffman, M. L., Ed.; Springer: New York, 2012; Chapter 17, pp 203–225.
- (4) Das, D.; Hong, J.; Chen, S.-H.; Wang, G.; Beigelman, L.; Seiwert, S. D.; Buckman, B. O. Recent advances in drug discovery of benzothiadiazine and related analogs as HCV NS5B polymerase inhibitors. *Bioorg. Med. Chem. Lett.* **2011**, *19*, 4690–4703.
- (5) Appel, N.; Schaller, T.; Penin, F.; Bartenschlager, R. From Structure to Function: New Insights into Hepatitis C Virus RNA Replication. *J. Mol. Biol. Chem.* **2006**, *281*, 9833–9836.
- (6) Clercq, E. D. The design of drugs for HIV and HCV. *Nat. Rev. Drug Discovery* **2007**, *6*, 1001–1018.
- (7) Hong, Z.; Cameron, C. E.; Walker, M. P.; Castro, C.; Yao, N.; Lau, J. Y. N.; Zhong, W. A Novel Mechanism to Ensure Terminal Initiation by Hepatitis C Virus NS5B Polymerase. *Virology* **2001**, *285*, 6–11.
- (8) De Francesco, R.; Tomei, L.; Altamura, S.; Summa, V.; Migliaccio, G. Approaching a new era for hepatitis C virus therapy: inhibitors of the NS3–4A serine protease and the NS5B RNA-dependent RNA polymerase. *Antiviral Res.* **2003**, *58*, 1–16.
- (9) Yamashita, T.; Kaneko, S.; Shirota, Y.; Qin, W.; Nomura, T.; Kobayashi, K.; Murakami, S. RNA-dependent RNA Polymerase Activity of the Soluble Recombinant Hepatitis C Virus NS5B Protein Truncated at the C-terminal Region. *J. Biol. Chem.* **1998**, *273*, 15479–15486.
- (10) Moradpour, D.; Brass, V.; Bieck, E.; Fribe, P.; Gosert, R.; Blum, H. E.; Bartenschlager, R.; Penin, F.; Lohmann, V. Membrane Association of the RNA-Dependent RNA Polymerase Is Essential for Hepatitis C Virus RNA Replication. *J. Virol.* **2004**, *78*, 13278–13284.
- (11) Ago, H.; Adachi, T.; Yoshida, A.; Yamamoto, M.; Habuka, N.; Yatsunami, K.; Miyano, M. Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Structure* **1999**, *7*, 1417–1426.
- (12) Bressanelli, S.; Tomei, L.; Roussel, A.; Incitti, I.; Vitale, R. L.; Mathieu, M.; De Francesco, R.; Rey, F. A. Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 13034–13039.
- (13) Lesburg, C. A.; Cable, M. B.; Ferrari, E.; Hong, Z.; Mannarino, A. F.; Weber, P. C. Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. *Nat. Struct. Mol. Biol.* **1999**, *6*, 937–943.
- (14) Bressanelli, S.; Tomei, L.; Rey, F. A.; De Francesco, R. Structural Analysis of the Hepatitis C Virus RNA Polymerase in Complex with Ribonucleotides. *J. Virol.* **2002**, *76*, 3482–3492.
- (15) Ding, Y.; Girardet, J.-L.; Smith, K. L.; Larson, G.; Prigaro, B.; Lai, V. C. H.; Zhong, W.; Wu, J. Z. Parallel synthesis of pteridine derivatives as potent inhibitors for hepatitis C virus NS5B RNA-dependent RNA polymerase. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 675–678.
- (16) Beaulieu, P. L.; Bös, M.; Bousquet, Y.; DeRoy, P.; Fazal, G.; Gauthier, J.; Gillard, J.; Goulet, S.; McKercher, G.; Poupart, M.-A.; Valois, S.; Kukolj, G. Non-nucleoside inhibitors of the hepatitis C virus NS5B polymerase: discovery of benzimidazole 5-carboxylic amide derivatives with low-nanomolar potency. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 967–971.
- (17) Beaulieu, P. L.; Gillard, J.; Bykowski, D.; Brochu, C.; Dansereau, N.; Duceppe, J.-S.; Haché, B.; Jakalian, A.; Lagacé, L.; LaPlante, S.; McKercher, G.; Moreau, E.; Perreault, S.; Stammers, T.; Thauvette, L.; Warrington, J.; Kukolj, G. Improved replicon cellular activity of non-nucleoside allosteric inhibitors of HCV NS5B polymerase: From benzimidazole to indole scaffolds. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4987–4993.
- (18) Biswal, B. K.; Wang, M.; Cherney, M. M.; Chan, L.; Yannopoulos, C. G.; Bilimoria, D.; Bedard, J.; James, M. N. G. Non-nucleoside Inhibitors Binding to Hepatitis C Virus NS5B Polymerase Reveal a Novel Mechanism of Inhibition. *J. Mol. Biol.* **2006**, *361*, 33–45.
- (19) Gopalsamy, A.; Shi, M.; Ciszewski, G.; Park, K.; Ellingboe, J. W.; Orlowski, M.; Feld, B.; Howe, A. Y. M. Design and synthesis of 2,3,4,9-tetrahydro-1H-carbazole and 1,2,3,4-tetrahydro-cyclopenta[b]indole derivatives as non-nucleoside inhibitors of hepatitis C virus NS5B RNA-dependent RNA polymerase. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 2532–2534.
- (20) Ishida, T.; Suzuki, T.; Hirashima, S.; Mizutani, K.; Yoshida, A.; Ando, I.; Ikeda, S.; Adachi, T.; Hashimoto, H. Benzimidazole inhibitors of hepatitis C virus NS5B polymerase: Identification of 2-[4-diarylmethoxyphenyl]-benzimidazole. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1859–1863.
- (21) De Francesco, R.; Carfi, A. Advances in the development of new therapeutic agents targeting the NS3–4A serine protease or the NS5B RNA-dependent RNA polymerase of the hepatitis C virus. *Adv. Drug Delivery Rev.* **2007**, *59*, 1242–1262.
- (22) Bhatt, A.; Gurukumar, K. R.; Basu, A.; Patel, M. R.; Kaushik-Basu, N.; Talele, T. T. Synthesis and SAR optimization of diketo acid pharmacophore for HCV NS5B polymerase inhibition. *Eur. J. Med. Chem.* **2011**, *46*, 5138–5145.
- (23) Jackson, R. W.; LaPorte, M. G.; Herbertz, T.; Draper, T. L.; Gaboury, J. A.; Rippin, S. R.; Patel, R.; Chunduru, S. K.; Benetatos, C. A.; Young, D. C.; Burns, C. J.; Condon, S. M. The discovery and structure–activity relationships of pyrano[3,4-b]indole-based inhibitors of hepatitis C virus NS5B polymerase. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 3227–3231.
- (24) Stankiewicz-Drogoń, A.; Dörner, B.; Erker, T.; Boguszewska-Chachulska, A. M. Synthesis of New Acridone Derivatives, Inhibitors of NS3 Helicase, Which Efficiently and Specifically Inhibit Subgenomic HCV Replication. *J. Med. Chem.* **2010**, *53*, 3117–3126.
- (25) Velázquez, F.; Venkatraman, S.; Blackman, M.; Pinto, P.; Bogen, S. p.; Sannigrahi, M.; Chen, K.; Pichardo, J.; Hart, A.; Tong, X.; Girjavallabhan, V.; Njoroge, F. G. Design, Synthesis, and Evaluation of Oxygen-Containing Macrocyclic Peptidomimetics as Inhibitors of HCV NS3 Protease. *J. Med. Chem.* **2009**, *52*, 700–708.
- (26) Wang, Q. M.; Heinz, B. Recent advances in prevention and treatment of hepatitis C virus infections. In *Progress in Drug Research*; Jucker, E., Ed.; Birkhäuser Basel, 2000; Vol. 55, Chapter 1, pp 1–32.
- (27) Chen, K. X.; Vibulbhan, B.; Yang, W.; Sannigrahi, M.; Velazquez, F.; Chan, T.-Y.; Venkatraman, S.; Anilkumar, G. N.; Zeng, Q.; Bennet, F.; Jiang, Y.; Lesburg, C. A.; Duca, J.; Pinto, P.; Gavalas, S.; Huang, Y.; Wu, W.; Selyutin, O.; Agrawal, S.; Feld, B.; Huang, H.-C.; Li, C.; Cheng, K.-C.; Shih, N.-Y.; Kozlowski, J. A.; Rosenblum, S. B.; Njoroge, F. G. Structure–Activity Relationship (SAR) Development and Discovery of Potent Indole-Based Inhibitors of the Hepatitis C Virus (HCV) NS5B Polymerase. *J. Med. Chem.* **2011**, *55*, 754–765.
- (28) Habibi-Yangjeh, A.; Pourbasheer, E.; Danandeh-Jenaghara, M. Prediction of basicity constants of various pyridines in aqueous solution using a principal component-genetic algorithm-artificial neural network. *Monatsh. Chem.* **2008**, *139*, 1423–1431.
- (29) Pourbasheer, E.; Riahi, S.; Ganjali, M. R.; Norouzi, P. QSAR study on melanocortin-4 receptors by support vector machine. *Eur. J. Med. Chem.* **2010**, *45*, 1087–1093.
- (30) Riahi, S.; Ganjali, M.; Pourbasheer, E.; Norouzi, P. QSRR Study of GC Retention Indices of Essential-Oil Compounds by Multiple Linear Regression with a Genetic Algorithm. *Chroma.* **2008**, *67*, 917–922.

- (31) Riahi, S.; Pourbasheer, E.; Ganjali, M. R.; Norouzi, P. Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine. *J. Hazard. Mater.* **2009**, *166*, 853–859.
- (32) Pourbasheer, E.; Alalizadeh, R.; Ganjali, M.; Norouzi, P. QSAR study of  $\alpha/\beta$  integrin inhibitors by GA-MLR and GA-SVM methods. *Struct. Chem.* **2013**, *25*, 355–370.
- (33) Pourbasheer, E.; Beheshti, A.; Khajehsharifi, H.; Ganjali, M. R.; Norouzi, P. QSAR study on hERG inhibitory effect of kappa opioid receptor antagonists by linear and non-linear methods. *Med. Chem. Res.* **2013**, *22*, 4047–4058.
- (34) Fratev, F.; Benfenati, E. 3D-QSAR and Molecular Mechanics Study for the Differences in the Azole Activity against Yeastlike and Filamentous Fungi and Their Relation to P450DM Inhibition. 1. 3-Substituted-4(3H)-quinazolinones. *J. Chem. Inf. Model.* **2005**, *45*, 634–644.
- (35) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.
- (36) Pourbasheer, E.; Amanlou, M. 3D-QSAR analysis of anti-cancer agents by CoMFA and CoMSIA. *Med. Chem. Res.* **2013**, *23*, 800–809.
- (37) Du, Q.-S.; Gao, J.; Wei, Y.-T.; Du, L.-Q.; Wang, S.-Q.; Huang, R.-B. Structure-Based and Multiple Potential Three-Dimensional Quantitative Structure–Activity Relationship (SB-MP-3D-QSAR) for Inhibitor Design. *J. Chem. Inf. Model.* **2012**, *52*, 996–1004.
- (38) Tosco, P.; Balle, T. A 3D-QSAR-Driven Approach to Binding Mode and Affinity Prediction. *J. Chem. Inf. Model.* **2011**, *52*, 302–307.
- (39) Tuccinardi, T.; Ortore, G.; Santos, M. A. I.; Marques, S. r. M.; Nuti, E.; Rossello, A.; Martinelli, A. Multitemplate Alignment Method for the Development of a Reliable 3D-QSAR Model for the Analysis of MMP3 Inhibitors. *J. Chem. Inf. Model.* **2009**, *49*, 1715–1724.
- (40) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (41) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.
- (42) HyperChem molecular modeling system, 7.03; Hypercube, Inc., Gainesville, FL, 2002.
- (43) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. DRAGON, 5.3; Talete srl, Milan, Italy, 2005.
- (44) Hocking, R. R. The Analysis and Selection of Variables in Linear Regression. *Biometrics* **1976**, *32*, 1–49.
- (45) Draper, N. R.; Smith, H. *Applied Regression Analysis*, 2nd ed.; John Wiley & Sons, Inc: New York, 1981.
- (46) SYBYL 7.3; Tripos International: St. Louis, MO, 2006.
- (47) Clark, M.; Cramer, R. D.; Van Opdenbosch, N. Validation of the general purpose tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982–1012.
- (48) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228.
- (49) Folkers, G.; Merz, A.; Rognan, D. *3D-QSAR in drug design, theory, methods and applications*; Kubinyi, H., Ed.; ESCOM: Leiden, the Netherlands, 1993; pp 583–618.
- (50) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Model.* **1989**, *29*, 163–172.
- (51) Wold, S.; Ruhe, A.; Wold, H.; Dunn, I. W. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *J. Sci. Stat. Comp.* **1984**, *5*, 735–743.
- (52) Bush, B.; Nachbar, R., Jr. Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA. *J. Comput. Aided Mol. Des.* **1993**, *7*, 587–619.
- (53) Martin, Y. C.; Willett, P. *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*; American Chemical Society: Washington, D.C., 1998.
- (54) Lin, L. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **1989**, *45*, 255–268.
- (55) Riahi, S.; Pourbasheer, E.; Dinavand, R.; Ganjali, M. R.; Norouzi, P. QSAR Study of 2-(1-Propylpiperidin-4-yl)-1H-Benzimidazole-4-Carboxamide as PARP Inhibitors for Treatment of Cancer. *Chem. Biol. Drug Des.* **2008**, *72*, 575–584.
- (56) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69–77.
- (57) Eriksson, L.; Johansson, E.; Müller, M.; Wold, S. On the selection of the training set in environmental QSAR analysis when compounds are clustered. *J. Chemometr.* **2000**, *14*, 599–616.