Effective Descriptions of Molecular Structures and the Quantitative Structure−Activity Relationship Studies

Lu Xu,* Jia-An Yang, and Ya-Ping Wu

Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, People's Republic of China

Received August 30, 2001

In this research, we found CoMFA alone could not obtain sufficiently a strong equation to allow confident prediction for aminobenzenes. When some other parameter, such as heat of molecular formation of the compounds, was introduced into the CoMFA model, the results were improved greatly. It gives us a hint that a better description for molecular structures will yield a better prediction model, and this hint challenged us to look for another method—the projection areas of molecules in 3D space for 3D-QSAR. It is surprising that much better results than that obtained by using CoMFA were achieved. Besides the CoMFA analysis, multiregression analysis and neural network methods for building the models were used in this paper.

INTRODUCTION

The basic assumption in quantitative structure−activity/property relationship (QSAR/QSPR) studies is that the molecular structures possess relations with their activities/properties. Thus, if we have methods to describe effectively the molecular structures, we can build a mathematical model to predict the activities/properties for unknown compounds. Since the 1960s, enormous efforts have been made by various investigators to develop quantitative parameters. During this period, Hansch and co-workers[1−3] made important breakthroughs for biological QSAR with electronic, stereo, and hydrophobic parameters to be known as the extrathermodynamic approach.

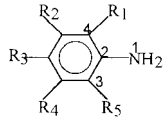
During the decades, though the stereo features, such as the Taft Es,[4] Verloop STERIMOL,[5] molecular volume, molecular weight, and so on, have been proposed for describing stereo effects in biological QSAR, they do not in general include consideration of 3D molecular structures.

Since the 1980s, several methods considering the reaction between three-dimensional (3D) molecular structure and receptor in QSAR have appeared. These methods are all called 3D-QSAR approaches. At the present, the most widely used 3D-QSAR technique is CoMFA (comparative molecular field analysis).[6] CoMFA assumes two basic tenets: at the molecular level, the interactions between a biomolecule and substrate molecule are usually noncovalent in nature, and a sampling of the stereo and electrostatic fields might provide the information necessary to understand their biological properties. However, recently, we found CoMFA alone cannot obtain a sufficiently strong equation to allow confident prediction for aminobenzenes. When some other parameter, such as heat of molecular formation of the compounds, was introduced into the CoMFA model, the results were improved greatly. It gives us a hint that a better description for molecular structures will give a better prediction model, and this hint challenged us to look for another method—the projection areas of molecules in 3D space for 3D-QSAR. It is surprising that much better results than that obtained by using CoMFA were achieved. The details in questions above will be given in this paper.

* Corresponding author phone: 0086-0431-5262239; fax: 0086-0431-5685653; e-mail: luxu@ns.ciac.jl.cn.

2. EXPERIMENTS

2.1. Data. The structures and their activities of aminobenzenes were obtained from RTES.[7] The toxicity LD_{50} is defined as the concentration required to cause the rats to die by 50% and -logLD_{50} as the toxic descriptor. The aminobenzene skeleton is



The different substitutions and their toxic activities are shown in Table 1.

2.2. Descriptors for Toxic Prediction. The descriptors (i.e., features) were derived from the structures of underlying compounds and used to create the predictive models. Abstraction of the features is the key step for a QSAR/QSPR study.

2.2.1. Quantum-Chemical Descriptors. Owing to the importance of quantum-chemical parameters for a QSAR/QSPR, 11 descriptors were calculated by using MOPAC of the option of SYBYL version 6.1 (Tripos Associates, 1699 S. Hanley Road, Suite 303, St. Louis, MO 63144). These descriptors are energies of highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), charge density on atom N, charge density on atom C connected to NH_2, dipole moment, net molecular charge, molecular total energy, nucleus−nucleus repulse energy, ionization potential, electronic energy, and heat of molecular formation. They are denoted respectively by E_H, E_L, Q_N, μ, Q_N, Q_C, Qt, E_p, E_i, E_e, and H_f.

10.1021/ci010092r CCC: $22.00
Published on Web 05/14/2002

Table 1. Substituted Aminobenzenes and Their Toxicity

no.	R_1	R_2	R_3	R_4	R_5	$-\log(LD_{50})$ (obs.)	$-\log(LD_{50})$ (cal.)	diff.
1	Me	H	H	H	H	−2.826	−2.854	0.028
2	H	Me	H	H	H	−2.653	−2.697	0.044
3[a]	H	H	Me	H	H	−2.817	−2.778	−0.039
4	Et	H	H	H	H	−3.100	−3.084	−0.016
5	isoprop	H	H	H	H	−3.072	−3.067	−0.005
6	Me	Me	H	H	H	−2.970	−2.980	0.010
7	Me	H	Me	H	H	−3.113	−3.097	−0.016
8	Me	H	H	Me	H	−2.924	−2.919	−0.005
9[a]	Me	H	H	H	Me	−2.924	−2.917	−0.007
10	H	Me	Me	H	H	−2.909	−2.915	0.006
11	H	Me	H	Me	H	−2.849	−2.834	−0.015
12	Me	H	H	H	Et	−3.072	−3.106	0.034
13	Et	H	H	H	Et	−3.255	−3.276	0.021
14[a]	Me	H	Me	H	Me	−2.871	−2.903	0.032
15	H	CF_3	H	H	H	−2.681	−2.685	0.004
16	H	Cl	H	H	H	−2.408	−2.419	0.011
17	H	H	Cl	H	H	−2.491	−2.501	0.010
18	H	Cl	Cl	H	H	−2.812	−2.767	−0.045
19	Me	Cl	H	H	H	−2.759	−2.694	−0.065
20[a]	Me	H	H	Cl	H	−2.667	−2.669	0.002
21	Cl	H	Me	H	H	−2.565	−2.580	0.015
22	H	H	F	H	H	−2.620	−2.634	0.014
23	H	H	Br	H	H	−2.659	−2.657	−0.002
24	H	NO_2	H	H	H	−2.728	−2.687	−0.041
25	H	H	NO_2	H	H	−2.875	−2.875	0.000
26[a]	Me	H	H	NO_2	H	−2.759	−2.760	0.001
27	NO_2	H	Cl	H	H	−2.602	−2.605	0.003
28	H	NO_2	Cl	H	H	−2.602	−2.601	−0.001
29	NO_2	H	NO_2	H	H	−2.621	−2.634	0.013

[a] Samples of the test set.

2.2.2. Topological Descriptors, Molecular Volume, and logP. Topological indices A_{m1}, A_{m2}, and A_{m3} were calculated, which were suggested by one of the authors based on the augmented distance matrix, and have been used with various kinds of compounds with satisfactory results.[8−10] In addition, we also calculated volume parameter V and hydrophobicity logP supported by the soft package developed by our laboratory.

2.2.3. Shape Descriptors. In many cases, the biological activity or physicochemical property of interest is related to the 3D shape of the tested compounds. The shape parameters presented here are molecular orthogonal projections. The projections were calculated by using a two-dimensional version of the point-encoded algorithm described by Stouch and Jurs.[11] To perform the calculation, the 3D atomic coordinates of the compounds must be available. These coordinates were computed by using SYBYL option BUILD, and the conformations of the compounds were minimized by using the ENERGY MINIMIZE option.

Once the molecular geometry has been defined, the structure is oriented in three-dimensional space according to some defined criterion. For this aim, the superposition procedure of CoMFA was adopted. The details will be given in section 3.2.2.

A molecule, represented by the 3D coordinates of its atoms and their van der Waals radii, is placed in a 3D grid of arbitrary density. Each point of intersection of the grid is checked to see if it lies within the molecule. If so, the point is put into a state, "1"; if not so, the point is put into a state, "0". After this, superimposed structures can be compared point by point, and the molecule can be viewed from three orthogonal directions along the x, y, and z coordinate axes.

Table 2. Results of Leaps-and-Bounds Regression for Best Combinations of the Descriptors

no.	descriptor	R	F	S
1	1	0.084	4.6	0.14
2	1,7	0.831	5.6	0.13
3	1,2,7	0.833	18.9	0.12
4	1,2,3,7	0.868	18.4	0.11
5	1,2,3,5,7	0.901	19.8	0.097
6	1,2,3,5,6,7	0.913	12.5	0.098
7	1,2,3,5,6,7,8	0.913	15.1	0.096

For each perspective, the coordinates are compressed into the plane defined by the remaining two axes. For example, for the perspective along the *z* axis, the *z* axis would be disregarded, and the molecule is projected onto the x−y plane. The area of this projection will be as a parameter of molecular shape to construct the projection model.

2.3. The Selection of the Descriptors. A good equation for structure and activity should possess high correlation coefficient R, low standard deviation S, and least variables. To this end, objective feature selection was done to weed out those descriptors that provide minimal or redundant information. During the process, one descriptor was eliminated from each pair exhibiting a high pairwise correlation, thereby greatly reducing information overlap. Then the remaining descriptors were further analyzed using leaps-and-bounds regression.[12] Because in the subsequent statistical analysis we want to find the best subsets of the descriptors, i.e., when we want to take one, two, or three descriptors, we should find which one or one of the combinations is the best for the predictive model. Leaps-and-bounds regression provides us an effective approach, which can give out the answers quickly. The results using leaps-and-bounds regression are given in Table 2.

3. RESULTS AND DISCUSSION

For comparisons, the different results in three cases will be given as follows.

3.1. Results with the Quantum-Chemical, Topological Parameters, and Volume Parameter V. 3.1.1. Best Subsets of the Descriptors. First, we made primary selection of the descriptors introduced as above. After this screening process, eight descriptors, E_f, E_e, E_p, μ, A_{m3}, E_H, E_L, and V, which are corresponding to the numbers 1, 2, 3, ..., 8, separately, remained. Table 2 shows the best combinations of the eight descriptors for different needs. For example, the best one-variable selection is 1 (E_f, heat of formation). The best two-variable selection is the combination of 1 and 7 (7 is E_H, HOMO energy), and the combination of 1, 2 (2 is E_p, pulse energy), and 7 is the best for three-variable selection, and so on. It is noteworthy to point out that the most important descriptor seems to be 1, i.e., heat of molecular formation, because descriptor 1 appeared in all the best subsets.

3.1.2. Multiple Regression. From statistical viewpoint, the ratio of the number of samples (N) to the number of variables (M) should not be too low. Usually, it is recommended that $N/M \geq 5$. In the situation of this study, we have 29 samples, so five variables should be selected based on the above rule. Table 2 shows that the combination of five descriptors is the best, whereas more descriptors cannot improve evidently the regression results.

These descriptors are 1, 2, 3, 5, and 7, and the regression model is

$$-\log(LD_{50}) = -0.760 - 1.744*10^{-3}E_f - 1.171*10^{-3}E_e + 1.556*10^{-3}E_p - 1.295*10^{-2}A_{m3} - 5.452*10^{-2}E_L \quad (1)$$

$$R = 0.901,\ F = 19.8,\ S = 0.097,\ N = 29$$

where R is correlation coefficient, F is significance test, S denotes standard deviation, and N stands for the number of samples. This equation shows that E_e (electronic energy) and E_p (nucleus-nucleus repulse energy) relate to the toxicity positively, whereas E_f (heat of formation), E_L(LUMO), and A_{m3} index relate to toxicity negatively. This means that compounds possessing higher E_e and E_p or lower E_f, E_L, and A_{m3} will be more toxic.

Evidently, the above equation is not sufficiently strong to allow confident prediction. Thus, we have made some efforts to try to improve the results. For instance, HOMO or LUMO was replaced by HOMO−LUMO, because the HOMO−LUMO gap, i.e., the difference in energy between the HOMO and LUMO, is an important stability index. A large HOMO−LUMO gap implies high stability for a molecule in the sense of its lower sensitivity in chemical reaction. But, the results could not be improved evidently. Therefore, neural networks were performed, since this approach always gives better results than that obtained by using multiple regression.

3.1.3. Neural Networks. In recent years, artificial neural networks have been used widely. Among the neural network learning algorithms, the back-propagation (BP) method is one of the most commonly used methods. The drawback of BP is that the training progresses slowly, because the gradient-descent algorithm is usually used for minimizing the sum-squared-error. In this research, the BFGS quasi-Newton method was used. The advantages of using the BFGS method are that specifying rate or momentum is not necessary and training progresses much more rapidly.[13,14]

The input nodes of the neural network are E_f, E_e, E_p, A_{m3}, and E_L, i.e. the total is 5. The number of the output neuron is one. To avoid overtraining, the test set was used to monitor the training process for networks; that is during the training of the network, the performance was monitored by predicting the values for the compounds in the test set. As long as test set results were improving, training was continued. However, when the test set results ceased to improve, the training was stopped. The results indicated that this method is an effective approach to avoiding overtraining.

Consequently, the entire data set was divided into two groups: 25 compounds as the training set and four compounds as the test set. As a usual rule of the thumb, the weights and bases should be less than the samples in number; thus the model achieved by the network is stationary. Therefore, the number of the hidden neurons should not be great than 3. The experiments showed that in the situations of 1, 2, and 3 hidden neurons, better results could be obtained by using three hidden neurons. So the architecture of an over network was 6:3:1. The results obtained by a neural network are $R = 0.967$, $F = 386.6$, $S = 0.053$. Obviously, these are much better than those obtained by using multiple regression analysis.

3.2. Results with CoMFA Method. For 3D-QSAR studies, CoMFA was carried out to try to improve the results.

3.2.1. Buildup of Model. The model was created by the SYBYL option, and its conformation was minimized using the ENERGY MINIMIZE option. Charges were then generated for each atom using the Gasteiger-Marsili and Hückel method.[15] The former is for σ charge, and the latter is for π charge. The total of molecular charges is the sum of these two parts. The electrostatic fields were calculated based on these charges.

3.2.2. Molecular Superposition for CoMFA Analysis. To compute the stereo and electrostatic fields, it is important to keep the angular orientation; that is, all the molecules to be put in the 3D space should be based on certain rules according to the active groups. This is very important, since the relative interaction energies depend strongly on relative molecular positions. RMS was used as a criterion for the superposition. The smaller the RMS, the better the superposition is. Compound 16 was chosen as the reference molecule for superposition because the molecule is the most toxic compound in the substituted aminobenzenes series. The compounds were superimposed on atoms 1−4 of the reference molecule. As stated above, the RMS should be minimized.

Usually, there are two rules can be used for molecular superposition. One is a superimposing analogue structure on a reference structure based on the rigid part in the molecule. The other one is based on the key atoms. In this work, we have used least-squares fitting to atoms 1−4, which may be important for the activities of aminobenzenes. Though the superposition procedure is arbitrary, the final results will be as the criterion. The other attempts have been tried by us, such as fitting to the six atoms in benzene cycle, but the results are not so good as fitting to atoms 1−4.

3.2.3. Experimental Conditions. CoMFA was performed by using the QSAR option of SYBYL version 6.1. Besides, default values were used. The CoMFA grid spacing was in the x (−22 to 10), y (−22 to 10), and z (−15 to 5) directions. The grid region was generated automatically by the software and was large enough to include all the molecules used to generate the model. The probe atom used to generate the interaction energies at each grid intersection was F^{-1}. At points where repulse stereo interaction energies exceeded 125.520 kJ/mol, these energies were truncated to 125.520 kJ/mol.

3.2.4. Improvement of the Descriptors and the CoMFA Results. As mentioned before, the basic assumption of CoMFA is that the interaction between drugs or toxic compounds and biomacromolecules are usually noncovalent but are stereo and electrostatic forces as well as hydrophobicities. However, the general CoMFA cannot improve the results in this study. The version of CoMFA used in this research includes only stereo and electrostatic parameters. The conventional correlation coefficient given by PLS (five latent variables) is 0.85; the cross-validation correlation coefficient is 0.64. After the addition of parameter logP, the correlation coefficient was not changed obviously.

As indicated in paragraph 3.1.1, H_f (heat of formations) is an important descriptor for predictive model of the underlying compounds. So H_f was taken as the third variable for CoMFA, and the results were improved greatly. The correlation coefficients of the conventional and the cross-

Table 3. Results of Leaps-and-Bounds Regression for Best Combinations of the Descriptors

no.	variable	R	F	S
1	1	0.804	101.0	0.083
2	1,4	0.871	71.1	0.071
3	1,4,7	0.903	55.8	0.065
4	1,4,6,7	0.939	44.8	0.056
5	1,4,5,6,8	0.951	43.2	0.051
6	1,4,5,6,7,8	0.959	41.5	0.049
7	1,2,4,5,6,7,8	0.960	35.2	0.050
8	1,2,3,4,5,6,7,8	0.961	30.0	0.051

validation analyses are, respectively, 0.975 and 0.88. This reveals that the interactions between aminobenzenes and biomacromolecules are determined not only by the stereo and electrostatic forces but also by the heat of formation. The results obtained in this work indicate that the higher the heat of formation of the compounds, the more toxic it is. And the results give us a hint that the even better results can be obtained if an even better method for descriptions of structures of the compounds is used. This hint prompts us to look for another method to try to improve ulteriorly the predictive results. The attempt results will be given in the next paragraph.

3.3. Results with Molecular Shapes. 3D molecular shape is often an important determinant of the physicochemical properties or biological activities of chemical compounds. A set of three descriptors that code the shapes of molecules by calculating the areas of three orthogonal projections was developed as introduced in paragraph 2.2.3. It should be pointed out that if the projection areas alone are being used, the good predictive model could not be obtained. Thus, the combinations of the projection areas with the quantum-chemical parameters and topological indices A_{mi} as well as volume V were observed.

Also as stated in the part of variable selection, at first, we made the primary selection of the descriptors. After this process, eight parameters, E_f, E_p, μ, E_L, A_{m3}, V, P_{xz}, and P_{yz}, remained. Then the eight remaining descriptors were further analyzed by using leaps-and-bounds regression, and Table 3 shows the different best combinations of the eight descriptors.

3.3.1. Multiple Regression. Based on the rule ($N/M \geq 5$) as mentioned above, in this part we also chose the best combination of five descriptors, i.e., 1, 4, 5, 6, and 8. The regression model is

$$-\log(LD_{50}) = -2.580 - 6.846*10^{-4}E_f - 3.648*10^{-2}E_L - 5.906*10^{-3}A_{m3} + 2.177*10^{-2}P_{xz} - 2.525*10^{-2}P_{yz} \quad (2)$$

$$R = 0.951,\ F = 43,\ S = 0.06,\ N = 29$$

where R is correlation coefficient, F stands for significance test, S denotes standard deviation, and N is the number of samples. It is worth noticing that there are three variables in common in eqs 1 and 2. These three variables are E_f (heat of molecular formation), E_L (LUMO energy), and topological index A_{m3} whereas E_e (electronic energy) and E_p (nucleus-nucleus repulse energy) in eq 1 were replaced by P_{xz} (projection area on x−z plane) and P_{yz} (projection area on y−z plane) in eq 2 to lead to much better results (correlation coefficient R from 0.901 ascending to 0.951) being obtained.

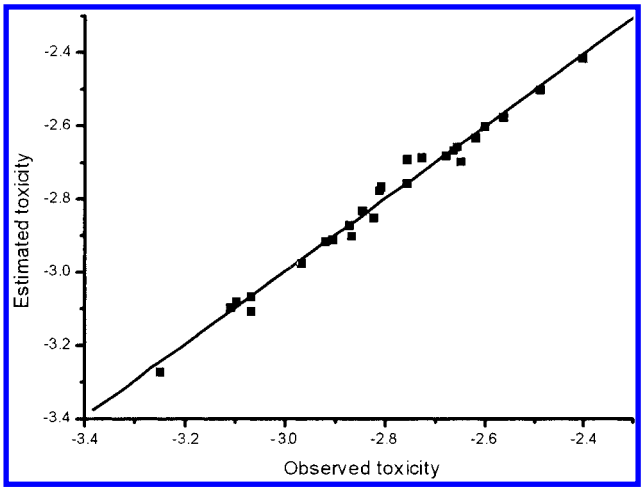


Figure 1. Plot of the toxic activity estimated by the model vs that observed for 29 amonobenzenes.

This means that projection areas P_{xz} and P_{yz} are more effective than E_e and E_p for building the predictive model for the aminobenzenes.

3.3.2. Neural Networks. Five input nodes of the neural network are the same as the multiple regression, i.e., E_f, E_L, A_{m3}, P_{xz}, and P_{yz}. The number of hidden neurons was 2, and the number of the output neurons was 1, so that the architecture of an over network was 5:2:1. The results obtained by a neural network are $R = 0.993$, $F = 1868.9$, $S = 0.023$. Evidently, these are much better than those obtained previously. In addition, cross-validation (leave one out) has been carried out, and the standard deviation S_{cv} is 0.032. This indicates that there seems no chance correlation to happen.

The toxic activities of the 29 compounds in this paper were calculated using the model obtained in this part, as shown in Table 1. These indicate that the largest absolute error for compound 19 is 0.065; the corresponding relative error is 2.3%, which is less than the experimental errors (Table 1). The toxic activities estimated by the QSAR model being plotted vs the toxic activities observed are shown in Figure 1.

4. CONCLUSIONS

From the results in this paper, the following conclusions can be obtained. (1) If we have good methods to describe effectively the molecular structures, such as using the projection areas of the compounds applied in this study to replace the stereo and electrostatic fields to represent the 3D information, the good predictive model can be constructed. (2) The reactions between aminobenzenes and the biomacromolecules are not only based on the stereo, electrostatic energies and quantum-chemical parameters (such as the heat of molecular formation) but also based on the molecular shapes (such as the projection areas). (3) As topological index Am3 represents the size of a molecule, and also the shape of this molecule, thus, again it reveals that shapes are very important for the toxic activities of the aminobenzenes.

ACKNOWLEDGMENT

The authors acknowledge the financial support of the National Natural Science Foundation of China.

REFERENCES AND NOTES

(1) Hansch, C.; Muir, R. M.; Fujita, T.; Maloncy, P. P.; Geiger, F.; Streich, M. The correlation of biological activity of plant growth regulators and chlormycetin derivatives with Hammet constants and partition coefficients. *J. Am. Chem. Soc.* **1963**, *85*, 2817−2824.

(2) Fujita, T.; Iwasa, J.; Hansch, C. A new substituent constant, π, derived from partition coefficients. *J. Am. Chem. Soc.* **1964**, *86*, 5175−5180.

(3) Ramsden, C. A. Quantitative Drug Design, Vol. 4. In *Comprehensive Medicinal Chemistry*; Hansch, C., Ed.; Pergamon Press: New York, 1990.

(4) Taft, R. W. In Steric Effects in Organic Chemistry, 2nd ed.; McGraw-Hill: New York, 1956; pp 556−675.

(5) Verloop, A.; Hoogenstraaten, W.; Tipker, J. *Drug Design* **1976**, *7*, 165−207.

(6) Cramer, III, 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.

(7) RTES (Register of Toxic Effects of Chemical Substances).

(8) Yao, Y. Y.; Xu, L.; Yang, Y. Q.; Yuan, X. S. Studies on structure−activity relationships of organic compounds: three new topological indices and their applications. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 590−594.

(9) Xu, L.; Yao, Y. Y.; Wang, H. M. New topological index and prediction of phase transfer energy for protonated amines and tetra alkylamines ions. *J Chem. Inf. Comput. Sci.* **1995**, *35*, 45−49.

(10) Li, H.; Xu, L.; Yang, Y. Q.; Su, Q. Quantitative structure−property relationships for colour reagents and their reactions with ytterbium using regression analysis and computational neural networks. *Anal. Chem. Acta* **1996**, *321*, 97−103.

(11) Stouch, T. R.; Jurs, P. C. A simple method for representation, and comparison of the volumes and shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4−12.

(12) Furnival, G. M.; Wilson, Jr., R. W. Regression by leaps and bounds. *Technometrics* **1974**, *16*, 499−504.

(13) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative structure−activity relationships for toxicity of phenols using regression analysis and computational neural networks. *Environ. Sci. Chem.* **1994**, *13*, 841−851.

(14) Anonymous. SYBYL molecular Modeling Software Version 5.4, Theory Manual; Tripos Associates, Inc.: St. Louis, MO, 1991; p 2070.

(15) Gasteiger, J.; Marsili, M. Interactive qualization of orbital electronegative−a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219.

CI010092R