Method for the improved semiempirical description of intermolecular interactions of biomolecules and their fragments

N. A. Anikin, V. L. Bugaenko, M. B. Kuzminskii, and A. S. Mendkovich*

N. D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences,
47 Leninsky prosp., 119991 Moscow, Russian Federation.
Fax: +7 (499) 135 5328. E-mail: nikan@swf.chem.ac.ru

A new method was proposed for the improvement of the semiempirical (PM3, *etc.* levels of theory) description of intermolecular potential energy surfaces in biomolecules, primarily hydrophobic dispersion-type interactions. The intermolecular interaction energy calculated by the PM3 method is supplemented with the sum of atom-atom corrections represented in a physically meaningful functional form. The corresponding empirical parameters were selected by the least-squares procedure minimizing the root-mean-square deviation of the intermolecular interaction energies from the reference values calculated by the high-accuracy *ab initio* MP2 method with the quadruple zeta aug-cc-pVTZ basis set. The empirical parameters depend on the valence environment of atoms. The root-mean-square deviation for 74079 reference calculations of small-molecule dimers (with molecular fragments typical of docking complexes) is ~1.6 kJ mol^{-1}, being about 2.5 times lower than that obtained from conventional PM3 calculations (~4.0 kJ mol^{-1}). It is important to take into account weak intermolecular atom-atom pairwise interactions because there is a lot of such interacting pairs in biomolecules.

Key words: intermolecular interactions, quantum chemical methods, semiempirical calculations, biomolecules, docking complexes.

The most widely used semiempirical PM3 method,[1] its predecessor AM1,[2] and the posteriors RM1,[3] PM5,[4] and PM6 (see Ref. 5) methods often provide the description of valence interactions with adequate accuracy. However, the accuracy of calculations of intermolecular interaction energies in biomolecules by these methods is obviously low.[6] For example, hydrogen bonds are incorrectly described by the AM1 method. All these methods poorly describe dispersion interactions, which play a significant role in the docking. For such interactions, it is necessary to substantially improve the available semiempirical methods.

As an example, let us consider the potential curves for the dispersion interaction between two methane molecules (Fig. 1). In the resulting complex, dispersion interactions substantially prevail, whereas Coulomb interactions are weak. Figure 1 shows that the PM3, RM1, and PM6 methods cannot reproduce the reference data calculated by the *ab initio* MP2 method; the simple inclusion of the dispersion component of the standard Lennard-Jones atom-atom potential also does not lead to the improvement of the accuracy.

The problem is complicated not only by a great diversity of geometric parameters of intermolecular contacts in real huge biomolecules (proteins, docking complexes, and so on) but also by the necessity of the correct description of the potential energy surfaces (PES). In these cases, it is not enough to accurately describe only the vicinity of the main intermolecular minimum. For example, only the main minimum for the $H_2O...H_2O$ dimer was well de-

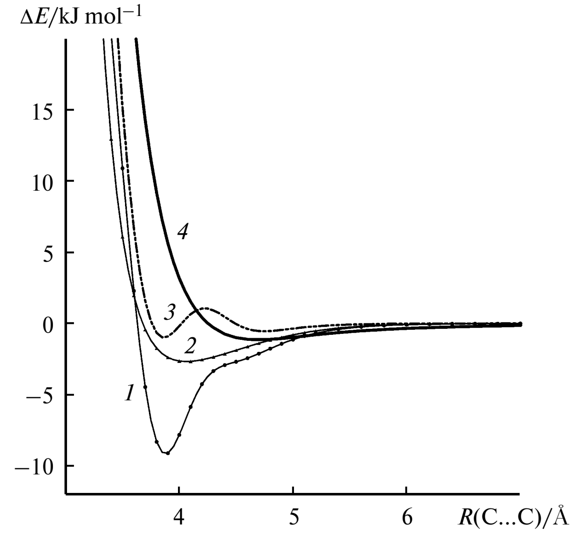

Fig. 1. Comparison of the accuracy of the semiempirical quantum chemical PM3 (*1*), PM6 (*2*), and RM1 (*3*) methods and the *ab initio* MP2/aug-cc-pVTZ approach (*4*) for the dispersion interaction between two methane molecules $H_3CH...HCH_3$ (D_{3h}).

Published in Russian in *Izvestiya Akademii Nauk. Seriya Khimicheskaya,* No. 1, pp. 12—16, January, 2012.

1066-5285/12/6101-12 © 2012 Springer Science+Business Media, Inc.

scribed by the PM3 method,[7] whereas the properties of large clusters and liquid water are described poorly.

Quite successful attempts were made to improve the semiempirical description of some types of intermolecular interactions by the addition of an intermolecular atom-atom potential (PM3-PIF, PM3-MAIS,[7] *etc.* methods) calibrated for the atoms having a specific (often, hydration) environment, primarily, for water clusters[7] and hydrates.[8,9] A similar improvement was proposed in the "semi-*ab initio*" SCC-DFTB method[10] and in the *ab initio* version of the DFT method (DFT-D).[11—13]

However, as shown below, the universal intermolecular atom-atom potential is insufficient for the PM3-type semiempirical description of multidimensional intermolecular PES (not only energy minima, as, for example, in the study[6]) for various biomolecular fragments with different atomic environments. It is necessary to take into account the specificity of the valence environment of atoms, by analogy with the specificity of the parameters in Refs 7—9. For example, the contribution of the H atom in the O—H fragment to dispersion interactions (long-range electron-electron correlation $\sim 1/R_{ij}^{6}$) is much smaller than that of the H atom in the C—H fragment due to a substantially lower electron density on the former atom and the lower polarizability of the O—H fragment combined with the higher activity of the lone pairs of the O atom.

The aim of the present study was to radically improve the description of intermolecular interactions in large biomolecules by semiempirical methods for systematic use. It is known that the intermolecular interaction of an isolated pair of atoms (ignoring hydrogen bonds) is very weak (<1 kJ mol^{-1}); however, there is a huge number of such pairs, *i.e.*, the corresponding contribution to the total energy is considerable. However, an accurate description of these weak interactions is difficult.

Computational methods

Basics of the method. The intermolecular interaction energy can be represented as the sum of the corresponding semiempirical energy calculated by the conventional PM3-type method and an energy correction. This correction $\Delta E_{\mathbf{AB}}{}^{\mathrm{Corr}}$ is represented in a pairwise additive form as the sum of the energies of intermolecular atom-atom interactions, *i.e.*, interactions of atom pairs, in which one atom belongs to molecule **A** and the other to molecule **B**:

$$\Delta E_{\mathbf{AB}}^{\mathrm{Corr}} = \sum_{i}^{\mathbf{A}} \sum_{j}^{\mathbf{B}} \Delta E_{ij}. \qquad (1)$$

In Eq. (1), the subscripts *i* and *j* refer to atoms, **A** and **B** are the interacting molecules (*i*th atom in molecule **A** and *j*th atom in molecule **B**). Meanwhile, the structures **A** and **B** can be fragments of a large biomolecule, whose atoms are involved in the interaction. The latter can be considered as an intermolecular interaction from the standpoint of its chemical nature.

A particular form of the expression for ΔE_{ij} is considered below. Our computational experiments showed that the influence of the valence environment of the *i*th and *j*th atoms on the ΔE_{ij} values should be taken into account when determining the correction $\Delta E_{\mathbf{AB}}{}^{\mathrm{Corr}}$.

The proposed method for calculations of intermolecular interactions can be employed not only in semiempirical schemes but also in various *ab initio* versions of the density functional theory (DFT), which poorly describe dispersion interactions.

Calculations with inclusion of the valence environment of atoms. In the method under consideration, the influence of the valence environment of the interacting *i*th and *j*th atoms, especially H atoms, on their intermolecular contacts is directly taken into account in the calculations of the energy ΔE_{ij}. In other words, the ΔE_{ij} values depend on the valence environment of the *i*th and *j*th atoms in the molecules (fragments) **A** and **B**, respectively. This effect can be taken into account due to a limited assortment of the valence environments typical of actual biomolecules, such as docking complexes, *i.e.*, for ligands and particularly for proteins consisting of different combinations of a limited number of amino acids.

The valence environment of atoms is described by the atom indices stored in the reference database containing the results of high-accuracy calculations (see below). These indices include the coordination index (type of the atom and the number of neighboring atoms) and the extended primary index (list of coordination indices of neighboring atoms).

Functional form for energies ΔE_{ij}. For physicochemical reasons, we represent the expression for ΔE_{ij} as a linear combination of three functions dependent on interatomic distances:

$$\Delta E_{ij} = -c_{ij}/R_{ij}^{6} + a_{ij}\mathrm{e}^{-\alpha(R_{ij}-1)} + b_{ij}\mathrm{e}^{-\beta(R_{ij}-1)}, \qquad (2)$$

where $R_{ij} = r_{ij}/(r_i + r_j)$ are the relative interatomic distances; r_i and r_j are the van der Waals radii of the interacting atoms *i* and *j*, respectively; and a_{ij}, b_{ij}, and c_{ij} are the linear coefficients to be selected. Using the relative distances R_{ij}, one can scale intermolecular interactions for different atom pairs and convert them to a common scale: R_{ij} is close to unity in the vicinity of the minimum of the energy ΔE_{ij}. All three functions dependent on R_{ij} in Eq. (2) are close to unity. Their derivatives with respect to R_{ij} are substantially different.

Hence, in this method, the pairwise atom-atom contributions ΔE_{ij} depend only on the types of the interacting *i*th and *j*th atoms, including their valence environment, and on the (relative) interatomic distance R_{ij} between these atoms. A linear combination of two exponential functions in Eq. (2) describes the short-range repulsion (the coefficients a_{ij} and b_{ij} are empirical parameters; the selection of the exponents α and β is described below), whereas the first term ($-c_{ij}/R_{ij}^{6}$) corresponds to the long-range dispersion attraction (this is indicated by the minus sign at the positive coefficients c_{ij}). The coefficients c_{ij}, like the coefficients a_{ij} and b_{ij}, are determined by the least-squares fitting of the calculated semiempirical interaction energies of the molecules $\Delta E_{\mathbf{AB}}{}^{\mathrm{SE}}$ (with inclusion of the corrections $\Delta E_{\mathbf{AB}}{}^{\mathrm{Corr}}$) to the reference interaction energies of the molecules $\Delta E_{\mathbf{AB}}{}^{\mathrm{MP2}}$, which are calculated by the high-accuracy *ab initio* MP2 method. The first term in Eq. (2) is analogous to the intermolecular energy corrections used in Refs 7—9.

Our computational experience showed that the dimensionless exponents α and β can be adequately fixed and assumed to be universal. This is due to the conversion to the dimensionless relative distances R_{ij} and the presence of two exponential functions covering the range of necessary values of the exponents: the moderate variation of these parameters can be approximately replaced by the variation of the linear coefficients a_{ij} and b_{ij} in front of different exponents. Hence, we chose the fixed values α and β, which, on the average, reproduce the derivative with respect to R_{ij} of our exponential functions in the vicinity of $R_{ij} \approx 1$ (in the vicinity of the energy minimum). This derivative is typical of the short-range repulsive component ($\sim R_{ij}^{-12}$) of conventional Lennard-Jones intermolecular potential. This corresponds to the exponent ~12 ($\mathrm{d}R_{ij}^{-12}/\mathrm{d}R_{ij} = \mathrm{d}\exp[-12(R_{ij} - 1)]/\mathrm{d}R_{ij}$ at $R_{ij} = 1.0$). We chose $\alpha = 12/\sqrt{2}$ and $\beta = 2\alpha$, so that $\sqrt{\alpha\beta} = 12$, *i.e.*, α and β harmonically enclose the value of 12 from above and below. At smaller α values, the exponential repulsive function $\exp[-\alpha(R_{ij} - 1)]$ almost linearly depends on the function $1/R_{ij}^{6}$ in the vicinity of $R_{ij} \approx 1$ due to the similarity of their derivatives (both derivatives are about −6).

For hydrogen bonds (H...N, H...O, H...F, and H...S), it is necessary to include the third exponential function in the expression (2). Three exponential funcitons were also used for H...H pairs because the description of this interaction in the semiempirical PM3 method is too inaccurate. In the triple zeta scheme, the coefficients at the interatomic distances in the exponents were taken, by analogy, equal to 8.4 (12•0.7), 14.4 (12•1.2), and 24.0 (12•2.0). In other cases, the double zeta approximation (2) appeared to be sufficient for obtaining results with satisfactory accuracy.

Selection of empirical parameters by the least-squares method. Parameters were selected by the least-squares procedure minimizing the weighted root-mean-square deviation (rmsd) of the semiempirical energies of interactions between molecules **A** and **B** calculated with inclusion of the correction (1) and the MP2/aug-cc-pVTZ reference energies $\Delta E_{\mathbf{AB}n}{}^{\mathrm{MP2}}$. The summation was performed over all points of all intermolecular PES (for each pair of compounds **A** and **B**, there is its own PES with different intermolecular geometric parameters enumerated with the index *n*):

$$\sum_{\mathbf{A}}\sum_{\mathbf{B}}\sum_{n}^{\mathbf{AB}} W_n(\Delta E^{\mathrm{PM3}}_{\mathbf{AB}n} + \Delta E^{\mathrm{Corr}}_{\mathbf{AB}n} - \Delta E^{\mathrm{MP2}}_{\mathbf{AB}n})^2 \to \min, \qquad (3)$$

where $\Delta E_{\mathbf{AB}n}{}^{\mathrm{PM3}}$ is the intermolecular interaction energy in the conventional semiempirical PM3 method and $\Delta E_{\mathbf{AB}n}{}^{\mathrm{Corr}}$ is the correction (see the expression (1)). The selection of the weights W_n for the optimal description of chemically meaningful interactions is considered below.

The parameters in the expressions for ΔE_{ij} were calibrated by the least-squares method based on the results of ~20000 highly accurate MP2 calculations of the points of the reference intermolecular PES for ~200 various pairs (**A**...**B**) of small monomeric molecules. We used monomers containing single and multiple bonds CH, CC, CN, CO, CF, CS, CCl, NH, NN, NO, OH, SH, and SO with the valence environments typical of proteins and ligands of docking complexes.

Reference calculations were performed by the *ab initio* MP2 method of electron correlation. The extended quadruple zeta aug-cc-pVTZ basis set was used. Our investigations showed that the employment of higher levels of theory has little effect on the accuracy of the description of intermolecular interactions (Fig. 2), but requires significantly greater (by orders of magnitude) computational resources, which is particularly critical in the case of large-scale calculations of tens of thousands (>70000) of reference complexes.

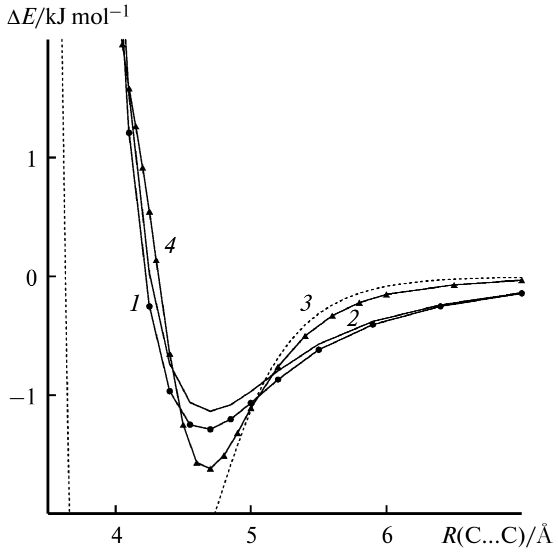

Fig. 2. Comparison of the accuracy of the semiempirical quantum chemical methods (conventional PM3 and improved PM3) and the *ab initio* MP4 and MP2 methods for the dispersion interaction between two methane molecules $H_3CH...HCH_3$ (D_{3h}): MP4/aug-cc-pVTZ (*1*), MP2/aug-cc-pVTZ (*2*), the conventional PM3 method (*3*), and the improved PM3 method (*4*).

In the calculations of intermolecular PES of **A**...**B** complexes, only the mutual arrangement of the monomers (**A** with respect to **B**) was varied, whereas the intramolecular geometry optimized in advance was kept fixed. To reliably determine the parameters responsible for the description of the repulsion, the mutual orientations of the monomers corresponding to the relatively close location of atom pairs of each monomer were taken into account in the reference data set.

For pairs of the nearest interacting atoms from different molecules (fragments), the geometric parameters with several different directions (orientations) of their approach to each other were specified. To improve the efficiency of the analysis, special sequences of points on the PES were generated. In these sequences, the selected, chemically meaningful, internuclear distance was varied, while the mutual orientation of the monomers remained unchanged, or, on the contrary, the distance was kept fixed, while one of the angles was varied.

To determine the parameters c_{ij} in the expression (2), one should have at hand a set of different conformations, including those in which monomers are far remote from each other. Hence, we calculated complexes, in which the nearest atoms of the monomers were at distances of up to 7 Å and more. As a rule, dimers contain one such pair of atoms. However, to check the additivity of interactions, we considered configurations, in which several such intermolecular interactions were present. To gain more detailed knowledge about the PES, an additional set of points for the least-squares fitting corresponding to an arbitrary selection of the dimer geometries (orientations of monomers and distances between them) was introduced.

Selection of the weights W_n. Based on the results of computational experiments and the requirement of their transferability to large docking complexes, a necessary system of the weights W_n for the least-squares fitting was developed for different dimers. The energies of the formation of dimers from monomers obtained in the reference calculations ($\Delta E_{\mathbf{AB}n}{}^{\mathrm{MP2}}$) can differ from those estimated by the PM3 method ($\Delta E_{\mathbf{AB}n}{}^{\mathrm{PM3}}$) by tens or hundreds of times. Hence, in the absence of properly selected weights, the contribution of small, but very important $\Delta E_{\mathbf{AB}n}{}^{\mathrm{MP2}}$ values in the least-squares will be unnoticeable against the background of large but less significant analogs of $\Delta E_{\mathbf{AB}n}{}^{\mathrm{MP2}}$.

At short internuclear distances, *i.e.*, in the repulsive range, the energy differences $\Delta E_{\mathbf{AB}n}{}^{\mathrm{PM3}} - \Delta E_{\mathbf{AB}n}{}^{\mathrm{MP2}}$ are large (up to tens or more kJ mol^{-1}), but the corresponding energetically unfavorable geometric configurations ($\Delta E_{\mathbf{AB}n}{}^{\mathrm{MP2}} >> 0$) are thermodynamically unlikely Hence, we introduced the Boltzmann term W_n dependent on $\Delta E_{\mathbf{AB}n}{}^{\mathrm{MP2}}$ for $\Delta E_{\mathbf{AB}n}{}^{\mathrm{MP2}} > 0.01$ au (~26 kJ mol^{-1}). Meanwhile, weak complexes corresponding to distant monomers (the minimum value $R_{ij} > 1$) are of great importance; there is a lot of such atom-atom interactions in docking complexes, and their total contribution to intermolecular interactions is large. Therefore, to increase the contribution of long-range intermolecular interactions, the expression for W_n includes the term $R_{ij}{}^6$ with the smallest R_{ij} value.

Results and Discussion

When selecting parameters, the detailed description of the atomic environment was done in the course of preliminary trials. In particular, the need for taking into account the effect of the donor-acceptor π-conjugation in the carboxyl (COOH) and amide (CONH) groups on the parameters of N and H atoms (of NH and OH groups) was revealed.

For H, C, N, O, F, S, and Cl atoms, the empirical parameters a_{ij}, b_{ij}, and c_{ij} in the expression (2) were selected for the calculation of the energy correction by the PM3 method according to Eq. (1).

Calculations were performed for 203 atom pairs (the specificity of the environment of these atoms was taken into account) and 96 dimers composed of 15 small molecules: CH_4, C_2H_4, C_2H_2, CH_3F, CH_3Cl, H_2O, CH_2O, NH_3, H_2CNH, H_2S, H_2SO, H_2SO_2, HCOOH, $HCONH_2$, and HCN. The total number of points for the least-squares fitting was 27759. The weighted rmsd for the proposed method is 1.52 kJ mol^{-1} (*cf.* 3.37 kJ mol^{-1} for the initial semiempirical PM3 method).

The achieved increase in the accuracy (by ~2 kJ mol^{-1}) is illustrated by Fig. 2. The potential curve calculated by the PM3 method with the proposed correction is much closer to the reference curve than the curve calculated by the conventional PM3 method (on this scale, some parts the curve calculated by the PM3 method go off scale). If it were sufficient to reproduce only the vicinities of individual energy minima rather than multidimensional PES (which were used to select the parameters for Eq. (2)), the problem could be solved much more easily and accurately.

To verify the transferability of the selected parameters to other complex structures, we additionally performed 46320 calculations. They included 311 other dimers composed of 67 relatively larger and chemically more interesting monomers: CH_4, C_2H_6, C_3H_8, CH_2CH_2, CH_3CHCH_2, $CH_2CHCHCH_2$, cyclopentadiene, HCCH, CH_3CCH, C_6H_6, CH_3F, CH_3Cl, CH_2F_2, CH_2Cl_2, CH_2FCl, C_2H_5F, CH_2CHF, CH_2CHCl, CH_3CF_3, H_2O, CH_3OH, CH_3CH_2OH, $(CH_3)_2CHOH$, $(CH_2OH)_2$, CH_2CHOH, $(CH_3)_2O$, CH_2O, CH_3CHO, CH_2CHCHO, $(CH_3)_2CO$, cyclopentadienone, NH_3, CH_3NH_2, $CH_3CH_2NH_2$, $(CH_3)_2NH$, $(CH_3)_3N$, CH_2NH, CH_2CHNH_2, CH_2NOH, CH_2NOCH_3, CH_3NNCH_3, H_2S, H_2S_2, CH_3SH, $(CH_3)_2S$, CH_3CH_2SH, $(CH_3)_2S_2$, $(CH_3)_2SO$, H_2SO_2, $(CH_3)_2SO_2$, HSO_2NH_2, $CH_3SO_2NH_2$, HCOOH, CH_3COOH, $HCOOCH_3$, HCN, $HCONH_2$, $HCONHCH_3$, CH_3CN, CH_3CONH_2, furan, thiophene, pyrrole, imidazole, pyridine, pyrazine, and uracil. The weighted rmsd for the proposed method was 2.30 kJ mol^{-1} (*cf.* 4.34 kJ mol^{-1} for PM3 calculations).

As should be expected, the accuracy for these more complex and diverse structures, which were not used in the least-squares optimization of the parameters, is lower, as well as in the calculations by the initial PM3 method; however, the loss in the accuracy is not too large (0.8—0.9 kJ mol^{-1}). In this case, the proposed method is also ~2 times more accurate than the initial PM3 method, *i.e.*, the transferability of the parameters is satisfactory.

In the case of the joint optimization of the parameters for all dimers (a total of ~74080 calculations of the structures both used for the least-squares optimization and included in the extended data set), the weighted rmsd for the proposed method is 1.56 kJ mol^{-1}; for the initial PM3 method, 4.03 kJ mol^{-1} (for 27759 calculations of more simple structures (see above) the weighted rmsd was 1.52 and 3.37 kJ mol^{-1}, respectively), *i.e.*, for more diverse and chemically more interesting structures, the gain in the accuracy is even larger (2.6 times).

Evidently, an analogous substantial increase in the accuracy of the description of intermolecular PES can be achieved by other widely used semiempirical methods and *ab initio* versions of DFT with the systematic consideration of the specificity of the valence environment of interacting atoms belonging to different molecules.

Thus, using the PM3 method as an example, we proposed and tested an approach, which substantially improves the description of intermolecular interactions by semiempirical methods. The novel approach is based on the addition of empirical atom-atom potentials calibrated by the least-squares method to perform efficient and correct large-scale calculations of intermolecular interaction energies for different parts of intermolecular PES chemically typical of protein—ligand docking complexes.

The method accounts for the strong effect of the valence environment of the interacting atoms (particularly H atoms) belonging to different molecules, on their intermolecular interactions, which is specific to docking complexes. This is combined with a limited assortment of these environments typical of docking complexes. The proposed approach involving the refinement of intermolecular parameters depending on the valence environment can be employed for substantial improvement of the accuracy of the description of intermolecular interactions by other widely used semiempirical methods and *ab initio* versions of DFT.

References

1. J. J. P. Stewart, *J. Comp. Chem.*, 1991, **12**, 320.
2. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902.
3. G. B. Rocha, O. R. Freire, A. M. Simas, J. J. P. Stewart, *J. Comput. Chem.*, 2006, **27**, 1101.
4. J. J. P. Stewart, *J. Mol. Model.*, 2004, **10**, 6.
5. J. J. P. Stewart, *J. Mol. Model.*, 2007, **13**, 1173.
6. J. P. McNamara, I. H. Hillier, *Phys. Chem. Chem. Phys.*, 2007, **9**, 2362.
7. G. Monard, M. I. Bernal-Uruchurtu, A. van der Vaart, K. M. Merz, M. F. Ruiz-Lopez, *J. Phys. Chem. A*, 2005, **109**, 3425.
8. M. I. Bernal-Uruchurtu, M. F. Ruiz-Lopez, *Chem. Phys. Lett.*, 2000, **330**, 118.
9. W. Harb, M. I. Bernal-Uruchurtu, M. F. Ruiz-Lopez, *Theor. Chem. Acc.*, 2004, **112**, 204.
10. M. Elstner, P. Hobza, T. Frauenheim, S. Suhai, E. Kaxiras, *J. Chem. Phys.*, 2001, **114**, 5149.
11. R. Sharma, J. P. McNamara, R. K. Raju, M. A. Vincent, I. H. Hillier, C. A. Morgado, *Phys. Chem. Chem. Phys.*, 2008, **10**, 2767.
12. P. Jurecka, J. Cerny, P. Hobza, D. R. Salahub, *J. Comput. Chem.*, 2007, **28**, 555.
13. Y. Bouteiller, J. C. Poully, C. Desfranois, G. Grgoire, *J. Phys. Chem. A*, 2009, **113**, 6301.

Received July 1, 2010;
in revised form March 21, 2011