Functional group corrections to the GFN2-xTB and PM6 semiempirical methods for noncovalent interactions in alkanes and alkenes

Enrique M. Cabaleiro-Lago, Berta Fernández, Roberto Rodríguez-Fernández, Jesús Rodríguez-Otero and Saulo A. Vázquez**

Departamento de Química Física, Facultade de Química, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain.

ABSTRACT

Analytical corrections were developed to improve the accuracy of the PM6 and GFN2-xTB semiempirical quantum mechanical (SQM) methods for the evaluation of noncovalent interaction energies in alkanes and alkenes. We followed the approach of functional group corrections (FGC), wherein the atom-atom pair corrections depend on the nature of the interacting functional groups. The training set includes 21 alkane and 13 alkene complexes taken from the Donchev's database [Scientific Data **8** (2021) 55], with interaction energies calculated at the CCSD(T)/CBS level, as well as our own data obtained for medium-size complexes (of 100 and 112 atoms). In general, for the systems included in the training and validation sets, the errors obtained with the PM6-FGC and xTB-FGC methods are within chemical accuracy.

I. INTRODUCTION

Intermolecular interactions or, more generally, noncovalent interactions, play a key role in many physical, chemical, and biological processes. For example, they determine states of matter, physical properties of substances, conformations and dynamics of proteins and other macromolecules, and so on. Even for processes involving exceedingly small systems, intermolecular interactions may play a crucial role. An example of the latter is the Cl + HD bimolecular reaction in the gas phase. For this system, Skouteris et al.[1] performed accurate calculations and found that the van der Waals forces in the entrance channel lead to a strong preference for DCl production. Using the same level of theory on a potential energy surface, but without considering the Cl···HD intermolecular interaction, they found similar abundances for the HCl and DCl products.

For the study of small-size systems, as the abovementioned Cl + HD reaction, one may employ electron-correlated ab initio methods, such as full configuration interaction or coupled cluster approaches, which provide a very accurate description of noncovalent interactions. When the system size increases, these methods become prohibitive from a computational point of view. In these cases, methods based on density functional theory may still provide a good description. For relatively large systems or when long time simulations are needed, it may be necessary to use semiempirical quantum mechanical (SQM) methods to reduce the computational demands, although at the cost of significantly decreasing the accuracy.

Traditionally, the most popular SQM methods have been those based on approximations to Hartree-Fock theory,[2-4] such as AM1,[5] PM6,[6] PM7,[7] and OM*x*.[8-11] More recently, the alternative approach based on density functional tight binding (DFTB) has attracted much attention,[2, 12-14] and the latest methods (e.g., GFN2-xTB[13]) appear to

be quite successful for the description of noncovalent interactions and other molecular properties. In general, however, the accuracy of SQM methods is low or moderate due to the strong approximations included in the corresponding derivations. Specifically, integral approximations, minimal basis sets, and the lack or poor treatment of electron correlation in the "parent" approaches lead not only to a deficient description of dispersion, but also of electrostatics, polarization, and exchange repulsion interactions.

The most common strategy for improving the description of noncovalent interactions by SQM methods has been the inclusion of analytical corrections, written as functions of atomic internal coordinates, which do not increase the computational cost significantly. An example is the D3H4X correction of Řezáč and Hobza,[15, 16] which was derived to improve the accuracy of PM6 and other SQM methods. The D3H4X correction comprises Grimme's D3 dispersion correction,[17] corrections for hydrogen and halogen bonding, and a hydrogen-hydrogen repulsion term for hydrocarbons. The analytical correction for hydrogen bonding was designed to be used for geometry optimizations and molecular dynamics simulations without any limitations. It includes radial and angular terms, a proton-transfer term, as well as scaling factors for charged H-bonds and for hydrogen bonds involving a water molecule as donor. The D3H4X correction is the latest of a series of corrections developed by Řezáč, Hobza, and their co-workers,[15, 16, 18, 19] and is perhaps the most advanced correction for SQM methods based on approximations to the Hartree-Fock theory. The D3H4 (i.e., without the correction for halogen bonding) was also parametrized for SCC-DFTB.[15] The corrections were developed using the S66 database[20-22] as the training set, which includes 66 molecular complexes that cover the most common types of noncovalent interactions in biomolecules. Among the methods considered by Řezáč and Hobza, PM6-D3H4, DFTB-D3H4, and RM1-D3H4 give errors lower than 4 kJ mol^{-1} (i.e., chemical accuracy) in several benchmark databases.[15]

As already mentioned, DFTB approaches have experienced significant improvements over the last years. Here, we give special attention to the recent GFN2-xTB method,[13] which has rapidly become very popular for the calculation of noncovalent interactions, structures, and conformational stabilities in systems composed of about 1000 atoms. GFN2-xTB is the first broadly parameterized DFTB method to include electrostatics and exchange-correlation terms beyond the monopole approximation. This results in a more physically sound method, which uses global and element-specific parameters, and which does not require any specific correction for hydrogen or halogen interactions. This method, therefore, represents a qualitative step forward in the development of SQM methods.

Very recently, we have proposed a simple approach to improve the accuracy of SQM methods for the description of noncovalent interactions.[23, 24] Basically, we use pairwise corrections in which the parameters depend on the nature of the interacting functional groups. For this reason, we use the acronym FGC (from functional group corrections) to name our method. For the development of the corrections, we chose small molecules as representatives of relevant functional groups (e.g., methane, formic acid, methylamine, acetamide), and we evaluated intermolecular potential energy curves (IPECs) for bimolecular complexes of these representatives, using a benchmark method and the selected SQM method, specifically, the PM6 Hamiltonian. Although PM6 precedes the more recent PM7 method,[7] the latter has not been recommended due to the damping and truncation of the dispersion energy at long range,[25] which leads to significant errors in the interaction energies calculated for large systems.[12, 26] As shown in our previous work,[23, 24] the inclusion of IPECs for different orientations of the interacting molecules in the training set is crucial for developing well-balanced corrections. As described in the next section, rather than using specific terms to correct

for dispersion or hydrogen bonding interactions, we employ simple and general pairwise functions that can account for the global deficiencies of the SQM methods. The corrections are parameterized by fits to the differences of interaction energies between the reference and the SQM data. The high parameter specificity of the FGC approach facilitates the development of rather accurate corrections. The cost, however, is the vast amount of work needed to extend the method to many of the functional groups relevant in chemistry.

Application of the PM6-FGC method to complexes of diglycine and dialanine provided interaction energies in good agreement with the reference method, showing a clear improvement over the PM6[6] and PM6-D3H4[15] methods.[23, 24] However, for other type of systems the agreement was not as good, indicating that further work is needed to improve our method. This is the case of alkanes, for which we only included the methane dimer in our previous training set.[23] Applying the PM6-FGC method to a series of alkane complexes reported in the BEGDB database,[27] we obtained very good agreement for simple alkanes, similar to that found for the PM6-D3H4 method;[15] but for branched alkanes (e.g., neopentane) the performance of PM6-FGC was worse than that of PM6-D3H4, although the improvement over PM6 was clear.[23] This points out that it is necessary to expand the training set by including additional complexes in order to improve the accuracy of the corrections.

In the present work, we developed corrections for PM6 and GFN2-xTB to improve the description of noncovalent interactions in alkanes and alkenes. To this end, for the training set we used interaction energies of complexes included in the recent benchmark database of Donchev et al.,[28] calculated using the coupled cluster singles and doubles with perturbative triples correction [CCSD(T)] correlation method[29] at the complete basis set (CBS) limit, which is accepted as the "gold standard" for noncovalent

interactions.[30] For validation, we employed our own data, as described in the next section. The performance of our methods, denoted here as PM6-FGC and xTB-FGC (for simplicity, we omit the acronym GFN2), is compared with those of PM6-D3H4 and GFN2-xTB.

II. COMPUTATIONAL DETAILS

As already mentioned, for the training set we employed interaction energies and geometries taken from the database reported by Donchev et al.[28] Specifically, we used interaction energies collected in the DES370K database, evaluated at the CCSD(T)/CBS level of theory. This database contains scans for complexes in which the monomer geometries were optimized at the second-order Møller-Plesset (MP2) perturbation theory,[31] as well as scans for complexes with out-of-equilibrium geometries selected from molecular dynamics simulations. As shown in the next section, the size of the systems included in this database is relatively small. One of the intended uses of SQM methods relays on their application to large systems, where more rigorous approaches are hardly applicable. Taking this into account, to validate the corrections, we considered complexes of large molecules. In our first validation tests, we found that, to improve the results in these systems, it is important to include in the training set a few structures of medium-size complexes. The medium-size and large-size systems selected for this study are described in the next paragraph.

We considered two alkane homodimers of 56 and 92 atoms per monomer, and two alkene homodimers of 50 and 80 atoms per monomer. The monomers are specified in Fig. S1 of the supplementary material. Since the hydrocarbons are largely flexible, we performed conformational searches to obtain suitable structures for the dimers of each hydrocarbon, using the CREST program[32] with the GFN-FF force field.[33] After these

searches, for each medium-size dimer (50 and 56 atoms per alkene and alkane monomer, respectively), a subset containing the 20 most stable structures were selected for further calculation of reference interaction energies. Ten of them (five for the alkane dimer and five for the alkene dimer) were included in the training set for the purpose of refining the long-range behaviour of the FGC correction, whereas the remaining values were employed for validation. For each large-size dimer (80 and 92 atoms per alkene and alkane monomer, respectively), we selected the ten most stable structures for energy refinement, and to be used for validation purposes only. For the medium-size complexes, the interaction energies were calculated at the domain-based local pair natural orbital[34] DLPNO-CCSD(T)/CBS level by extrapolating the interaction energy at the MP2 level, using the cc-pVTZ and cc-pVQZ basis sets[35] and the procedure of Halkier[36] and Helgaker.[37] The MP2 calculations were carried out with the resolution-of-the-identity RI approach,[38, 39] whereas the TightPNO setting was employed for DLPNO-CCSD(T)/cc-pVTZ. The RIJCOSX approach[40] was used in both cases. For comparison purposes, we also used several DFT functionals, namely, the B3LYP, PW6B95, and PBE0 hybrid functionals, the B97M meta-GGA,[41] and the wB97M range-separated hybrid meta-GGA.[41] In all cases, we used the def2-TZVPP basis set and included Grimme's D3 dispersion correction, Becke-Johnson damping and three-body terms, D3(BJ,ABC).[17, 42] We also considered, for comparison purposes, the recent r^{2}SCAN-3c composite method of Grimme and co-workers.[43] For the large-size complexes, the interaction energies were only computed with the DFT methods. All these calculations were performed with the Orca 5.0 program.[44]

The noncovalent potential-energy corrections to the PM6 and GFN2-xTB methods are given as the pairwise sum of terms in Eq. (1). The selection of this functional form was justified in our proof-of-concept work.[23] As shown there, the differences

between the reference and the semiempirical IPECs have, in general, the form of typical intermolecular potential energy curves or of decaying exponentials with negative amplitudes, in agreement with Eq. (1),

$$E_{\text{corr}} = \sum_i \sum_j f_{\text{cut}}(r_{ij}) \times \left\{ A_{ij} e^{-B_{ij} r_{ij}} + \frac{C_{ij}}{r_{ij}^{D_{ij}}} \right\} \qquad (1)$$

where indexes i and j refer to atoms belonging to different interacting molecules, and r_{ij} is the interatomic distance between atoms i and j. The parameters A_{ij}, B_{ij}, C_{ij} and D_{ij} depend on the nature of the considered pair of atoms. The B_{ij} and D_{ij} parameters are restricted to real positive numbers. However, the A_{ij} and C_{ij} parameters may be either positive or negative. $f_{\text{cut}}(r_{ij})$ is a cutoff function introduced to remove the correction at very short r_{ij} distances.

$$f_{\text{cut}}(r_{ij}) = \left(1 + \tanh\left(s_{ij}\left(r_{ij} - d_{ij}\right)\right)\right)/2 \qquad (2)$$

In Eq. (2), s_{ij} is a parameter, set to 10 as in our previous work, which controls the strength of the damping for the interaction between atoms i and j, and d_{ij} is the distance at which the cutoff function takes the value ½. We notice that at very short intermolecular distances, for which one or more interatomic distances are close to the corresponding d_{ij} parameter values, the errors in the calculation of energies and especially forces may be substantial. However, considering the fittings performed in the present work, this possibility may only occur at high energies (e.g., higher than 50-100 kJ/mol), which are not relevant for biochemical processes.

In the present work, the parameters were obtained through fits to differences between the interaction energies evaluated at the CCSD(T)/CBS level and those

computed with each of the semiempirical methods. The PM6 calculations were carried out with the MOPAC2016 program.[45] We used a least-squares nonlinear fitting procedure based on a genetic algorithm,[46, 47] as implemented in a new version of the GAFit code,[48] modified to facilitate simultaneous fittings for different systems. The fittings were conducted by minimizing the following objective function, χ^2,

$$\chi^2(\boldsymbol{a}) = \sum_{i=1}^{N}[y_i - \mathrm{f}(\boldsymbol{x_i}; \boldsymbol{a})]^2 \times w_i \qquad (3)$$

where i runs over all the considered intermolecular geometries for all the studied complexes, y_i denotes the CCSD(T)/CBS − SQM energy difference for geometry i, $\boldsymbol{x_i}$ represents the matrix of the internuclear-intermolecular distances for geometry i, $\boldsymbol{a}$ is the collective variable formed by the total number of parameters, and $\mathrm{f}(\boldsymbol{x_i}; \boldsymbol{a})$ is the correction evaluated by Eqs. (1) and (2) for geometry i using the set of parameters $\boldsymbol{a}$. The square of the difference between the CCSD(T)/CBS − SQM energy difference for point i (i.e., y_i) and the corresponding $\mathrm{f}(\boldsymbol{x_i}; \boldsymbol{a})$ value is multiplied by a weighting factor (w_i) for each data point. For each of the Donchev's complexes, we used $w_i = \exp(E_i/E_0)$, where E_0 is the energy of the complex in the global minimum and E_i is the energy corresponding to geometry i. For the medium-size complexes, we used $w = 20$. When the number of fitting parameters is large, genetic algorithms provide an efficient way to search for near-optimal solutions. Since genetic algorithms may lead to many solutions that can be equally valid, our corrections should better be viewed as whole functional group corrections, rather than as individual pairwise corrections.

III. RESULTS

A. Training set

As already mentioned, for the training set we employed geometries and interaction energies taken from the database reported by Donchev et al.,[28] as well as five geometries and interaction energies of a medium-size alkane homodimer, and five of a medium-size alkene homodimer. The systems taken from the Donchev's database are specified in Table I. We use capital letters A and E to denote the complexes of alkanes and alkenes, respectively. To specify the nature of the complexes, the corresponding simplified molecular-input line-entry system (SMILES) strings[49] are reported. For alkanes, we considered 21 different systems, which comprise a total of 1873 geometries. The number of systems for alkenes is smaller than for alkanes (13), but the total number of points is larger (2191).

In addition to the Donchev's data, we included a few structures of medium-size homodimers, as anticipated in the previous section. The selected monomers are depicted in Fig. S1 of the supplementary material. The interaction energies calculated in this work for these medium-size complexes are collected in Table SI of the supplementary material. In particular, for the training set we selected the $C_{18}H_{38}$-1 to $C_{18}H_{38}$-5 (alkanes) and the $C_{18}H_{32}$-1 to $C_{18}H_{32}$-5 (alkenes) geometries. The remaining geometries were used for validation purposes.

Table I. Alkane and alkene complexes included in the training set, together with the SMILES string, and the number of points (N_p) and orientations (N_o) for each complex.

Code	SMILES string	N_p	N_o	Code	SMILES string	N_p	N_o
A1	CC_CC	280	11	**E1**	C=C_C=C	286	11
A2	CCC_CCC	295	11	**E2**	CC=C_CC=C	286	11
A3	CCCC_CCCC	284	11	**E3**	CC=CC_CC=CC	191	7
A4	C_CCCC	26	1	**E4**	CCC=C_CCC=C	259	10
A5	CC_CCCC	25	1	**E5**	CCC=CC_CCC=CC	207	8
A6	CCC_CCCC	26	1	**E6**	CC(=C)C_CC(=C)C	293	11
A7	CCCC_CCCCC	27	1	**E7**	CC=C(C)C_CC=C(C)C	293	11
A8	CCCC_CCCCCC	27	1	**E8**	CCC(=C)C_CCC(=C)C	241	9
A9	CCCCC_CCCCC	27	1	**E9**	C=C_CC=C	27	1
A10	CC(C)C_CCCC	27	1	**E10**	C=C_CC=CC	27	1
A11	CC(C)(C)C_CCCC	28	1	**E11**	C=C_CC(=C)C	27	1
A12	CC(C)C_CC(C)C	291	11	**E12**	C=C_CC=C(C)C	27	1
A13	CC(C)(C)C_CC(C)(C)C	30	1	**E13**	C=C_CC(=C(C)C)C	27	1
A14	C1CCCCC1_CC(C)C	28	1				
A15	C1CCCC1_C1CCCC1	285	11				
A16	C1CCCC1_C1CCCCC1	29	1				
A17	C_C1CCCCC1	27	1				
A18	C1CCCCC1_CC	27	1				
A19	C1CCCCC1_CCC	28	1				
A20	C1CCCCC1_CCCC	28	1				
A21	C1CCCC1_CCCC	28	1				

B. Parameterization

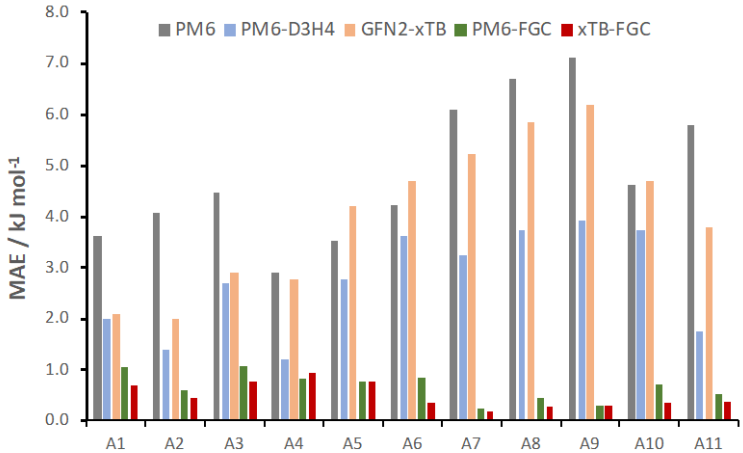
The parameterization of the FGC corrections involved several series of independent fits, following an iterative process. In the first iteration, we set the lower and upper limits of the parameters such that we constrained the parameter ranges to be relatively wide. Typically, parameter A_{ij} is constrained to be between -10^6 and 10^6 kJ·mol^{-1}, parameter B_{ij} between 2.0 and 6.0 Å^{-1}, parameter C_{ij} between –3000 and 3000 kJ·Å$^{D_{ij}}$·mol^{-1}, and parameter D_{ij} between 2 and 10. For parameter d_{ij}, we considered, for each type of atom pair, an interval of 4 Å wide centered around the corresponding minimum internuclear distance found in the geometries used for the fittings. With these parameter ranges, we run 20 or more independent fits, each involving around 1 million

evaluations of the objective function, i.e., Eq. (3). After the analysis of the results, the lower and upper parameter limits are modified, narrowing the parameter ranges to find better solutions in the subsequent runs. The process is repeated several times until convergence is achieved and only a marginal gain in the quality of the fit is obtained with the new parameters as compared with the ones from the previous iteration. In general, convergence may be achieved after five iterations. We applied this whole iterative process to different data sets to select a final training set for this study. The final training set was specified in the previous section. As mentioned above, in general, the application of genetic algorithms in optimization problems involving many parameters leads to multiple solutions that can be equally valid. Discussions on the use of genetic algorithms for parameterizations of SQM methods are reported elsewhere.[50, 51]

In this work, to develop the FGC corrections for the PM6 and GFN2-xTB methods, we defined two types of alkane atoms, namely, alkane carbon (denoted as C3) and alkane hydrogen (denoted as H3). Therefore, in all we have 15 parameters, which were fit simultaneously. In our first trials, we only included the Donchev's complexes in the training set. For the fittings, we found that the corrections derived for PM6 overestimated the strength of the noncovalent interactions in medium-size systems. By contrast, the corrections parameterized for the GFN2-xTB method led to underestimation. Considering this, we included five geometries of the medium-size alkane complex, using a weight of 20 for each point. In all, we considered 1878 points for the 22 alkane systems.

The parameters obtained in the best fits for the PM6 and GFN2-xTB corrections are collected in Table SII of the supplementary material, respectively. With our corrections, we calculated the interaction energies for the systems included in the training set, as well as the mean absolute errors (MAEs) and mean signed errors (MSEs), using the CCSD(T)/CBS interaction energies as reference. The MSE is calculated as the mean

of the differences between the SQM and the reference values. The MAEs and MSEs for the Donchev's systems included in the training set are shown graphically in Figs. 1 and 2, respectively. For comparison, we also include the results obtained with PM6, PM6-D3H4, and GFN2-xTB. As can be seen, the improvement over the PM6-D3H4 and GFN2-xTB results is clear. Most of the MAE values calculated by the PM6-FGC and xTB-FGC are below 1 kJ mol^{-1}, which indicates excellent agreement with the reference data. It is important to notice that all the systems shown in the comparison of Fig. 1 were included in the FGC parameterizations, so that the better performance of the FGC approach can be expected. However, the fact that the training set includes a large number of small-size alkanes suggests that, in general, the PM6-FGC and xTB-FGC methods significantly improve the description of noncovalent interactions in this type of systems with respect to the PM6-D3H4 and GFN2-xTB approaches. For the methane-cyclohexane complex (A17) and the neopentane dimer (A13), the PM6-D3H4 slightly outperforms our methods. In general, the PM6-D3H4 results are better than those provided by the GFN2-xTB method.



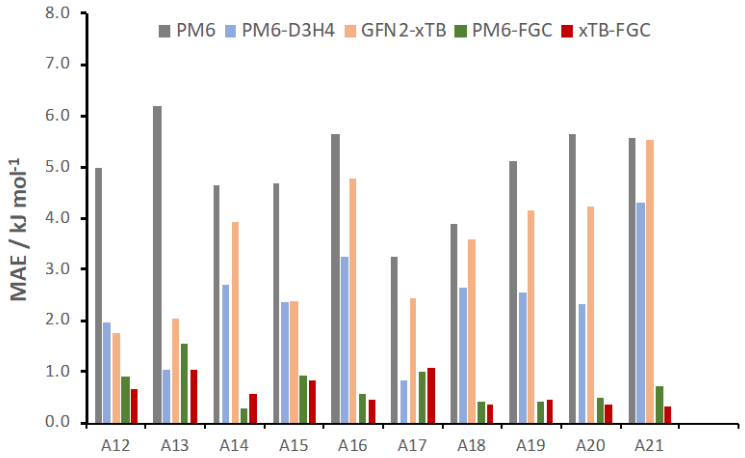
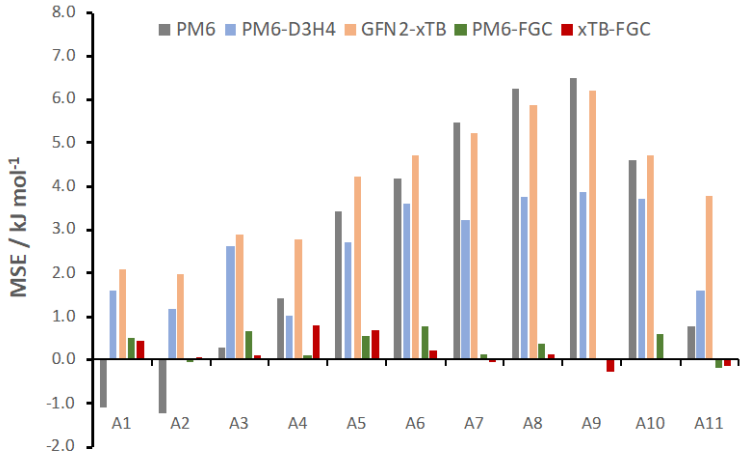


Fig. 1. Mean absolute errors obtained with the PM6-FGC and xTB-FGC methods for the Donchev's complexes included in the alkane training set, and comparison with the results calculated with the PM6, PM6-D3H4 and GFN2-xTB methods.

To show whether there is a bias towards overestimation or underestimation of the interaction energies, we depict in Fig. 2 the corresponding MSEs. As can be seen, the PM6-FGC and xTB-FGC methods do not exhibit a clear bias. By contrast, both the PM6-D3H4 and the GFN2-xTB methods clearly underestimate the strength of the noncovalent interactions in these systems. Notice that the MSE is calculated as the mean of the differences between the SQM results and the reference values.



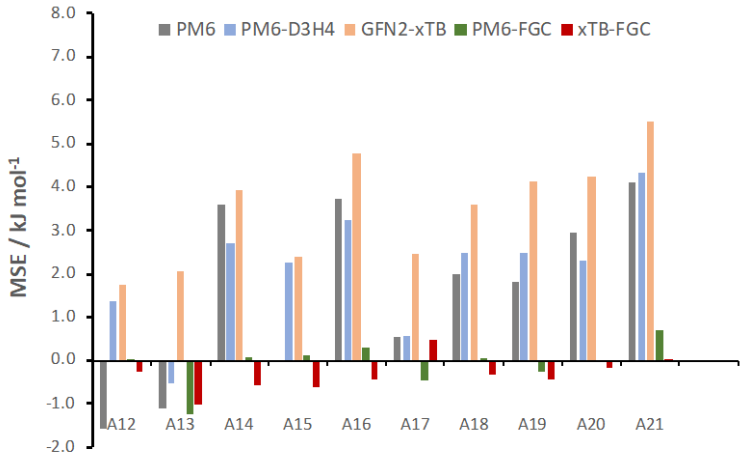


Fig. 2. Mean signed errors obtained with the PM6-FGC and xTB-FGC methods for the Donchev's complexes included in the alkane training set, and comparison with the results calculated with PM6, PM6-D3H4 and GFN2-xTB.

Most of the alkene systems included in the training set also contain saturated carbon atoms. Therefore, for the parameterization of the alkene corrections, we used the values of the parameters involving C3 and H3 atoms obtained in the alkane fits and given in Table SII. As for alkanes, we defined two atom types for the alkene groups, namely, C2 and H2 atoms. Consequently, for each method, the alkene parameterizations involved the simultaneous fitting of 35 parameters. The parameters obtained in the best fits performed for the PM6 and GFN2-xTB corrections are listed in Table SII of the supplementary material.

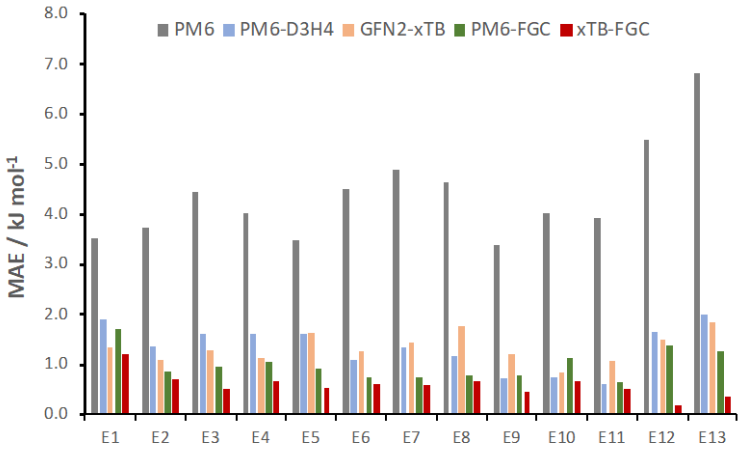


Fig. 3. Mean absolute errors obtained with the PM6-FGC and xTB-FGC methods for the Donchev's complexes included in the alkene training set, and comparison with the results calculated with the PM6, PM6-D3H4 and GFN2-xTB methods.

We calculated the interaction energies with the PM6-FGC and xTB-FGC methods and compared them with the reference values to obtain the MAEs and MSEs, which are displayed in Figs. 3 and 4, respectively. For comparison, we also include the results obtained with the PM6-D3H4 and GFN2-xTB approaches. Inspection of Fig. 3 shows that the performance of these last two methods for alkenes is much better than for alkanes. In general, the results determined with the PM6-FGC and xTB-FGC methods are better than those obtained with PM6-D3H4 and GFN2-xTB, but now the differences are smaller than in alkanes. Except for E1, which corresponds to the ethylene dimer, the PM6-FGC and xTB-FGC MAEs are about 1 kJ mol^{-1} or smaller, providing xTB-FGC the best results. The calculated MSEs (Fig. 4) indicate that, in general, the PM6-D3H4 and GFN2-xTB methods underestimate the strength of the interaction energy in these alkene systems, resembling the results obtained for alkanes, although they show better performance.

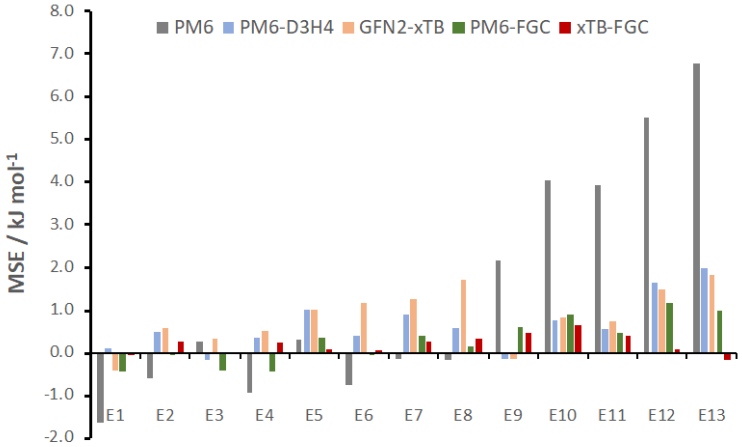


Fig. 4. Mean signed errors obtained with the PM6-FGC and xTB-FGC methods for the Donchev's complexes included in the alkene training set, and comparison with the results calculated with the PM6, PM6-D3H4 and GFN2-xTB approaches.

So far, we have analyzed the performance of our corrections by using the MAE and MSE statistical estimators. It is also interesting to investigate how well the PM6-FGC and xTB-FGC methods describe the interaction energy as a function of the intermolecular distance. To this end, we evaluated the intermolecular potential energy curves corresponding to the orientations obtained by quantum mechanical optimizations in the Donchev's database,[28] where these curves are denoted as "Dimer scans based on QM optimization". For a given complex, the IPEC is constructed by a rigid scan starting from the MP2 optimized structure and changing the intermolecular distance along an axis defined as the line connecting weighted atomic centers of the two molecules. The atom weight was C/R^6, where R is the distance to the nearest atom of the other molecule, and C was set to 1.0 for heavy atoms and to 0.1 for hydrogens.

In our training set, we have 21 IPECs of this kind for alkanes and ten for alkenes. The IPECs for the alkane complexes are depicted in Figs. S2 - S5 of the supplementary material. Here we show in Fig. 5 the IPECs for six selected alkane complexes, those for

which we observed the most important deviations from the reference CCSD(T)/CBS curves. In these graphs, the quantity R is a relative distance (in Å). Specifically, it is the difference between the intermolecular distance define in the Donchev's database and the distance at the optimized structure; therefore, the value $R = 0$ corresponds to the optimized structure. Notice that the minima may not correspond to the optimized structure because the curves are evaluated at the CCSD(T)/CBS level. For comparison, we also include the IPECs evaluated with PM6, PM6-D3H4, and GFN2-xTB. Expectedly, the PM6 curves show a strong underestimation of the interaction strength in the regions of the minima. The GFN2-xTB results exhibit some degree of underestimation of the interaction strength in the well regions, and more repulsion in the short-range regions compared to the reference. The PM6-D3H4 method works quite well in general. However, for many systems the repulsive region is somewhat displaced relative to the reference, which slightly affects the location and well depth of the minima. This occurs, for example, for the butane dimer, propane-butane, and butane-pentane complexes, shown in Plots a-c of Fig. 5 (see also Figs. S2 and S3). The PM6-FGC and xTB-FGC methods do not exhibit this disagreement in the repulsive regions. In general, they reproduce the reference curves quite well, although in a few cases they show some deviations in the vicinity of the minima. For PM6-FGC, the worst disagreement occurs for the complexes of methane with butane (Fig. 5e) and with cyclohexane (Fig. 5f). In both cases, our method overestimates somewhat the well depths of these curves. The PM6-D3H4 also overestimates these well depths, but slightly less than the PM6-FGC method does. The xTB-FGC method overestimates the well depth of the neopentane dimer, as does the PM6-D3H4 method (Fig. 5d). For a few other systems involving cyclohexane (see Fig. S4), the xTB-FGC also overestimates the well depths of the IPECs, but only slightly.

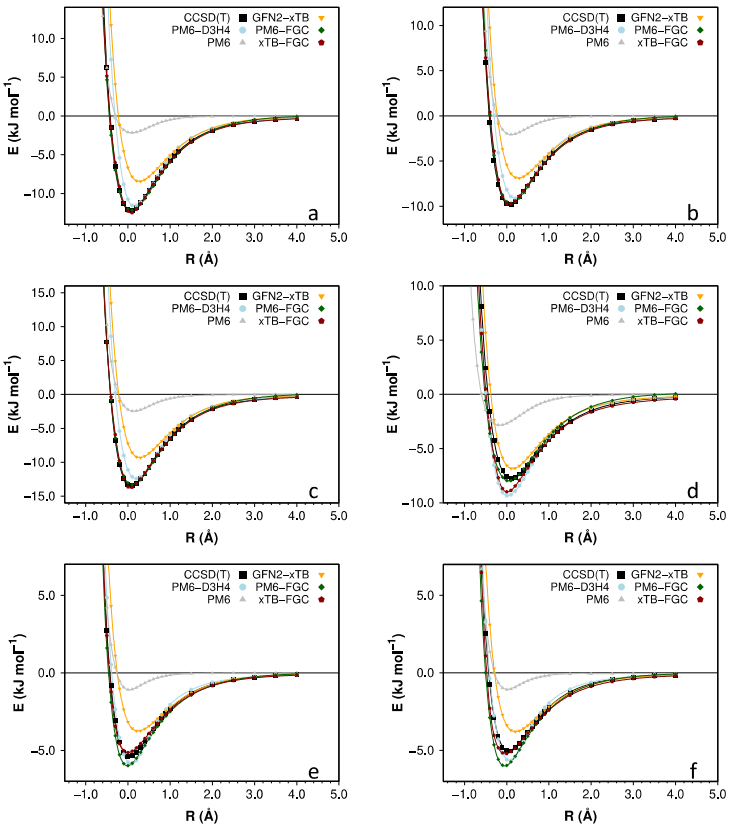


Fig. 5. Comparison of IPECs calculated for (a) butane dimer, (b) propane-butane complex, (c) butane-pentane complex, (d) neopentane dimer, (e) methane-butane complex, and (f) methane-cyclohexane complex.

The IPECs determined for the alkene complexes are displayed in Figs. S5 and S6 of the supplementary material. Here, in Fig. 6, we show the curves for six selected complexes. In general, the agreement is better than for alkanes, but the GFN2-xTB method still exhibits low accuracy. However, for the ethylene dimer, only the xTB-FGC curve agrees quite well with the CCSD(T)/CBS curve; the remaining methods clearly underestimate the well depth. The xTB-FGC method, however, slightly underestimates the strength of the interaction energy in the vicinity of the minima of the propene dimer (Fig. 6a) and the ethylene-isobutene complex (Fig. S7c).

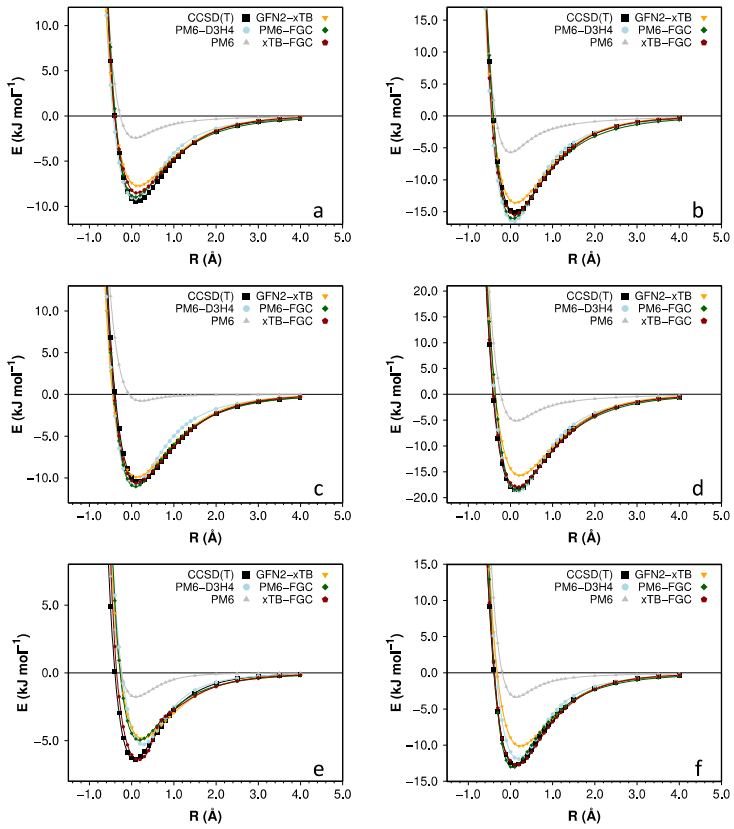


Fig. 6. Comparison of IPECs calculated for (a) propene dimer, (b) isobutene dimer, (c) 2-butene dimer, (d) 2-methylbut-2-ene dimer, (e) ethylene dimer, and (f) ethylene-tetramethylethylene complex.

The results obtained for the medium-sized complexes included in the training set ($C_{18}H_{38}$-1 to $C_{18}H_{38}$-5 and $C_{18}H_{32}$-1 to $C_{18}H_{32}$-5) are listed in Table SI of the supplementary material. As can be seen, the interaction energies calculated with the PM6-FGC and xTB-FGC methods agree very well with the DLPNO-CCSD(T)/CBS data. The largest deviations are only 1.7 kJ mol^{-1} (PM6-FGC for $C_{18}H_{38}$-5) and 1.3 kJ mol^{-1} (xTB-FGC for $C_{18}H_{32}$-5). Since these systems were included in the training set, this agreement essentially indicates that the fits were good. The performance of the PM6-D3H4 and GFN2-xTB methods is discussed in the next section.

C. Validation

So far, we have analyzed the performance of the PM6-FGC and xTB-FGC methods on the systems included in the training set. In other words, we assessed the quality of the fittings. Therefore, it is not a surprise the good agreement observed, in general, between the FGC methods and the reference data. As already mentioned, to validate our corrections, we chose two medium-size homodimers and two large-size homodimers, composed by alkane monomers of 56 and 92 atoms, and alkene monomers of 50 and 80 atoms (see Fig. S1 of the supplementary material).

The interaction energies calculated for the medium-size complexes are listed in Table SI of the supplementary material. As mentioned previously, the five most stable alkane and alkene homodimers were included in the training set. Here, we focus the attention on the complexes not included in the training set. In general, the PM6-FGC and xTB-FGC results agree very well with the reference data. To visualize the agreement more easily, we display in Fig. 7 the MAEs and MSEs calculated for our methods, as well as for PM6-D3H4, GFN2-xTB, and the DFT methods mentioned in the Section on computational details. The values of these statistical estimators, as well as the upper and lower bounds of the error intervals are collected in Table SIII of the supplementary material.

As can be seen from Fig. 7, for the alkanes, the performance of our FGC methods is remarkable. The calculated MAEs are 0.5 kJ mol^{-1} for both methods, and the MSEs are 0.3 and 0.4 kJ mol^{-1} for PM6-FGC and xTB-FGC, respectively. For these methods, the largest deviations from the reference are 1.6 and 1.2 kJ mol^{-1}, respectively. The PM6-D3H4 approach also affords chemical accuracy (i.e., within 4 kJ mol^{-1}), although the

results are not as good as those provided by the FGC methods. Specifically, the calculated MAE and MSE are 2.8 and –2.8 kJ mol^{-1}, respectively, and the largest deviation from the reference is –4.2 kJ mol^{-1}. The PM6-D3H4 method, therefore, displays a slight bias towards overestimation of the strength of the (attractive) interaction energy. This trend appears to be contrary to that found, for this method, for the small systems included in the training set. The reason for this behaviour is that, for the small systems, the resulting MSE is dominated by the contributions from the repulsive regions, wherein the PM6-D3H4 method overestimates the repulsion interaction in comparison with the benchmark. The results indicate that, in general, the GFN2-xTB method does not provide good accuracy for alkanes. The MAE calculated for this method is 16.6 kJ mol^{-1} (same value for the MSE), which indicates that the GFN2-xTB method clearly underestimates the strength of the attractive noncovalent interaction in alkanes. This also justifies the convenience of including the FGC corrections to improve the accuracy of this method for this type of systems.

For the medium-size alkanes, PM6-FGC and xTB-FGC give results comparable to those obtained at the B3LYP-D3(BJ,ABC)/def2-TZVPP level. The other functionals afford slightly less accurate results, when compared with the DLPNO-CCSD(T)/CBS reference. Interestingly, the r^{2}SCAN-3c composite method provides chemical accuracy for this type of systems. This corroborates the results of Grimme and co-workers, who recommend this method because of its efficiency and robustness.[43]

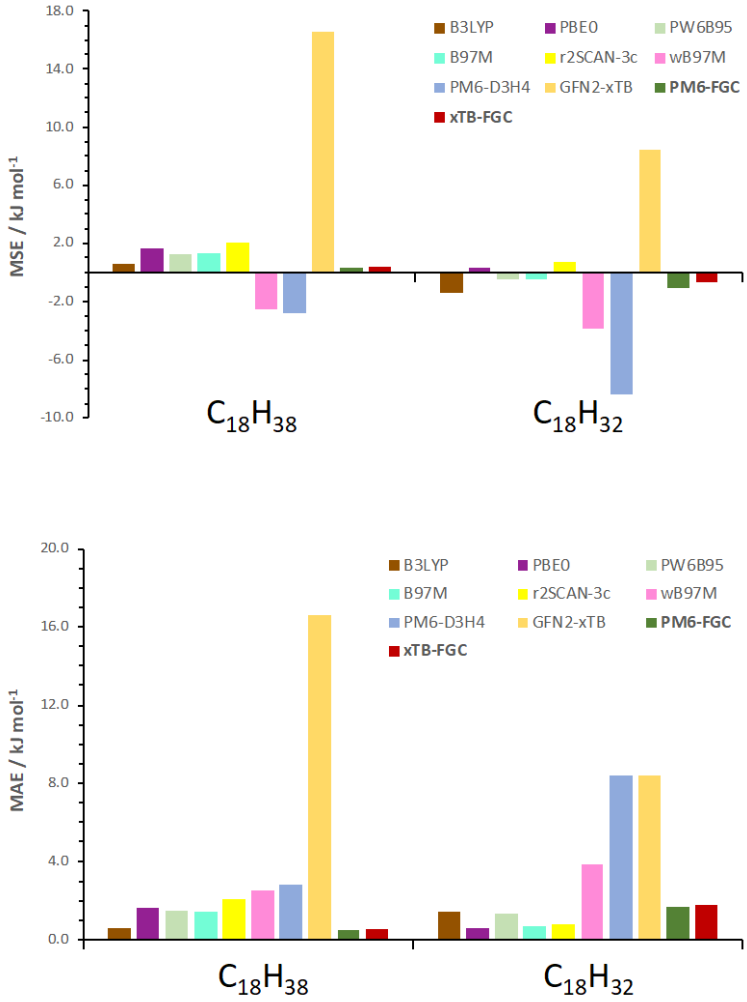


Fig. 7. Mean signed errors (MSEs) and mean absolute errors (MAEs) obtained with the PM6-FGC and xTB-FGC methods for the medium-size complexes (not included in the training set), and comparison with the results calculated with PM6-D3H4, GFN2-xTB, and some DFT methods described in the Computational Details.

For the medium-size alkenes, PM6-FGC and xTB-FGC also provide good accuracy. The calculated MAEs are 1.7 and 1.8 kJ mol^{-1}, respectively, and the MSEs are –1.1 and –0.7 kJ mol^{-1}, respectively. These results are comparable to those obtained by the B3LYP-D3(BJ,ABC) calculations. The best results, in comparison with the reference data, are obtained by B97M-D3(BJ,ABC), PBE0-D3(BJ,ABC), and r^{2}SCAN-3c. Again, it is remarkable the accuracy provided by the r^{2}SCAN-3c. The results obtained with the PM6-D3H4 are not as good as for alkanes. For this method, the MAE and MSE values

are 8.4 and −8.4 kJ mol^{-1}, respectively, which indicates overestimation of the strength of the noncovalent interaction, following the trend found for the medium-size alkanes. For the GFN2-xTB method, the results improve with respect to those obtained for alkanes. Actually, the corresponding MAE is similar to that calculated for PM6-D3H4 (8.4 kJ mol^{-1}), and the MSE has the same value, pointing out that the GFN2-xTB method underestimates the strength of the attractive noncovalent interaction.

For the validation, it is important to investigate how PM6-FGC and xTB-FGC work for rather large systems. To this end, we considered the $C_{30}H_{62}$ and $C_{30}H_{50}$ monomers (see Fig. S1) to form an alkane homodimer of 184 atoms in all, and an alkene homodimer of 160 atoms in all, respectively. For each dimer, we selected ten different complexes obtained by random samplings of the corresponding conformational spaces, as explained above. Due to the size of these complexes, we did not perform DLPNO-CCSD(T)/CBS calculations. Instead, the comparison is made with the results of the DFT methods specified previously. Therefore, this is not a rigorous validation, but it serves to analyze whether the above trends are maintained.

The interaction energies for these large complexes are listed in Table SIV of the supplementary material. To analyze the trends more easily, we have considered the B3LYP-D3(BJ,ABC) results as a reference, and we calculated the MAEs and MSEs statistical estimators, as well as the upper and lower bounds of the error intervals. These data are collected in Table SV of the supplementary material. For the alkane complexes, the PM6-FGC and xTB-FGC methods show small MAE and MSE values, thus following the results obtained for the medium-size complexes. Specifically, the MAEs calculated for PM6-FGC and xTB-FGC are 2.1 and 1.2 kJ mol^{-1}, respectively, and the MSEs are −1.4 and −0.1 kJ mol^{-1}, respectively. The largest deviations from the B3LYP-D3 data are −5.3 and 2.8 kJ mol^{-1}, respectively. Therefore, considering that, for the medium-size

alkanes, B3LYP-D3 and our FGC methods give very good results in comparison with the DLPNO-CCSD(T)/CBS reference, we conclude that the PM6-FGC and xTB-FGC methods provide good accuracy for alkanes in general. As for the medium-size alkanes, the strength of the interaction is overestimated by PM6-D3H4 (MSE = −7.3 kJ mol^{-1}) and underestimated by the GFN2-xTB method (37.7 kJ mol^{-1}). Notice that the results of the PM6-D3H4 method are very similar to those of the wB97M-D3 method, therefore resembling the trend found for the medium-size alkanes (see Table SIII).

We found more deviations for the large-size alkenes. The PM6-FGC method has a MSE value of = –4.3 kJ mol^{-1}, similar to that calculated for the wB97M-D3 method (–4.8 kJ mol^{-1}), indicating overestimation of the strength of the interaction. The MSE value of the PM6-D3H4 method (–16.8 kJ mol^{-1}) indicates a larger deviation, also overestimating the strength of the attractive noncovalent interaction. The PBE0-D3, PW6B95-D3, and B97M-D3 methods give MSE values of 5.2, 4.8, and 2.9 kJ mol^{-1}, respectively, thus underestimating somewhat the strength of the interaction. The MSE calculated with the r^{2}SCAN-3c composite method is 4.6 kJ mol^{-1}, similar to the above values. These values follow the same trends found for the medium-size alkenes, if we consider the B3LYP-D3 results as reference, although the differences are somewhat larger. This is a consequence of the fact that, to a large extent, interaction energies scale with system size. The MSE obtained with the xTB-FGC method is 7.1 kJ mol^{-1}, close to most of the DFT values and indicating underestimation of the strength of the attractive interaction in comparison with the B3LYP-D3 reference. For xTB-FGC, the largest deviation with respect to the reference is 10.2 kJ mol^{-1}. Nevertheless, for these large-size alkenes, the interaction energies calculated with this method are quite close to those predicted by the DFT methods. The MSE calculated for the GFN2-xTB method is 26.2 kJ mol^{-1}, which shows a clear underestimation of the strength of the noncovalent

interaction. This is the same trend as that found for the alkanes and justifies the need to include corrections to improve its accuracy.

IV. CONCLUSIONS

We have developed pairwise analytical corrections to improve the accuracy of the PM6 and GFN2-xTB methods for the calculation of noncovalent interactions in alkanes and alkenes. The key idea of our method is to use atom-pair corrections that depend on the nature of the functional groups involved in the interaction. Therefore, to name our method we used the acronym FGC, from functional group corrections. The functional form of each atom-pair correction term is very simple, since it only contains a decaying exponential and a term proportional to r^{-D}, thus resembling a generic pairwise interaction potential. The correction should be viewed as a global correction that takes account of deficiencies of the SQM methods in describing not only dispersion but also electrostatics, polarization, and exchange-repulsion terms.

For the parameterizations, we used interaction energies taken from the database reported by Donchev et al.,[28] as well as those evaluated for five medium-size alkane and five medium-size alkene structures. This set of additional structures was necessary to derive accurate corrections for large systems. For validation, we employed other structures of these medium-size complexes and a series of different conformations of large alkanes (184 atoms in all) and alkenes (160 atoms in all).

The quality of the fits is reflected by the good agreement found between the interaction energies calculated with the PM6-FGC and xTB-FGC methods and the CCSD(T)/CBS reference data for the complexes included in the training set. The results obtained for the systems used for the validation of our corrections are also in good

agreement with the benchmark data. In general, the errors obtained with the PM6-FGC and xTB-FGC methods are within or close to chemical accuracy. We may expect good performance of our corrections for applications to alkanes and alkenes similar to those investigated here. However, it is important to keep in mind that, for systems with structural features very different from those included in the training set (e.g., aromatic systems), the application of these corrected methods may lead to significant errors.

We plan to extend the method to other functional groups relevant to biological compounds and implement the corrections in the MOPAC[52] and xtb[53] programs. A Python script to calculate PM6-FGC corrections is available from the authors upon request.

Acknowledgments

We thank the Galician Supercomputer Center (CESGA) for the use of their computational facilities. This research was funded by Ministerio de Ciencia e Innovación, grants number PID2019-107307RB-100 and PID2020-117605GB-100, and Xunta de Galicia, grant number ED431C 2021/40. We are also thankful for the financial support of the EU Doctoral Network PHYMOL 101073474 (project call reference HORIZON-MSCA-2021-DN-01) and of the COST Action CA21101 "Confined molecular systems: from a new generation of materials to the stars" (COSY) supported by COST (European Cooperation in Science and Technology).

REFERENCES

[1] D. Skouteris, D. E. Manolopoulos, W. Bian, H.-J. Werner, L.-H. Lai, and K. Liu, Science **286** (1999) 1713-1716.
[2] A. S. Christensen, T. Kubař, Q. Cui, and M. Elstner, Chem. Rev. **116** (2016) 5301-5337.
[3] T. Husch, A. C. Vaucher, and M. Reiher, Int. J. Quantum Chem. **118** (2018) e25799.
[4] W. Thiel, WIREs Computational Molecular Science **4** (2014) 145-157.
[5] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, J. Am. Chem. Soc. **115** (1993) 5348-5348.
[6] J. J. P. Stewart, J. Mol. Model. **13** (2007) 1173-1213.
[7] J. J. P. Stewart, J. Mol. Model. **19** (2013) 1-32.
[8] M. Kolb, and W. Thiel, J. Comp. Chem. **14** (1993) 775-789.
[9] W. Weber, and W. Thiel, Theor. Chem. Acc. **103** (2000) 495-506.
[10] P. O. Dral, X. Wu, L. Spörkel, A. Koslowski, W. Weber, R. Steiger, M. Scholten, and W. Thiel, J. Chem. Theory Comput. **12** (2016) 1082-1096.
[11] M. Korth, and W. Thiel, J. Chem. Theory Comput. **7** (2011) 2929-2936.
[12] J. G. Brandenburg, M. Hochheim, T. Bredow, and S. Grimme, J. Phys. Chem. Lett. **5** (2014) 4275-4284.
[13] C. Bannwarth, S. Ehlert, and S. Grimme, J. Chem. Theory Comput. **15** (2019) 1652-1671.
[14] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, and S. Grimme, WIREs Computational Molecular Science **11** (2021) e1493.
[15] J. Řezáč, and P. Hobza, J. Chem. Theory Comput. **8** (2012) 141-151.
[16] J. Řezáč, and P. Hobza, Chem. Phys. Lett. **506** (2011) 286-289.
[17] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys. **132** (2010) 154104.
[18] J. Řezáč, J. Fanfrlík, D. Salahub, and P. Hobza, J. Chem. Theory Comput. **5** (2009) 1749-1760.
[19] M. Korth, M. Pitoňák, J. Řezáč, and P. Hobza, J. Chem. Theory Comput. **6** (2010) 344-352.
[20] J. Řezáč, K. E. Riley, and P. Hobza, J. Chem. Theory Comput. **7** (2011) 2427-2438.
[21] J. Řezáč, K. E. Riley, and P. Hobza, J. Chem. Theory Comput. **7** (2011) 3466-3470.
[22] J. Řezáč, K. E. Riley, and P. Hobza, J. Chem. Theory Comput. **10** (2014) 1359-1360.
[23] S. Pérez-Tabero, B. Fernández, E. M. Cabaleiro-Lago, E. Martínez-Núñez, and S. A. Vázquez, J. Chem. Theory Comput. **17** (2021) 5556-5567.
[24] M. Ríos-García, B. Fernández, J. Rodríguez-Otero, E. M. Cabaleiro-Lago, and S. A. Vázquez, Molecules **27** (2022) 1678.
[25] S. Grimme, A. Hansen, J. G. Brandenburg, and C. Bannwarth, Chem. Rev. **116** (2016) 5105-5154.
[26] R. Sure, and S. Grimme, J. Chem. Theory Comput. **11** (2015) 3785-3801.
[27] J. Řezáč, P. Jurečkab, K. E. Riley, J. Černý, H. Valdes, K. Pluháčková, K. Berka, T. Řezáč, M. Pitoňák, J. Vondrášek, and P. Hobza, Collect. Czech. Chem. Commun. **73** (2008) 1261-1270.
[28] A. G. Donchev, A. G. Taube, E. Decolvenaere, C. Hargus, R. T. McGibbon, K.-H. Law, B. A. Gregersen, J.-L. Li, K. Palmo, K. Siva, M. Bergdorf, J. L. Klepeis, and D. E. Shaw, Scientific Data **8** (2021) 55.
[29] G. D. Purvis, and R. J. Bartlett, J. Chem. Phys. **76** (1982) 1910-1918.

[30] J. Řezáč, and P. Hobza, J. Chem. Theory Comput. **9** (2013) 2151-2155.
[31] C. Møller, and M. S. Plesset, Phys. Rev. **46** (1934) 618-622.
[32] P. Pracht, F. Bohle, and S. Grimme, Phys. Chem. Chem. Phys. **22** (2020) 7169-7192.
[33] S. Spicher, and S. Grimme, Angew. Chem. Int. Ed. **59** (2020) 15665-15673.
[34] Y. Guo, C. Riplinger, U. Becker, D. G. Liakos, Y. Minenkov, L. Cavallo, and F. Neese, J. Chem. Phys. **148** (2018) 011101.
[35] DunningT.H, Jr., J. Chem. Phys. **90** (1989) 1007-1023.
[36] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen, and A. K. Wilson, Chem. Phys. Lett. **286** (1998) 243-252.
[37] T. Helgaker, W. Klopper, H. Koch, and J. Noga, J. Chem. Phys. **106** (1997) 9639-9646.
[38] F. Weigend, M. Häser, H. Patzelt, and R. Ahlrichs, Chem. Phys. Lett. **294** (1998) 143-152.
[39] F. Weigend, Phys. Chem. Chem. Phys. **4** (2002) 4285-4291.
[40] S. Kossmann, and F. Neese, J. Chem. Theory Comput. **6** (2010) 2325-2338.
[41] A. Najibi, and L. Goerigk, J. Chem. Theory Comput. **14** (2018) 5725-5738.
[42] S. Grimme, S. Ehrlich, and L. Goerigk, J. Comp. Chem. **32** (2011) 1456-1465.
[43] S. Grimme, A. Hansen, S. Ehlert, and J.-M. Mewes, J. Chem. Phys. **154** (2021) 064103.
[44] F. Neese, WIREs Computational Molecular Science **12** (2022) e1606.
[45] J. J. P. Stewart, (Steward Computational Chemistry, web-site: http://OpenMOPAC.net, 2016).
[46] J. M. C. Marques, F. V. Prudente, F. B. Pereira, M. M. Almeida, A. M. Maniero, and C. E. Fellows, J. Phys. B: At. Mol. Opt. Phys. **41** (2008) 085103.
[47] M. M. Almeida, F. V. Prudente, C. E. Fellows, J. M. C. Marques, and F. B. Perieira, J. Phys. B: At. Mol. Opt. Phys. **44** (2011) 225102.
[48] R. Rodríguez-Fernández, F. B. Pereira, J. M. C. Marques, E. Martínez-Núñez, and S. A. Vázquez, Comput. Phys. Commun. **217** (2017) 89-98.
[49] D. Weininger, J. Chem. Inf. Comput. Sci. **28** (1988) 31-36.
[50] L. Fiedler, H. R. Leverentz, S. Nachimuthu, J. Friedrich, and D. G. Truhlar, J. Chem. Theory Comput. **10** (2014) 3129-3139.
[51] P. O. Dral, X. Wu, and W. Thiel, J. Chem. Theory Comput. **15** (2019) 1743-1760.
[52] https://github.com/openmopac/mopac, Molecular Orbital PACkage (MOPAC).
[53] https://github.com/grimme-lab/xtb, Semiempirical extended tight-binding program package xtb.