Efficient model chemistries for peptides. II. Basis set convergence in the B3LYP method.

Pablo ECHENIQUE
Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), and Departamento de Física Teórica, Universidad de Zaragoza, Pedro Cerbuna 12, E-50009 Zaragoza, Spain
E-mail: echenique.p@gmail.com

Gregory A. CHASS
Global Institute Of COmputational Molecular and Materials Science (GIOCOMMS), and School of Chemistry, University of Wales, Bangor, Gwynedd, LL57 2UW United Kingdom, and College of Chemistry, Beijing Normal University, Beijing, 100875, China

PACS: 07.05.Tp; 31.15.Ar; 31.50.Bc; 87.14.Ee; 87.15.Aa; 89.75.-k

Keywords: peptides, quantum chemistry, PES, B3LYP, basis set convergence

Abstract

Small peptides are model molecules for the amino acid residues that are the constituents of proteins. In any bottom-up approach to understand the properties of these macromolecules essential in the functioning of every living being, to correctly describe the conformational behaviour of small peptides constitutes an unavoidable first step. In this work, we present an study of several potential energy surfaces (PESs) of the model dipeptide HCO-L-Ala-NH_2. The PESs are calculated using the B3LYP density-functional theory (DFT) method, with Dunning's basis sets cc-pVDZ, aug-cc-pVDZ, cc-pVTZ, aug-cc-pVTZ, and cc-pVQZ. These calculations, whose cost amounts to approximately 10 years of computer time, allow us to study the basis set convergence of the B3LYP method for this model peptide. Also, we compare the B3LYP PESs to a previous computation at the MP2/6-311++G(2df,2pd) level, in order to assess their accuracy with respect to a higher level reference. All data sets have been analyzed according to a general framework which can be extended to other complex problems and which captures the nearness concept in the space of model chemistries (MCs).

1 Introduction

In any bottom-up attempt to understand the behaviour of protein molecules (in particular, the still elusive protein folding process [1–5]), the characterization of the conformational preferences of short peptides [6–13] constitutes an unavoidable first step. Due to the lower numerical effort required and also to the manageability of their conformational space, the most frequently studied peptides are the shortest ones: the *dipeptides* [14–17], in which a single amino acid residue is capped at both the N- and C-termini with neutral peptide groups. Among them, the most popular choice has been the *alanine* dipeptide [6, 18–34], which, being the simplest chiral residue, shares many similarities with most of the rest of dipeptides for the minimum computational price.

Although classical force fields [35–43] are the only feasible choice for simulating large molecules

at present, they have been reported to yield inaccurate *potential energy surfaces* (PESs) for dipeptides [29, 44–47] and short peptides [6, 48]. Therefore, it is not surprising that they are widely recognized as being unable to correctly describe the intricacies of the whole protein folding process [44, 49–55]. On the other hand, albeit prohibitively demanding in terms of computational resources, ab initio quantum mechanical calculations [56–58] are not only regarded as the correct physical description that in the long run will be the preferred choice to directly tackle proteins (given the exponential growth of computer power and the advances in the search for pleasantly scaling algorithms [59, 60]), but they are also used in small peptides as the reference against which less accurate methods must be compared [6, 29, 44, 45, 47, 61, 62] in order to, for example, parameterize improved generations of additive, classical force fields for polypeptides.

However, despite the sound theoretical basis, in practical quantum chemistry calculations a plethora of approximations must be typically made if one wants to obtain the final results in a reasonable human time. The exact 'recipe' that includes all the assumptions and steps needed to calculate the relevant observables for any molecular system has been termed *model chemistry* (MC) by John Pople. In his own words, a MC is an "approximate but well-defined general and continuous mathematical procedure of simulation" [63].

After assuming that the particles involved move at non-relativistic velocities and that the greater weight of the nuclei allows to perform the Born-Oppenheimer approximation, we are left with the problem of solving the non-relativistic electronic Schrödinger equation [60]. The two starting approximations to its exact solution that a MC must contain are, first, the truncation of the N-electron space (in wavefunction-based methods) or the choice of the functional (in DFT) and, second, the truncation of the one-electron space, via the LCAO scheme (in both cases). The extent up to which the first truncation is carried (or the functional chosen in the case of DFT) is commonly called the *method* and it is denoted by acronyms such as RHF, MP2, B3LYP, CCSD(T), FCI, etc., whereas the second truncation is embodied in the definition of a finite set of atom-centered Gaussian functions termed *basis set* [57, 58, 60, 64, 65], which is also designated by conventional short names, such as 6-31+G(d), TZP or cc-pVTZ(–f). If we denote the method by a capital M and the basis set by a B, the specification of both is conventionally denoted by $L := M/B$ and called a *level of the theory*. Typical examples of this are RHF/3-21G or MP2/cc-pVDZ [56–58].

Note that, apart from these approximations, which are the most commonly used and the only ones that are considered in this work, the MC concept may include a lot of additional features: the heterolevel approximation (explored in a previous work in this series [34]), protocols for extrapolating to the infinite-basis set limit [66–70], additivity assumptions [71–74], extrapolations of the Møller-Plesset series to infinite order [75], removal of the so-called *basis set superposition error* (BSSE) [76–82], etc. The reason behind most of these techniques being the urging need to reduce the computational cost of the calculations.

Now, although general applicability is a requirement that all MCs must satisfy, general accuracy is not mandatory. Actually, the fact is that the different procedures that conform a given MC are typically parameterized and tested in very particular systems, which are often small molecules. Therefore, the validity of the approximations outside that native range of problems must be always

questioned and checked. However, while the approximate computational cost of a given MC for a particular system is rather easy to predict on the basis of simple scaling relations, its expected accuracy on a particular problem could be difficult to predict a priori, specially if we are dealing with large molecules in which interactions in very different energy scales are playing a role. The description of the conformational behaviour of peptides (or, more generally, flexible organic species), via their PESs in terms of the soft internal coordinates, is one of such problems and the one that is treated in this work.

To this end, we first describe, in sec. 2, the computational and theoretical methods used throughout the rest of the document. Then, in sec. 3, we introduce a basic framework that rationalizes the actual process of evaluating the efficiency of any MC for a complex problem. These general ideas are used, in sec. 4, to perform an study of the *density-functional theory* (DFT) B3LYP [83–86] method with the cc-pVDZ, aug-cc-pVDZ, cc-pVTZ, aug-cc-pVTZ, and cc-pVQZ Dunning's basis sets [87, 88]. To this end, we apply these levels of the theory to the calculation the PES of the model dipeptide HCO-L-Ala-NH_2 (see fig. 1), and assess their efficiency by comparison with a reference PES. Finally, in sec. 5, the most important conclusions are briefly summarized.

2 Methods

All ab initio quantum mechanical calculations have been performed using the GAMESS-US program [89, 90] under Linux and on 2.2 GHz PowerPC 970FX machines with 2 GB RAM memory.

The internal coordinates used for the Z-matrix of the HCO-L-Ala-NH_2 dipeptide in the GAMESS-US input files are the *Systematic Approximately Separable Modular Internal Coordinates* (SASMIC) ones introduced in ref. 91. They are presented in table 1 (see also fig. 1 for reference).

All PESs in this study have been discretized into a regular 12×12 grid in the bidimensional space spanned by the Ramachandran angles ϕ and ψ, with both of them ranging from -165^{o} to 165^{o} in steps of 30^{o}. To calculate the PES at a particular level of the theory, we have run constrained energy optimizations at each point of the grid, freezing the two Ramachandran angles ϕ and ψ at the corresponding values. In order to save computational resources, the starting structures were taken, when possible, from PESs previously optimized at a lower level of the theory. All the basis sets used in the study have been taken from the GAMESS-US internally stored library, and spherical Gaussian-type orbitals (GTOs) have been preferred, thus having 5 d-type and 7 f-type functions per shell.

We have computed 5 PESs, using the DFT B3LYP [83–86] method with the cc-pVDZ, aug-cc-pVDZ, cc-pVTZ, aug-cc-pVTZ, and cc-pVQZ Dunning's basis sets [87, 88]. The total cost of these calculations in the machines used is around 10 years of computer time.

Also, let us note that the correcting terms to the PES coming from mass-metric tensors determinants and from the determinant of the Hessian matrix have been recently shown to be relevant for the conformational behaviour of peptides [18]. (The latter may be regarded as a residual entropy

Atom name	Bond length	Bond angle	Dihedral angle
H_1			
C_2	(2,1)		
N_3	(3,2)	(3,2,1)	
O_4	(4,2)	(4,2,1)	(4,2,1,3)
C_5	(5,3)	(5,3,2)	(5,3,2,1)
H_6	(6,3)	(6,3,2)	(6,3,2,5)
C_7	(7,5)	(7,5,3)	ϕ :=**(7,5,3,2)**
C_8	(8,5)	(8,5,3)	(8,5,3,7)
H_9	(9,5)	(9,5,3)	(9,5,3,7)
H_{10}	(10,8)	(10,8,5)	(10,8,5,3)
H_{11}	(11,8)	(11,8,5)	(11,8,5,10)
H_{12}	(12,8)	(12,8,5)	(12,8,5,10)
N_{13}	(13,7)	(13,7,5)	ψ :=**(13,7,5,3)**
O_{14}	(14,7)	(14,7,5)	(14,7,5,13)
H_{15}	(15,13)	(15,13,7)	(15,13,7,5)
H_{16}	(16,13)	(16,13,7)	(16,13,7,15)

Table 1: Internal coordinates in Z-matrix form of the protected dipeptide HCO-L-Ala-NH_2 according to the SASMIC scheme introduced in ref. 91. The numbering of the atoms is that in fig. 1, and the soft Ramachandran angles ϕ and ψ are indicated.

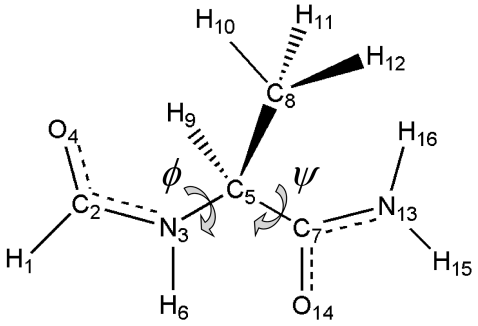

Figure 1: Atom numeration of the protected dipeptide HCO-L-Ala-NH_2 according to the SASMIC scheme introduced in ref. 91. The soft Ramachandran angles ϕ and ψ are also indicated.

arising from the elimination of the hard coordinates from the description.) Although, in this study, we have included none of these terms, the PES calculated here is the greatest part of the effective free energy [18], so that it may be considered as the first ingredient for a further refinement of the study in which the correcting terms are taken into account. The same may be said about another important source of error in the calculation of relatives energies in peptide systems: the already mentioned BSSE [31].

In order to compare the PESs produced by the different MCs, a statistical criterium (distance)

introduced in ref. 92 has been used. Let us recall here that this *distance*, denoted by d_{12}, profits from the complex nature of the problem studied to compare any two different potential energy functions, V_1 and V_2. From a working set of conformations (in this case, the 144 points of each PES), it statistically measures the typical error that one makes in the *energy differences* if V_2 is used instead of the more accurate V_1, admitting a linear rescaling and a shift in the energy reference.

Despite having energy units, the quantity d_{12} approximately presents all properties characteristic of a typical mathematical metric in the space of MCs (hence the word 'distance'), such as the possibility of defining a symmetric version of it and a fulfillment of the triangle inequality (see ref. 92 for the technical details and sec. 3 for more about the importance of these facts). It also presents better properties than other quantities customarily used to perform these comparisons, such as the energy RMSD, the average energy error, etc., and it may be related to the Pearson's correlation coefficient r_{12} by

$$d_{12} = \sqrt{2}\,\sigma_2(1 - r_{12}^2)^{1/2}\,, \tag{1}$$

where σ_2 is the standard deviation of V_2 in the working set.

Moreover, due to its physical meaning, it has been argued in ref. 92 that, if the distance between two different approximations of the energy of the same system is less than RT, one may safely substitute one by the other without altering the relevant dynamical or thermodynamical behaviour. Consequently, we shall present the results in units of RT (at 300$^{\text{o}}$ K, so that $RT \simeq 0.6$ kcal/mol).

Finally, if one assumes that the effective energies compared will be used to construct a polypeptide potential and that it will be designed as simply the sum of mono-residue ones (more complex situations may be found in real problems [93]), then, the number N_{res} of residues up to which one may go keeping the distance d_{12} between the two approximations of the the N-residue potential below RT is [92]

$$N_{\text{res}} = \left(\frac{RT}{d_{12}}\right)^2 . \tag{2}$$

According to the value taken by N_{res} for a comparison between a fixed reference PES, denoted by V_1, and a candidate approximation, denoted by V_2, we shall divide the whole accuracy range in sec. 4 in three regions depending on the accuracy: the *protein region*, corresponding to $0 < d_{12} \leq 0.1RT$, or, equivalently, to $100 \leq N_{\text{res}} < \infty$; the *peptide region*, corresponding to $0.1RT < d_{12} \leq RT$, or $1 \leq N_{\text{res}} < 100$; and, finally, the *inaccurate region*, where $d_{12} > RT$, and even for a dipeptide it is not advisable to use V_2 as an approximation to V_1. Of course, these are only approximate regions based on the general idea that we are not interested on the dipeptides as a final system, but only as a mean to approach protein behaviour from the botton-up. Therefore, not only the error in the dipeptides must be measured, but it must also be estimated how this discrepancy propagates to polypeptide systems.

3 General framework

The general abstract framework behind the investigation presented in this study (and also implicitly behind most of the works found in the literature), may be described as follows:

The objects of study are the *model chemistries* defined by Pople [63] and discussed in the introduction. The MCs under scrutiny are applied to a particular *problem* of interest, which may be thought to be formed by three ingredients: the *physical system*, the *relevant observables* and the *target accuracy*. The MCs are then selected according to their ability to yield numerical values of the relevant observables for the physical system studied within the target accuracy. The concrete numerical values that one wants to approach are those given by the *exact model chemistry* MC_ε, which could be thought to be either the experimental data or the exact solution of the non-relativistic electronic Schrödinger equation [60]. However, the computational effort needed to perform the calculations required by MC_ε is literally infinite, so that, in practice, one is forced to work with a *reference model chemistry* $\mathrm{MC}^{\mathrm{ref}}$, which, albeit different from MC_ε, is thought to be close to it. Finally, the set of MCs that one wants to investigate are compared to $\mathrm{MC}^{\mathrm{ref}}$ and the nearness to it is seen as approximating the nearness to MC_ε.

These comparisons are commonly performed using a numerical quantity $\mathcal{D}$ that is a function of the relevant observables. In order for the intuitive ideas about relative proximity in the $\mathcal{M}$ space to be captured and the above reasoning to be meaningful, this numerical quantity $\mathcal{D}$ must have some of the properties of a mathematical distance. In particular, it is advisable that the *triangle*

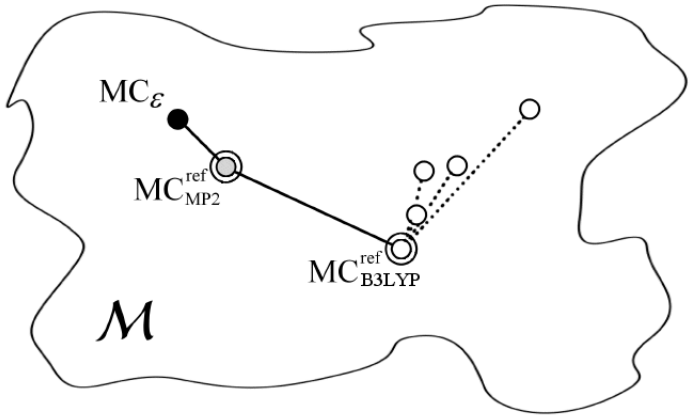

Figure 2: Space $\mathcal{M}$ of all model chemistries. The exact model chemistry MC_ε is shown as a black circle, the MP2 reference MC is shown as a grey-filled circle, and B3LYP MCs as white-filled ones. Both reference PESs are indicated with an additional circle around the points. The situation depicted is (schematically) the one found in this study, assuming that MP2 is a more accurate method than B3LYP to account for the conformational preferences of peptide systems. The positions of the different MCs have no relevance, and only the relative measured distances among them are qualitatively depicted.

inequality is obeyed, so that, for any model chemistry MC, one has that

$$\mathcal{D}(\mathrm{MC}_\varepsilon, \mathrm{MC}) \leq \mathcal{D}(\mathrm{MC}_\varepsilon, \mathrm{MC}^{\mathrm{ref}}) + \mathcal{D}(\mathrm{MC}^{\mathrm{ref}}, \mathrm{MC}) \,, \tag{3a}$$

$$\mathcal{D}(\mathrm{MC}_\varepsilon, \mathrm{MC}) \geq \left|\mathcal{D}(\mathrm{MC}_\varepsilon, \mathrm{MC}^{\mathrm{ref}}) - \mathcal{D}(\mathrm{MC}^{\mathrm{ref}}, \mathrm{MC})\right| \,, \tag{3b}$$

and, assuming that $\mathcal{D}(\mathrm{MC}_\varepsilon, \mathrm{MC}^{\mathrm{ref}})$ is small (and $\mathcal{D}$ is a positive function), we obtain

$$\mathcal{D}(\mathrm{MC}_\varepsilon, \mathrm{MC}) \simeq \mathcal{D}(\mathrm{MC}^{\mathrm{ref}}, \mathrm{MC}) \,, \tag{4}$$

which is the sought result in agreement with the ideas stated at the beginning of this section.

The distance d_{12} introduced in ref. 92 and summarized in the previous section, measured in this case on the conformational energy surfaces (the relevant observable) of the model dipeptide HCO-L-Ala-NH_2 (the physical system), approximately fulfills the triangle inequality and thus captures the *nearness* concept in the space $\mathcal{M}$ of model chemistries.

This space, $\mathcal{M}$, containing all possible MCs, is a rather complex and multidimensional one. For example, two commonly used 'dimensions' which may be thought to parameterize $\mathcal{M}$ are the size of the basis set and the amount of electron correlation in the model (or the quality of the DFT functional used). However, since there are many ways in which the size of a basis set or the electron correlation may be increased and there are additional approximations that can be included in the MC definition (see sec. 1), the 'dimensions' of $\mathcal{M}$ can be considered to be many more than two.

The definition of a distance, such as the one described in the previous lines, for a given problem of interest helps to provide a certain degree of structure into this complex space. In fig. 2 a two-dimensional scheme of the overall situation found in this study is presented.

4 Results

Before starting with the results of the calculations, let us introduce the concept of *efficiency* of a particular MC that shall be used: It is laxly defined as a balance between accuracy (in terms of

MCs	d_{12}/RT [a]	a_{12} [b]	N_{res} [c]	t [d]
B3LYP/aug-cc-pVTZ	0.079	15.2	159.8	79.09%
B3LYP/cc-pVTZ	0.191	21.1	27.4	9.78%
B3LYP/aug-cc-pVDZ	0.172	82.8	33.7	5.27%
B3LYP/cc-pVDZ	1.045	109.4	0.9	1.29%

Table 2: Basis set convergence results for the B3LYP MCs investigated in this work. [a]Distance with the B3LYP/cc-pVQZ reference in units of RT at 300° K. [b]Energy offset with the reference MC in kcal/mol. [c]Maximum number of residues in a polypeptide potential up to which the corresponding MC may correctly approximate the reference (under the assumptions in sec. 2). [d]Required computer time, expressed as a fraction of t_{ref}.

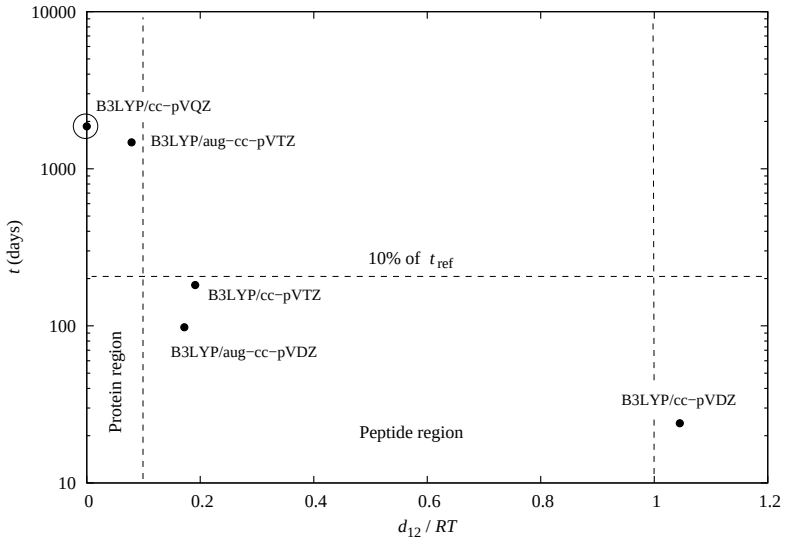

Figure 3: Efficiency plot of all the B3LYP MCs studied. In the x-axis, we show the distance d_{12}, in units of RT at 300^{o} K, between any given MC and the B3LYP/cc-pVQZ reference (indicated by an encircled point), while, in the y-axis, we present the computer time needed to compute the whole 12×12 grid in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2. The different accuracy regions are labeled, and the 10% of the time t_{best} taken by the reference MC is also indicated.

the distance introduced in sec. 2) and computational cost (in terms of computer time). It can be graphically extracted from the *efficiency plots*, where the distance d_{12} between any given MC and a reference one is shown in units of RT in the x-axis, while, in the y-axis, one can find the computer time taken for each MC (see the following pages for two examples). As a general thumb-rule, *we shall consider a MC to be more efficient for approximating the reference when it is placed closer to the origin of coordinates in the efficiency plot*. This approach is intentionally non-rigorous due to the fact that many factors exist that influence the computer time but may vary from one practical calculation to another; such as the algorithms used, the actual details of the computers (frequency of the processor, size of the RAM and cache memories, system bus and disk access velocity, operating system, mathematical libraries, etc.), the starting guesses for the SCF orbitals or the starting structures in geometry optimizations.

Taking all this into account, the only conclusions that shall be drawn in this work about the relative efficiency of the MCs studied are those deduced from strong signals in the plots and, therefore, those that can be extrapolated to future calculations; in other words, *the small details shall be typically neglected*.

In the first part of the study, we compare all B3LYP MCs to the one with the largest basis set, B3LYP/cc-pVQZ (the highest level of the theory calculated for this work, depicted in fig. 4) using the distance introduced in sec. 2. All mentions to the accuracy of any given MC in this part must be understood as relative to this reference. However, it has been reported that MP2 is a superior method to B3LYP to account for the conformational behaviour of peptide systems [94]. Therefore,

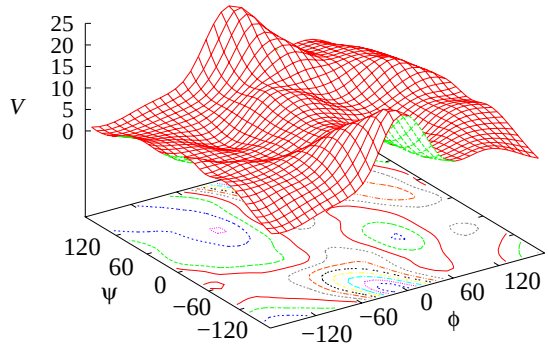

Figure 4: Potential energy surface of the model dipeptide HCO-L-Ala-NH_2 computed at the B3LYP/cc-pVQZ level of the theory. The PES has been originally calculated in a 12×12 discrete grid in the space spanned by the Ramachandran angles ϕ and ψ and later smoothed with bicubic splines for visual convenience. The energy reference has been set to zero. (At this level of the theory, the absolute energy of the minimum point in the 12×12 grid, located at $(-75^o, 75^o)$, is −417.199231353 hartree).

the absolute accuracy of the B3LYP MCs calculated here is probably closer to the relative accuracy with respect to the MP2/6-311++G(2df,2pd) reference in what follows. In this spirit, this part of the study should be regarded as an investigation of the convergence to *the infinite basis set B3LYP limit*, for which the best B3LYP MC here is probably a good approximation.

The results are depicted in fig. 3, and in table 2. We can extract several conclusions from them:

- Regarding the convergence to the infinite basis set limit, we observe that only the most expensive MC, B3LYP/aug-cc-pVTZ, correctly approximates the reference for peptides of more than 100 residues. On the other hand, for only 5.27% of the computer time t_{ref} taken by the reference MC, we can use B3LYP/aug-cc-pVDZ, which correctly approximates it up to 30-residue peptides. Finally, the MC with the smallest basis set, B3LYP/cc-pVDZ cannot properly replace the reference even in dipeptides.

- In ref. [34], using Pople's basis sets [95–102], we saw that "the general rule that is sometimes assumed when performing quantum chemical calculations, which states that 'the more expensive, the more accurate', is rather coarse-grained and relevant deviations from it may be found." We recognized that "One may argue that this observation is due to the unsystematic way in which Pople basis sets can be enlarged and that the correlation between accuracy and cost will be much higher if, for example, only Dunning basis sets are used.", which is definitely observed in fig. 3, but we argued that this was something to be expected, since "there are two few Dunning basis sets below a reasonable upper bound on the number of elements to see anything but a line in the efficiency plot". In the results presented in this work, we can see that, even if the correlation between accuracy and cost is higher in the case of Dunning's basis sets than in the case of Pople's, due to the smaller number of the former,

MCs	d_{12}/RT [a]	a_{12} [b]	N_{res} [c]	t [d]
B3LYP/cc-pVQZ	1.008	-457.2	0.98	1861
B3LYP/aug-cc-pVTZ	1.029	-442.0	0.94	1472
B3LYP/cc-pVTZ	1.058	-436.1	0.89	182
B3LYP/aug-cc-pVDZ	1.006	-374.4	0.99	98
B3LYP/cc-pVDZ	1.533	-347.8	0.43	24

Table 3: Comparison of all the B3LYP MCs investigated in this work with the MP2/6-311++G(2df,2pd) in ref. 34. [a]Distance with the MP2/6-311++G(2df,2pd) reference in units of RT at 300° K. [b]Energy offset with the reference MC in kcal/mol. [c]Maximum number of residues in a polypeptide potential up to which the corresponding MC may correctly approximate the reference (under the assumptions in sec. 2). [d]Computer time needed for the calculation of the whole PES, in days.

we can still observe that the thumb-rule 'the more expensive, the more accurate' breaks also in this case, since the B3LYP/aug-cc-pVDZ MC is, at the same time, more accurate and less costly than B3LYP/cc-pVTZ. In general, this idea applies to all the approximations that a MC may contain (see the introduction for a partial list), and justifies the systematic search for the most efficient combination of them for a given problem. This work is our second step (ref. [34] is the first one) in that path for the particular case of the conformational behaviour of peptide systems.

- The observation in the previous point also suggests that it may be efficient to include diffuse functions (the 'aug-' in aug-cc-pVDZ) in the basis set for this type of problems.

- The error of the studied MCs regarding the differences of energy (as measured by d_{12}) is much smaller than the error in the absolute energies (as measured by a_{12}), suggesting that the largest part of the discrepancy must be a systematic one.

In the second part of the study, we assess the absolute accuracy of the B3LYP MCs by comparing them to the (as far as we are aware) highest homolevel in the literature, the MP2/6-311++G(2df,2pd) PES in ref. [34]. If one assumes that this level of the theory may be close enough to the exact result for the given problem at hand, then this comparison measures the 'absolute' accuracy of the B3LYP MCs, and not only their relative accuracy with respect to the B3LYP infinite basis set limit, as we did in the previous part. This is the fundamental difference between figs. 3 and 5.

The results of this part of the study are depicted in fig. 5, and in table 3. We can extract several conclusions from them:

- All B3LYP MCs, including the largest one, B3LYP/cc-pVQZ, lie in the inaccurate region of the efficiency plot in fig. 5, meaning that they cannot be reliably used to approximate the MP2/6-311++G(2df,2pd) reference even in the smallest dipeptides.

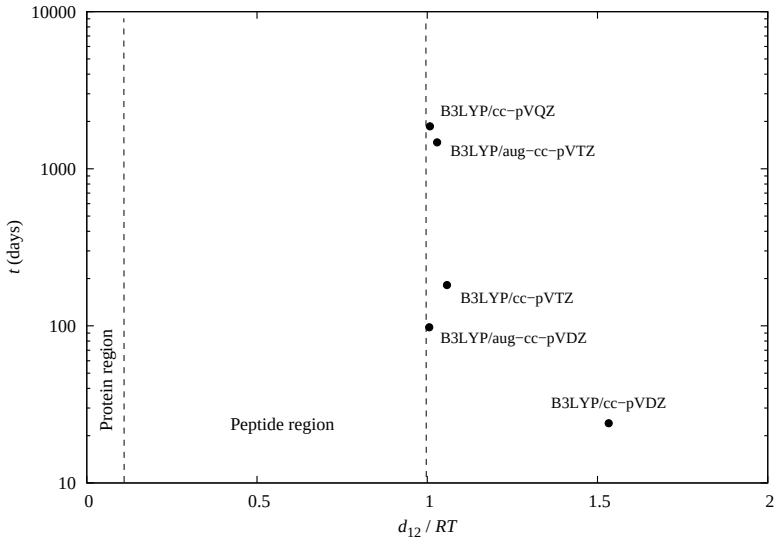

Figure 5: Efficiency plot of all the B3LYP MCs studied. In the x-axis, we show the distance d_{12}, in units of RT at 300^o K, between any given MC and the MP2/6-311++G(2df,2pd) reference calculated in ref. 34, while, in the y-axis, we present the computer time needed to compute the whole 12×12 grid in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2. The different accuracy regions are labeled

- Related with the observations in the previous part of the study, we see that there is no point, if one is worried about absolute accuracy, in going beyond the aug-cc-pVDZ basis set in B3LYP.
- The B3LYP/cc-pVDZ MC again performs significantly worse than the rest, agreeing with the results in the previous part of the study, and suggesting that cc-pVDZ may be a too small basis set for the problem tackled here.
- Again, the error of the MCs in the differences of energy (as measured by d_{12}) is much smaller than the error in the absolute energies (as measured by a_{12}).

5 Conclusions

In this study, we have investigated 5 PESs of the model dipeptide HCO-L-Ala-NH_2, calculated with the B3LYP method, and the cc-pVDZ, aug-cc-pVDZ, cc-pVTZ, aug-cc-pVTZ, and cc-pVQZ Dunning's basis sets. We have first assessed the convergence of the B3LYP MCs to the infinite basis set limit, and then we have evaluated their absolute accuracy by comparing them to the (as far as we are aware) highest homolevel in the literature, the MP2/6-311++G(2df,2pd) PES in ref. [34]. All the comparisons have been performed according to a general framework which is extensible to further studies, and using a distance between the different PESs that correctly captures

the nearness concept in the space of MCs. The calculations performed here have taken around 10 years of computer time.

The main conclusions of the study are the following:

- The complexity of the problem (the conformational behaviour of peptides) renders the correlation between accuracy and computational cost of the different quantum mechanical algorithms imperfect. This ultimately justifies the need for systematic studies, such as the one presented here, in which the most efficient MCs are sought for the particular problem of interest.
- Assuming that the MP2/6-311++G(2df,2pd) level of the theory is closer to the exact solution of the non-relativistic electronic Schrödinger equation than B3LYP/cc-pVQZ, B3LYP is not a reliable method to study the conformational behaviour of peptides. Even if, as we emphasize at the end of this section, it may be dangerous to state that a method that performs well in the particular model of an alanine residue studied here will also be recommendable for longer and more complex peptides, we can clearly *reject* any method that already fails in HCO-L-Ala-NH_2.
- If B3LYP is still needed to be used, due to, for example, computational constraints, aug-cc-pVDZ represents a good compromise between accuracy and cost.
- The error of the studied MCs regarding the differences of energy (as measured by d_{12}) is much smaller than the error in the absolute energies (as measured by a_{12}), suggesting that the largest part of the discrepancy must be a systematic one.

Finally, let us stress again that the investigation performed here have used one of the simplest dipeptides. The fact that we have treated it as an isolated system, the small size of its side chain and also its aliphatic character, all play a role in the results obtained. Hence, for bulkier residues included in polypeptides, and, specially for those that contain aromatic groups, those that are charged or may participate in hydrogen-bonds, the methods that have proved to be efficient here must be re-tested and the conclusions drawn about the B3LYP convergence to the infinite basis set limit, as well as those regarding the comparison between B3LYP and MP2, should be re-evaluated.

Acknowledgments

The numerical calculations in this work have been performed thanks to a computer time grant at the Zaragoza node (Caesaraugusta) of the Spanish Supercomputing Network (RES). We thank all the support staff there, for the efficiency at solving the problems encountered. We also thank J. L. Alonso for illuminating discussions.

This work has been supported by the research projects DGA (Aragón Government, Spain) E24/3 and MEC (Spain) FIS2006-12781-C02-01. P. Echenique is supported by a MEC (Spain) postdoctoral contract.

References

[1] C. B. Anfinsen, *Science* **181** (1973) 223–230.

[2] V. Daggett and A. R. Fersht, *Trends Biochem. Sci.* **28** (2003) 18–25.

[3] P. Echenique, *Contemp. Phys.* **48** (2007) 81–108.

[4] B. Honig, *J. Mol. Biol.* **293** (1999) 283–293.

[5] J. Skolnick, *Proc. Natl. Acad. Sci. USA* **102** (2005) 2265–2266.

[6] M. Beachy, D. Chasman, R. Murphy, T. Halgren, and R. Friesner, *J. Am. Chem. Soc.* **119** (1997) 5908–5920.

[7] R. A. DiStasio Jr., Y. Jung, and M. Head-Gordon, *J. Chem. Theory Comput.* **1** (2005) 862–876.

[8] M. Elstner, K. J. Jalkanen, M. Knapp-Mohammady, T. Frauenheim, and S. Suhai, *Chem. Phys.* **256** (2001) 15–27.

[9] R. Hegger, A. Altis, P. Nguyen, and G. Stock, *Phys. Rev. Lett.* **98** (2007) 028102.

[10] A. Perczel, I. Jákli, and I. G. Csizmadia, *Chem. Eur. J.* **9** (2003) 5332–5342.

[11] A. Perczel, P. Hudáky, A. K. Füzéry, and I. G. Csizmadia, *J. Comput. Chem.* **25** (2004) 1084–1100.

[12] D. Toroz and T. van Mourik, *Mol. Phys.* **104** (2006) 559–570.

[13] H. Zhong and H. A. Carlson, *J. Chem. Theory Comput.* **2** (2006) 342–353.

[14] A. G. Császár and A. Perczel, *Prog. Biophys. Mol. Biol.* **71** (1999) 243–309.

[15] P. Hudáky, I. Jákli, A. G. Császár, and A. Perczel, *J. Comput. Chem.* **22** (2001) 732–751.

[16] J. C. P. Koo, G. A. Chass, A. Perczel, Ö. Farkas, L. L. Torday, A. Varro, J. G. Papp, and I. G. Csizmadia, *J. Phys. Chem. A* **106** (2002) 6999–7009.

[17] A. Láng, I. G. Csizmadia, and A. Perczel, *PROTEINS: Struct. Funct. Bioinf.* **58** (2005) 571–588.

[18] P. Echenique, I. Calvo, and J. L. Alonso, *J. Comput. Chem.* **27** (2006) 1748–1755.

[19] M. Elstner, K. J. Jalkanen, M. Knapp-Mohammady, and S. Suhai, *Chem. Phys.* **263** (2001) 203–219.

[20] G. Endrédi, A. Perczel, O. Farkas, M. A. McAllister, G. I. Csonka, J. Ladik, and I. G. Csizmadia, *J. Mol. Struct. (Theochem)* **391** (1997) 15-26.

[21] R. F. Frey, J. Coffin, S. Q. Newton, M. Ramek, V. K. W. Cheng, F. A. Momany, and L. Schäfer, *J. Am. Chem. Soc.* **114** (1992) 5369–5377.

[22] I. R. Gould, W. D. Cornell, and I. H. Hillier, *J. Am. Chem. Soc.* **116** (1994) 9250–9256.

[23] T. Head-Gordon, M. Head-Gordon, M. J. Frisch, C. Brooks III, and J. Pople, *Intl. J. Quant. Chem.* **16** (1989) 311-322.

[24] T. Head-Gordon, M. Head-Gordon, M. J. Frisch, C. L. Brooks III, and J. A. Pople, *J. Am. Chem. Soc.* **113** (1991) 5989–5997.

[25] M. Iwaoka, M. Okada, and S. Tomoda, *J. Mol. Struct. (Theochem)* **586** (2002) 111–124.

[26] M. Mezei, P. K. Mehrotra, and D. L. Beveridge, *J. Am. Chem. Soc.* **107** (1985) 2239–2245.

[27] A. Perczel, J. G. Angyán, M. Kajtar, W. Viviani, J.-L. Rivail, J.-F. Marcoccia, and I. G. Csizmadia, *J. Am. Chem. Soc.* **113** (1991) 6256-6265.

[28] A. Perczel, Ö. Farkas, I. Jákli, I. A. Topol, and I. G. Csizmadia, *J. Comput. Chem.* **24** (2003) 1026–1042.

[29] A. M. Rodríguez, H. A. Baldoni, F. Suvire, R. Nieto Vázquez, G. Zamarbide, R. D. Enriz, Ö. Farkas, A. Perczel, M. A. McAllister, L. L. Torday, J. G. Papp, and I. G. Csizmadia, *J. Mol. Struct. (Theochem)* **455** (1998) 275–301.

[30] P. J. Rossky and M. Karplus, *J. Am. Chem. Soc.* **101** (1979) 1913.

[31] R. Vargas, J. Garza, B. P. Hay, and D. A. Dixon, *J. Phys. Chem. A* **106** (2002) 3213–3218.

[32] Z.-X. Wang and Y. Duan, *J. Comput. Chem.* **25** (2004) 1699–1716.

[33] C.-H. Yu, M. A. Norman, L. Schäfer, M. Ramek, A. Peeters, and C. van Alsenoy, *J. Mol. Struct.* **567–568** (2001) 361–374.

[34] P. Echenique and J. L. Alonso, *J. Comput. Chem.* **29** (2008) 1408–1422.

[35] J. W. Ponder and D. A. Case, *Adv. Prot. Chem.* **66** (2003) 27–85.

[36] A. D. MacKerell Jr., B. Brooks, C. L. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus, *CHARMM: The energy function and its parameterization with an overview of the program*, in *The Encyclopedia of Computational Chemistry*, edited by P. v. R. Schleyer, P. R. Schreiner, N. L. Allinger, T. Clark, J. Gasteiger, P. Kollman, and H. F. Schaefer III, pp. 217–277, (John Wiley & Sons, Chichester, 1998).

[37] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.* **4** (1983) 187–217.

[38] W. F. Van Gunsteren and M. Karplus, *Macromolecules* **15** (1982) 1528–1544.

[39] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, J. Merz, K. M., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *J. Am. Chem. Soc.* **117** (1995) 5179–5197.

[40] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. R. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman, *Comp. Phys. Commun.* **91** (1995) 1–41.

[41] W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.* **110** (1988) 1657–1666.

[42] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, *J. Am. Chem. Soc.* **118** (1996) 11225–11236.

[43] T. A. Halgren, *J. Comput. Chem.* **17** (1996) 490–519.

[44] A. R. MacKerell Jr., M. Feig, and C. L. Brooks III, *J. Comput. Chem.* **25** (2004) 1400–1415.

[45] A. R. MacKerell Jr., M. Feig, and C. L. Brooks III, *J. Am. Chem. Soc.* **126** (2004) 698–699.

[46] Y. K. Kang and H. S. Park, *J. Mol. Struct. (Theochem)* **593** (2002) 55–64.

[47] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, *J. Phys. Chem. B* **105** (2001) 6476–6487.

[48] T. Wang and R. Wade, *J. Chem. Theory Comput.* **2** (2006) 140–148.

[49] C. D. Snow, E. J. Sorin, Y. M. Rhee, and V. S. Pande, *Annu. Rev. Biophys. Biomol. Struct.* **34** (2005) 43–69.

[50] O. Schueler-Furman, C. Wang, P. Bradley, K. Misura, and D. Baker, *Science* **310** (2005) 638–642.

[51] K. Ginalski, N. V. Grishin, A. Godzik, and L. Rychlewski, *Nucleic Acids Research* **33** (2005) 1874–1891.

[52] A. V. Morozov, T. Kortemme, K. Tsemekhman, and D. Baker, *Proc. Natl. Acad. Sci. USA* **101** (2004) 6946–6951.

[53] C. Gómez-Moreno Calera and J. Sancho Sanz, editors, *Estructura de Proteínas*, (Ariel ciencia, Barcelona, 2003).

[54] M. Karplus and J. A. McCammon, *Nat. Struct. Biol.* **9** (2002) 646–652.

[55] R. Bonneau and D. Baker, *Annu. Rev. Biophys. Biomol. Struct.* **30** (2001) 173–189.

[56] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, (John Wiley & Sons, Chichester, 2nd edition, 2002).

[57] F. Jensen, *Introduction to Computational Chemistry*, (John Wiley & Sons, Chichester, 1998).

[58] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduced to Advanced Electronic Structure Theory*, (Dover Publications, New York, 1996).

[59] Y. Shao, L. F. Molnar, Y. Jung, J. Kussmann, C. Ochsenfeld, S. T. Brown, A. T. B. Gilbert, L. V. Slipchenko, S. V. Levchenko, D. P. Oneill, R. A. Distasio, R. C. Lochan, T. Wang, G. J. O. Beran, N. A. Besley, J. M. Herbert, C. Y. Lin, T. Van Voorhis, S. H. Chien, A. Sodt, R. P. Steele, V. A. Rassolov, P. E. Maslen, P. P. Korambath, R. D. Adamson, B. Austin, J. Baker, E. F. C. Byrd, H. Dachsel, R. J. Doerksen, A. Dreuw, B. D. Dunietz, A. D. Dutoi, T. R. Furlani, S. R. Gwaltney, A. Heyden, S. Hirata, C.-P. Hsu, G. Kedziora, R. Z. Khalliulin, P. Klunzinger, A. M. Lee, M. S. Lee, W. Liang, I. Lotan, N. Nair, B. Peters, E. I. Proynov, P. A. Pieniazek, Y. M. Rhee, J. Ritchie, E. Rosta, D. C. Sherrill, A. C. Simmonett, J. E. Subotnik, L. H. Woodcock, W. Zhang, A. T. Bell, and A. K. Chakraborty, *Phys. Chem. Chem. Phys.* **8** (2006) 3172–3191.

[60] P. Echenique and J. L. Alonso, *Mol. Phys.* **105** (2007) 3057–3098.

[61] P. Maurer, A. Laio, H. W. Hugosson, M. C. Colombo, and U. Rothlisberger, *J. Chem. Theory Comput.* **3** (2007) 628–639.

[62] Y. A. Arnautova, A. Jagielska, and H. A. Scheraga, *J. Phys. Chem. B* **110** (2006) 5025–5044.

[63] J. A. Pople, *Rev. Mod. Phys.* **71** (1999) 1267–1274.

[64] J. M. García de la Vega and B. Miguel, *Basis sets for computational chemistry*, in *Introduction to Advanced Topics of Computational Chemistry*, edited by L. A. Montero, L. A. Díaz, and R. Bader, chapter 3, pp. 41–80, (Editorial de la Universidad de la Habana, 2003).

[65] T. Helgaker and P. R. Taylor, *Gaussian basis sets and molecular integrals*, in *Modern Electronic Structure Theory. Part II*, edited by D. R. Yarkony, pp. 725–856, (World Scientific, Singapore, 1995).

[66] P. Jurečka, J. Šponer, J. Černý, and P. Hobza, *Phys. Chem. Chem. Phys.* **8** (2006) 1985–1993.

[67] G. A. Petersson, D. K. Malick, M. J. Frisch, and M. Braunstein, *J. Chem. Phys.* **123** (2005) 074111.

[68] F. Jensen, *Theo. Chem. Acc.* **113** (2005) 267–273.

[69] Z.-H. Li and M. W. Wong, *Chem. Phys. Lett.* **337** (2001) 209–216.

[70] M. R. Nyden and G. A. Petersson, *J. Chem. Phys.* **75** (1981) 1843–1862.

[71] P. Jurečka and P. Hobza, *Chem. Phys. Lett.* **365** (2002) 89–94.

[72] E. W. Ignacio and H. B. Schlegel, *J. Comput. Chem.* **12** (1991) 751–760.

[73] J. S. Dewar and A. J. Holder, *J. Comput. Chem.* **3** (1989) 311–313.

[74] R. H. Nobes, W. J. Bouma, and L. Radom, *Chem. Phys. Lett.* **89** (1982) 497–500.

[75] J. A. Pople, M. J. Frisch, B. T. Luke, and J. S. Binkley, *Intl. J. Quant. Chem.* **17** (1983) 307–320.

[76] R. Crespo-Otero, L. A. Montero, W.-D. Stohrer, and J. M. García de la Vega, *J. Chem. Phys.* **123** (2005) 134107.

[77] M. L. Senent and S. Wilson, *Intl. J. Quant. Chem.* **82** (2001) 282–292.

[78] I. Mayer and P. Valiron, *J. Chem. Phys.* **109** (1998) 3360–3373.

[79] F. Jensen, *Chem. Phys. Lett.* **261** (1996) 633–636.

[80] I. Mayer, *Theo. Chem. Acc.* **72** (1987) 207–210.

[81] S. F. Boys and F. Bernardi, *Mol. Phys.* **19** (1970) 553–566.

[82] H. B. Jansen and P. Ros, *Chem. Phys. Lett.* **3** (1969) 140–143.

[83] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, *Journal of Physical Chemistry A* **98** (1994) 11623–11627.

[84] A. D. Becke, *J. Chem. Phys.* **98** (1993) 5648.

[85] C. Lee, W. Yang, and R. G. Parr, *Phys. Rev. B* **37** (1988) 785–789.

[86] S. H. Vosko, L. Wilk, and M. Nusair, *Can. J. Phys.* **58** (1980) 1200-1211.

[87] T. H. Dunning Jr., *J. Chem. Phys.* **90** (1989) 1007–1023.

[88] R. A. Kendall, T. H. Dunning Jr., and R. J. Harrison, *J. Chem. Phys.* **96** (1992) 6796–6806.

[89] M. W. Gordon, M. S. ans Schmidt, *Advances in electronic structure theory: GAMESS a decade later*, in *Theory and Applications of Computational Chemistry: The first forty years*, edited by C. E. Dykstra, G. Frenking, K. S. Kim, and Scuseria, pp. 1167–1189, (Elsevier, Amsterdam, 2005).

[90] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, H. J. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery, *J. Comput. Chem.* **14** (1993) 1347–1363.

[91] P. Echenique and J. L. Alonso, *J. Comput. Chem.* **27** (2006) 1076–1087.

[92] J. L. Alonso and P. Echenique, *J. Comput. Chem.* **27** (2006) 238–252.

[93] P. Echenique, *A note on the accuracy of free energy functions in protein folding: Propagation of errors from dipeptides to polypeptides*, In progress, 2008.

[94] J. Kaminský and F. Jensen, *J. Chem. Theory Comput.* **3** (2007) 1774–1788.

[95] R. Ditchfield, W. J. Hehre, and J. A. Pople, *J. Chem. Phys.* **54** (1971) 724–728.

[96] W. J. Hehre, R. Ditchfield, and J. A. Pople, *J. Chem. Phys.* **56** (1972) 2257–2261.

[97] M. J. Frisch, J. A. Pople, and J. S. Binkley, *J. Chem. Phys.* **80** (1984) 3265–3269.

[98] R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople, *J. Chem. Phys.* **72** (1980) 650–654.

[99] J. S. Binkley, J. A. Pople, and W. J. Hehre, *J. Am. Chem. Soc.* **102** (1980) 939–947.

[100] G. W. Spitznagel, T. Clark, J. Chandrasekhar, and P. v. R. Schleyer, *J. Comput. Chem.* **3** (1982) 363–371.

[101] T. Clark, J. Chandrasekhar, G. W. Spitznagel, and P. v. R. Schleyer, *J. Comput. Chem.* **4** (1983) 294–301.

[102] P. C. Hariharan and J. A. Pople, *Theor. Chim. Acta* **28** (1973) 213–222.