JOURNAL OF PHYSICAL ORGANIC CHEMISTRY
J. Phys. Org. Chem. 2006; **19:** 196–205
Published online 30 January 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/poc.1019

Prediction of experimentally unknown r_e distances of organic molecules from Dunning basis set extrapolations for *ab initio* post-HF calculations

Alexander Neugebauer[1]* and Günter Häfelinger[2]

[1]Saarland University, Pharmaceutical and Medicinal Chemistry, Im Stadtwald, Building 32, D-66123 Saarbruecken, Germany
[2]University of Tuebingen, Institute of Organic Chemistry, Auf der Morgenstelle 18, D-72076 Tuebingen, Germany

Received 25 June 2005; revised 12 December 2005; accepted 13 December 2005

ABSTRACT: An approach to estimate equilibrium r_e bond lengths of organic molecules which contain standard bonding situations for CC, CH, CO and CN distances from only one equation is presented. For this, optimizations of molecular geometries using correlated post-Hartree–Fock and density functional methods have been performed. A selection scheme was developed to determine the most reliable methodology for prediction of equilibrium r_e distances of covalent bonds from a set of investigated theoretical methods. Consequently, distances computed in the CCSD(T) procedure via exponential extrapolation from a consecutive set of Dunning cc-pVXZ basis sets by use of Eqn (2) are accurate up to ± 0.0005 Å in comparison to experimentally available r_e distances. Applications for predictions of the experimentally unknown r_e distances of methanol, methylamine and methylenimine are presented.

Additionally the estimation of r_e distances of larger, chemically more interesting molecules is possible by lower order calculations (e.g. DFT B3LYP/cc-pVDZ) via linear correlation statistics using the results from our r_e reference model system via Eqn (3). Copyright © 2006 John Wiley & Sons, Ltd.
Supplementary electronic material for this paper is available in Wiley Interscience at http://www.interscience.wiley.com/jpages/0894-3230/suppmat/

KEYWORDS: experimental and calculated equilibrium r_e distances; r_e distance predictions for methanol, methylamine and methylenimine; CCSD(T) and MP4(SDQ) optimizations; DFT B3LYP optimizations; linear least-squares regressions

INTRODUCTION

Since the development of capable computer facilities, numerical quantum chemistry has been very successful. A major aspect in this field is still the calculation of geometries of molecules performed routinely by *ab initio* gradient methods, which deliver distances at the minimum of the corresponding potential energy curve denoted as r_e. Because of the approximate nature of such calculations none of the calculated results[1] agree completely with accurate experimental gas-phase determinations[2] of molecular structures that deliver method-dependent parameters, as reviewed in Refs 3 and 4. The numerical results are dependent on the selected method of quantum chemical calculations (HF, post-HF MP or CI, or DFT) and the quality of the basis set expansion chosen for evaluation of the target system. Another serious problem is the scant availability of highly accurate experimental r_e distances. Determination of experimental gas-phase r_e values that refer to minima of the energy hypersurface implies the use of electron diffraction[3] or microwave spectroscopy[4] involving corrections for anharmonicity in both cases. These procedures are rather complicated and applicable only for small molecules that contain mostly CC and CH bond lengths if organic compounds are considered.[2] Experimental equilibrium r_e distances for C–O and C–N single bonds are unknown up to now.

One aim of our research is to determine how accurately we can calculate molecular r_e distances by various quantum chemical methods, such as Hartree-Fock (HF), density functional theory (DFT) and post-HF methods, and how precisely we can predict known and unknown r_e distances of organic molecules. In previous studies we treated CC bonds[5,6] as well as CH bond lengths[6,7] by linear regression statistics. In Ref. 6 we evaluated the accuracy of CH and CC bond length calculations for various density functional methods and introduced scaling factors to generate results of higher order basis set optimizations on r_e distances. Using this scaling scheme one can approximate r_e distances by relatively inexpensive DFT calculations with double-zeta basis sets. Thus larger and chemically more interesting molecules can be treated at this approximate level. In another study[8] the performance of a wide range of theoretical methods was evaluated for CO bond lengths and a general ranking scheme was introduced that allows distance-dependent determination of the best method/basis set combination

**Correspondence to:* A. Neugebauer, Saarland University, Pharmaceutical and Medicinal Chemistry, Im Stadtwald, Building 32, D-66123 Saarbruecken, Germany.
E-mail: a.neugebauer@mx.uni-saarland.de

of a given set of methods (HF, MP2, DFT), various basis sets and five molecules for which experimental CO r_e distances are available. The general behaviour of quantum chemical methods for calculations of molecular geometries is discussed in Refs 6 and 8 and references therein.

Here we concentrate on treating types of r_e distances of organic molecules for which no experimental r_e data (e. g. the types C–O, C–N and C=N) are available. This means determining a level of theory that fits best to available experimental r_e distances for all main bond types between carbon and H, C, N and O appearing in organic molecules. A very important property at this level of theory is the feasibility of the computational efforts. Very advanced models such as CCSDT (all electrons)/cc-pCV5Z, as described by Helgaker *et al.*,[9] are quite unusable for this because at that demanding level of theory only very small molecules such as CO or CH_2 can be treated.

In addition, some test cases for reference models that include core correlation contributions were investigated[10] but are not included in this paper. Specific studies on the core-correlation effect of electron correlation have been performed by Helgaker *et al.*[9] and by Martin[11] for molecules **1**, **3**, **6**, **7** and **10** of our study. Koput[12] studied the equilibrium structure of methanol by CCSD(T) calculations using consecutive Dunning basis sets quite comparable to our calculations.

In Refs 6 and 8 we studied classes of organic bond types (CC, CH, CO) for each class separately, but now we try to find a more universal level of theory representing all common types of bonds occurring in organic molecules (but limited to C, H, N and O) in a reasonable way. We call this level of theory the reference model because data delivered at this level of theory are intended as an exact reference (near r_e distances) for experimentally accessible or inaccessible r_e distances. For this post-HF correlation, interaction methods such as fourth-order Møller–Plesset perturbation calculations (MP4(SDQ)) and coupled cluster (CC) calculations with single, double and approximate triple excitations (CCSD(T)) have been performed using Dunning's correlation-consistent polarized valence X-tuple basis sets[13,14] (cc-pVXZ) from $X = 2$ up to $X = 4$ or even $X = 5$. The correlation methods mentioned above allow the calculation of very reliable geometries of organic compounds but they are also very demanding in computing resources. Thus only very small molecules can be calculated. Nevertheless a small number of organic molecules are sufficient to derive a stable reference for prediction of near r_e distances of larger molecules containing standard binding situations.

We also tried to derive analogous reference models from DFT B3LYP calculations in a similar way to the study of Martin *et al.*[15]

Our results obtained by the best reference model (best implies economy and accuracy) can be used as a replacement for experimentally inaccessible r_e distances. Finally, such findings can be utilized to predict experimentally unavailable r_e distances from the reference level of theory and for larger molecules, for which a treatment at the reference level of theory is impossible, from lower order calculations as described in Refs 6 and 8 and here in the section on the prediction of r_e distances. Thus it is possible to predict equilibrium distances of larger and chemically more interesting molecules.

CALCULATIONAL PROCEDURE

Ground-state molecular geometries of ten organic molecules—methane (**1**), ethane (**2**), ethene (**3**), ethyne (**4**), methanol (**5**), formaldehyde (**6**), carbon monoxide (**7**), methylamine (**8**), methylenimine (**9**) and hydrogen cyanide (**10**)—with single, double and triple bonds to carbon were determined by optimizations using fourth-order Møller–Plesset perturbation theory[1] with single, double and quadruple excitation (MP4(SDQ)) as well as coupled cluster theory[1] with single, double and approximate triple excitations (CCSD(T)). Further, density functional theory (DFT)[16,17] with the popular method B3LYP (*B*ecke *3*-parameter-*L*ee-*Y*ang-*P*arr hybrid exchange-correlation functional[18,19]) was used. Calculations have been carried out with the standard correlation-consistent polarized valence Gaussian basis sets (cc-pVXZ, $X = $ D,T,Q,5) of Dunning,[13,14] which present systematic consistent improvements in size from double-zeta to quintuple-zeta quality. For estimations of the basis set limit geometries we used exponential extrapolations obtained from cc-pVTZ, cc-pVQZ and cc-pV5Z (TQ5-limits) as well as from cc-pVDZ, cc-pVTZ and cc-pVQZ (DTQ-limits) calculations. All calculations and optimizations were performed using the Gaussian 98 program system (Revision A.7).[20] For DFT calculations we used the 'fine' integration grid of the Gaussian 98 program. A reviewer pointed our attention to the study of Martin *et al.*,[21] referring to problems with typical grid sizes for integrations in DFT calculations, but we did not check possible changes due to the use of 'ultrafine' grids. All computations were carried out on two dual-processor PC-systems (2×1.53 GHz) running under Linux.

RESULTS AND DISCUSSION

Determination of basis set limits

One main problem of calculations of highly accurate electronic structures and properties of molecules is the truncation of the Gaussian AO basis set, which is the most important source of errors. Unfortunately the computational costs grow very fast if one increases the basis set expansion. The convergence to the basis set limit (which means that the total energy will not change if one adds some more Gaussian basis functions) is generally very

slow, thus computing complete basis set energies or properties of a chemical system is very demanding. A well-known solution solving the slow convergence problem is to consider a series of basis sets that comprise systematic improvements to the ground-state energy or property and to develop an extrapolation based on the asymptotic form of these series. Basis set series with systematic increase of the one-particle basis set were developed by Dunning *et al.*,[13,14,22–24] so-called correlation-consistent polarized basis sets (cc-pVXZ), or by Almlöf, Helgaker and Taylor,[25–27] denoted as atomic natural orbital (ANO) basis sets.

The standard Gaussian basis sets[28] of Pople's group (e.g. the series 6-31G, 6-31G*, 6-31G** and 6-311G, 6-311G*, 6-311G**, where one star denotes polarization d-functions on heavy atoms and the second star denotes polarization p-functions on hydrogen) for total HF energies, in agreement with the variation principle, lead to a lowering of the total energies with increase of basis sets but the correspondingly optimized bond lengths do not follow a regular change towards a limiting HF distance, which is also different from the experimental r_e value (see Fig. 1 for CH distances of **1** and **4**). In contrast to this, the increasing Dunning basis sets follow a regular order for energies as well as for distances towards corresponding limiting values, which may be approximated by an exponential function of the general form of Eqn (1)

$$r(X) = r(\infty) + a \cdot e^{-b \cdot X} \quad (1)$$

which was used first by Feller[29] in 1992 for total energies. We also tried other approximate functions such as polynomial functions or potential functions (data not shown), but corresponding basis set limits of r_e distances differ only slightly from the basis set limits obtained by the exponential function of Eqn (1), denoted there as $r(\infty)$.

Here, we concentrate on the correlation-consistent polarized valence basis set series (cc-pVXZ) of Dunning[13] from double-zeta ($X = 2$) to quintuble-zeta ($X = 5$). Table 1 summarizes our calculated CCSD(T) distances and basis set limits derived by extrapolation for a set of ten organic molecules **1**–**10** containing a wide range of different bond types: C–H for C_{sp3}–H, C_{sp2}–H and C_{sp}–H; C–C single, C=C double and C≡C triple bonds; C–O single, C=O double and C≡O triple bonds; C–N single, C=N double and C≡N triple bonds; N–H and O–H bonds. Corresponding tables for distances from MP4(SDQ) and B3LYP calculations can be found in the Supplementary material. Experimental r_e values[30–36] are also given in Table 1 as far as available. MP4(SDQ) and CCSD(T) distances generally show a lowering upon going to larger basis sets. Thus exponential determination of the basis set limits via Eqn (1) are successfully applicable except in the case of methanol (**5**) for the CCSD(T) and DFT B3LYP method.

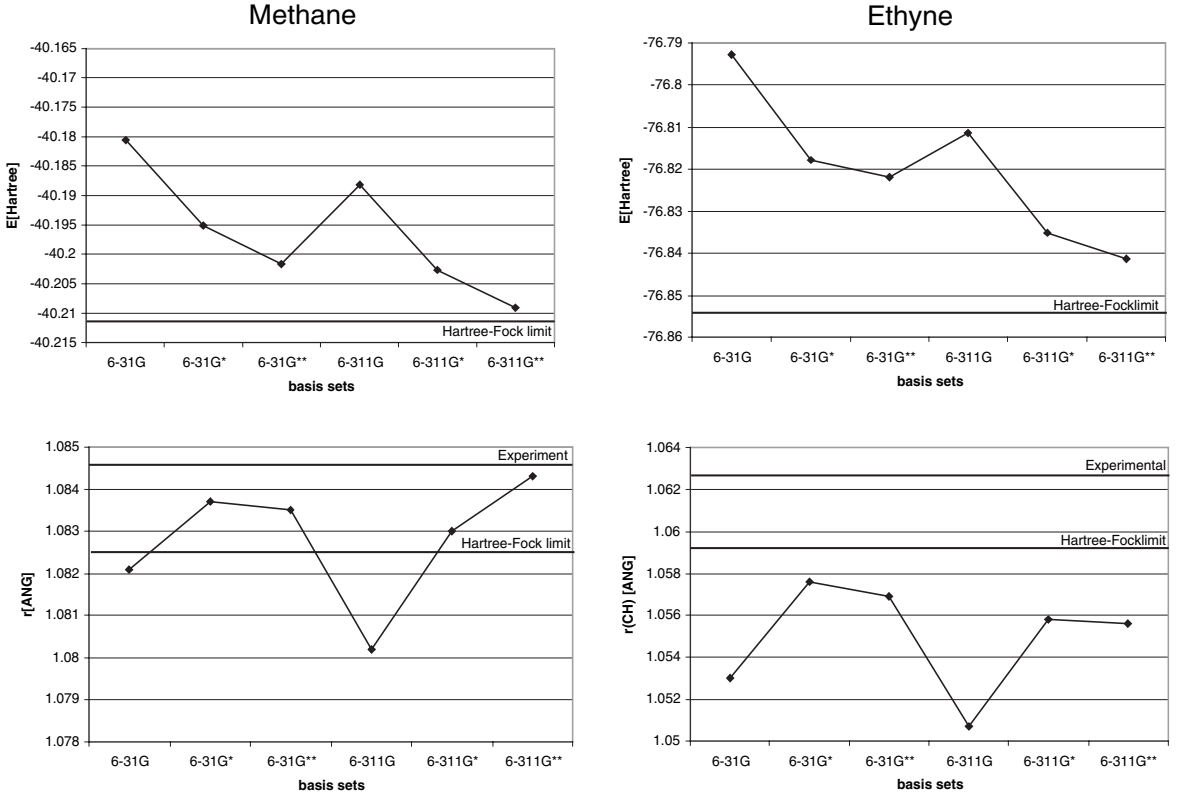

Figure 1. Behaviour of standard Pople basis sets from Hartree–Fock calculations for total energies and equilibrium CH distances of **1** and **4**. The Hartree–Fock limit energies and distances are given by a bold line

Table 1. Calculated (CCSD(T)/cc-pVXZ, $X = 2$–5) and experimental equilibrium r_e distances (Å) of molecules **1**–**10**

Molecule	No.	Bond	cc-pVDZ	cc-pVTZ	cc-pVQZ	cc-pV5Z	Limit[a,b]	Experiment
CH_4	**1**	C–H	1.1038	1.0890	1.0879	1.0876	1.0874[a]	1.0858 (Ref. 32)
H_3CCH_3	**2**	C–C	1.5359	1.5288	1.5260	—	1.52420[b]	1.5220 (Ref. 35)
		C–H	1.1066	1.0919	1.0910	—	1.09090[b]	1.0895 (Ref. 35)
H_2CCH_2	**3**	C=C	1.3516	1.3370	1.3342	1.3335	1.33327[a]	1.3307 (3)[c]
		C–H	1.0984	1.0832	1.0823	1.0824	1.08235[a]	1.0809 (3) (Ref. 34)
HCCH	**4**	C≡C	1.2287	1.2096	1.2065	1.2057	1.20542[a]	1.2027 (Ref. 33)
		C–H	1.0789	1.0638	1.0634	1.0633	1.06327[a]	1.06208 (Ref. 33)
H_3COH	**5**	C–O	1.4206	1.4206	1.4194	—	—	—
		H^1–C[c]	1.1043	1.0888	1.0877	—	1.0876[b]	—
		H^2–C[d]	1.1114	1.0952	1.0936	—	1.0934[b]	—
		H–O	0.9666	0.9595	0.9577	—	0.9571[b]	—
H_2CO	**6**	C=O	1.2156	1.2096	1.2066	1.2051	1.20360[a]	1.2031(5) (Ref. 31)
		C–H	1.1199	1.1033	1.1022	1.1021	1.10209[a]	1.1003(5) (Ref. 31)
CO	**7**	C≡O	1.1446	1.1358	1.1314	1.1307	1.13057[a]	1.1284 (Ref. 30)
H_3CNH_2	**8**	C–N	1.4736	1.4691	1.4685	—	1.4684[b]	—
		H^1–C[e]	1.1133	1.0970	1.0938	—	1.0930[b]	—
		H^2–C[f]	1.1060	1.0909	1.0879	—	1.0872[b]	—
		N–H	1.0266	1.0136	1.0111	—	1.0105[b]	—
H_2CNH	**9**	C=N	1.2890	1.2776	1.2753	—	1.2747[b]	—
		H^1–C[g]	1.1039	1.0881	1.0850	—	1.0842[b]	—
		H^2–C[h]	1.1084	1.0923	1.0891	—	1.0883[b]	—
		N–H	1.0349	1.0217	1.0192	—	1.0186[b]	—
HCN	**10**	C≡N	1.1754	1.1601	1.1564	1.1556	1.15538	1.15324(2) (Ref. 36)
		C–H	1.0825	1.0670	1.0668	1.0666	1.06660	1.06501(8) (Ref. 36)

[a] $r(X) = r(\infty) + a \cdot e^{-b \cdot X}$; $X = 3$–5 (TQ5-limit). [b] $r(X) = r(\infty) + a \cdot e^{-b \cdot X}$; $X = 2$–4 (DTQ-limit).
[c] H atom in the C–O plane.
[d] H atoms out of the C–O plane.
[e] H atom in the C–N plane.
[f] H atoms out of the C–N plane.
[g] H atom *trans* to the N-hydrogen.
[h] H atom *cis* to the N-hydrogen.
[i] Experimental uncertainties in parentheses refer to the last digit.

Because the CCSD(T) basis set limit values are slightly too long and MP4(SDQ) basis set limit values are often too short, these two values form an upper and a lower bound bracketing the experimental r_e value. This means that there is a high possibility of finding the experimentally available or unknown r_e values within these boundaries. Figure 2 gives an impression of this basis set limit behaviour for these two procedures for a selection of molecules and bond types.

Absolute average errors

Absolute average errors (from differences between calculated and experimental r_e bond lengths) and standard deviations for all distances of molecules **1**–**4**, **6**, **7** and **10** compared with experimental r_e bond lengths for all examined methods and basis sets, as well as the extrapolated basis set limit, are collected in Table 2. Values for all CCSD(T) calculations are above experimental r_e determinations, while results obtained by MP4(SDQ) and B3LYP overestimate and underestimate experimental determinations. The absolute average errors of the correlated methods are remarkably small (0.0018 Å and 0.0014 Å for CCSD(T) and MP4(SDQ) in its basis set limit), but these are smallest for the MP4(SDQ)/cc-pVQZ level of theory. The absolute average error of the density functional method in the extrapolated limit (0.0045 Å) is substantially larger than for the correlated methods, therefore the DFT B3LYP method is unsuitable for a reliable reference model. A similar behaviour is given by the standard deviations of the absolute errors (Table 2). These are smallest for the CCSD(T)/limit level of theory. Convergence of absolute average errors and standard deviations of absolute errors can be observed solely for the CCSD(T) methods. Figure 3 shows this decreasing trend of absolute average errors with increasing basis set size of the CCSD(T) method. The remaining error of 0.0018 Å in the basis set limit of the CCSD(T) method is caused by neglect of core electron correlation, incompleteness of the CCSD(T) method itself and relativistic effects. Another source of error is the inexactness of the experimental determinations of equilibrium r_e structures.

Correlation of experimental and calculated r_e distances

All bond lengths of molecules **1**–**4**, **6**, **7** and **10** for which highly precise experimental r_e distances are available

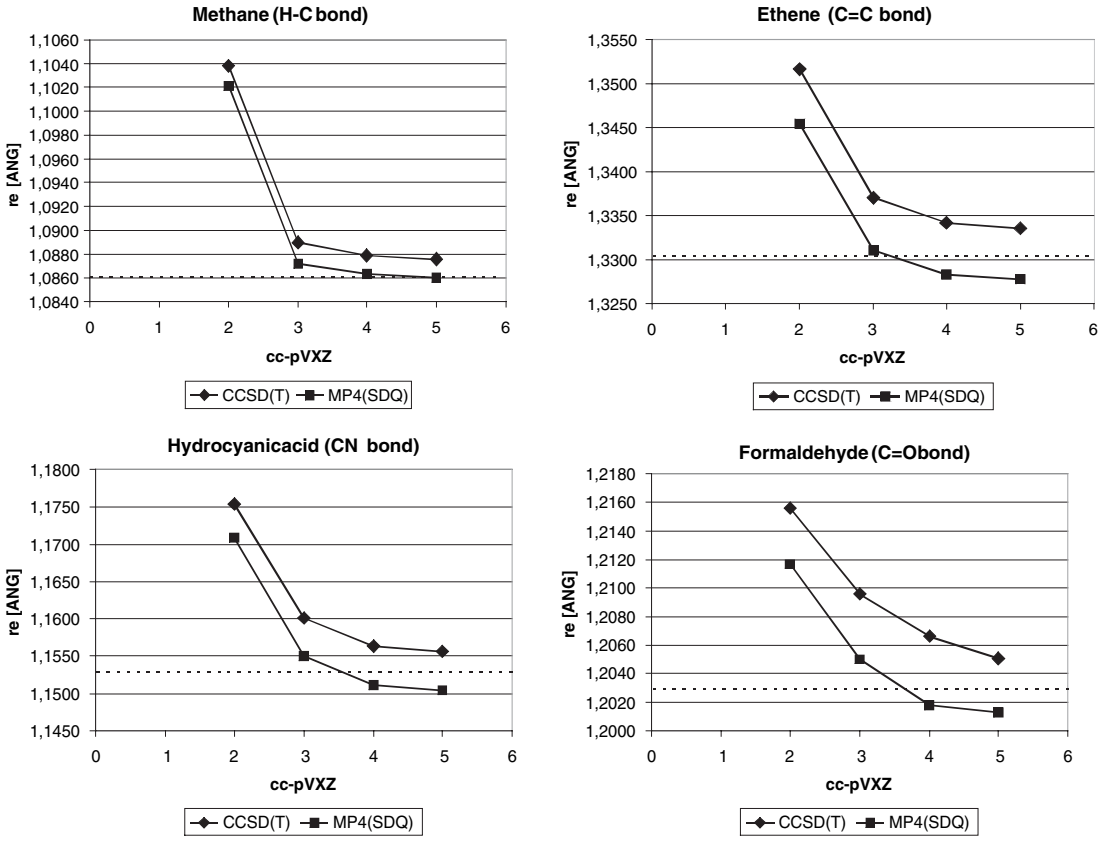

Figure 2. Behaviour of equilibrium r_e distances for MP4(SDQ) and CCSD(T) calculations with different molecules and bond types for basis set expansion from cc-pVDZ to cc-pV5Z. Dashed lines indicate the experimental values

Table 2. Absolute average errors and standard deviations compared with experimentally determined r_e bond lengths of all distances for molecules **1**–**4**, **6**, **7** and **10** with successive expansion of the size of the basis set

Method	Basis set	Absolute average error (Å)	Independent determinations	Standard deviation of absolute errors (Å)
CCSD(T)	cc-pVDZ	0.0182	12	0.0035
CCSD(T)	cc-pVTZ	0.0046	12	0.0022
CCSD(T)	cc-pVQZ	0.0026	12	0.00096
CCSD(T)	cc-pV5Z	0.0020	10	0.0005
CCSD(T)	Limit	0.0018	12	0.0005
MP4(SDQ)	cc-pVDZ	0.0154	12	0.0028
MP4(SDQ)	cc-pVTZ	0.0015	12	0.0014
MP4(SDQ)	cc-pVQZ	0.0011	12	0.0008
MP4(SDQ)	cc-pV5Z	0.0014	12	0.0010
MP4(SDQ)	Limit	0.0014	12	0.0011
B3LYP	cc-pVDZ	0.0191	12	0.0197
B3LYP	cc-pVTZ	0.0038	12	0.0025
B3LYP	cc-pVQZ	0.0040	12	0.0027
B3LYP	cc-pV5Z	0.0039	12	0.0027
B3LYP	Limit	0.0045	12	0.0027

were correlated by linear least-squares regressions against calculated distances. Table 3 shows the correspondingly derived statistical parameters: the linear regression coefficients (R), estimated standard deviations (esd), slope (m) and intercept (b). The esd value is always a positive value, thus only absolute deviations are represented. Figure 4 presents a diagram of experimental r_e bond lengths with regard to the MP4(SDQ) method in its extrapolated basis set limits. The regression line with $m = 1.0009$ and $b = 0.0003$ Å is closest to unit slope, which means a satisfactorily absolute agreement of experimental and calculated values with an estimated

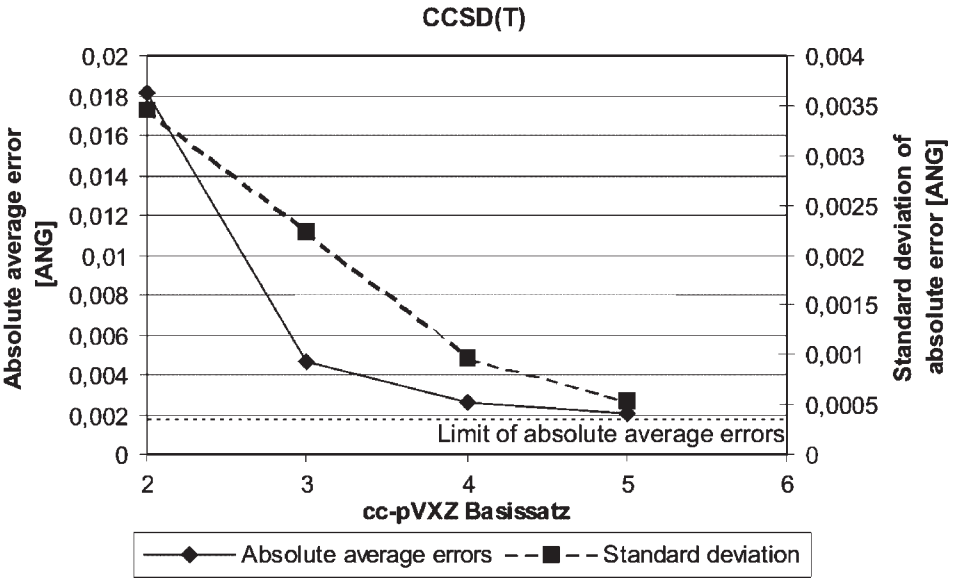

Figure 3. Convergence of CCSD(T) calculations for absolute average errors and standard deviations of absolute errors with increasing basis set size. The estimated basis set limit of the absolute average errors is indicated by a dashed line

Table 3. Parameters of least-squares regressions (R = correlation coefficient, esd = standard deviation, m = slope and b = intercept) of experimentally available and correspondingly calculated distances for 12 data points of molecules **1**–**4**, **6**, **7** and **10**

Basis set	Method	R	esd (Å)	m	b (Å)
cc-pVDZ	MP4(SDQ)	0.99979	0.00291	1.0071	−0.0237
	CCSD(T)	0.99965	0.00378	1.0028	−0.0215
	B3LYP	0.99906	0.00618	1.0251	−0.0398
cc-pVTZ	MP4(SDQ)	0.99995	0.00143	0.9941	0.0055
	CCSD(T)	0.99992	0.00184	0.9890	0.0083
	B3LYP	0.99942	0.00487	1.0005	0.0002
cc-pVQZ	MP4(SDQ)	0.99996	0.00131	0.9978	0.0033
	CCSD(T)	0.99999	0.00064	0.9942	0.0043
	B3LYP	0.99938	0.00503	0.9989	0.0028
cc-pV5Z	MP4(SDQ)	0.99995	0.00136	0.9992	0.0021
	CCSD(T)	0.99999	0.00034	0.9945	0.0043
	B3LYP	0.99941	0.00491	0.9994	0.0023
Limit[a]	MP4(SDQ)	0.99996	0.00127	1.0009	0.0003
	CCSD(T)	0.99997	0.00063	0.9967	0.0020
	B3LYP	0.99950	0.00489	0.9990	0.0022

[a] $r(X) = r(\infty) + a \cdot e^{-b \cdot X}$; X = 3–5 (TQ5-limit).

standard error (esd) of 0.0013 Å. Statistically best is the CCSD(T) limit value with an esd of 0.00063 Å.

Dependence of estimated standard error (esd) on basis set expansion

The behaviour of the esd from linear regressions for each considered method upon expansion to larger basis sets is quite different. Figure 5 shows esd values for the cc-pVXZ (X = 2–5) series of basis sets for all investigated methods. The esd values of MP4(SDQ) calculated distances strongly decrease from cc-pVDZ to cc-pVTZ and only slightly from cc-pVTZ to cc-pVQZ. Further basis set expansion to cc-pV5Z leads to a slight increase of the esd value, but this value is smallest for the cc-pVQZ basis set. An explanation for this behaviour is the fact that cc-pVDZ and cc-pVTZ calculated distances are too long compared with experimental bond lengths, cc-pVQZ calculated distances are within experimental values and cc-pV5Z calculated distances are too short in comparison to experiment. Therefore, calculations using this quadruple basis set are better than the extrapolated basis set limit esd concerning experimentally determined r_e values. This behaviour is inconsistent and so the MP4(SDQ) method is not the best choice in this case for a

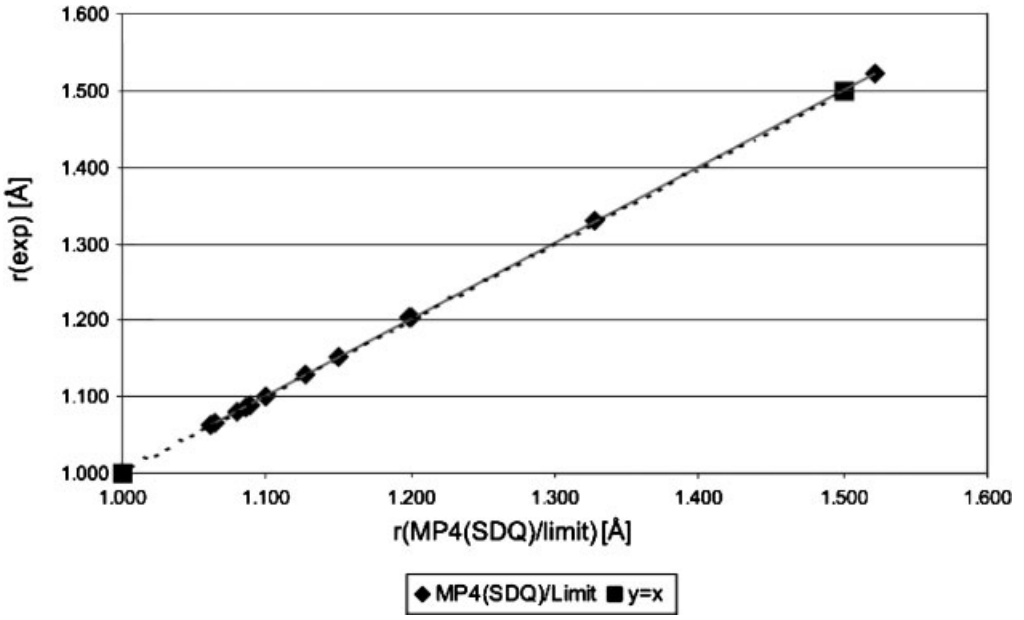

Figure 4. Linear regression of r_e bond lengths with regard to the MP4(SDQ) method in its basis set limit

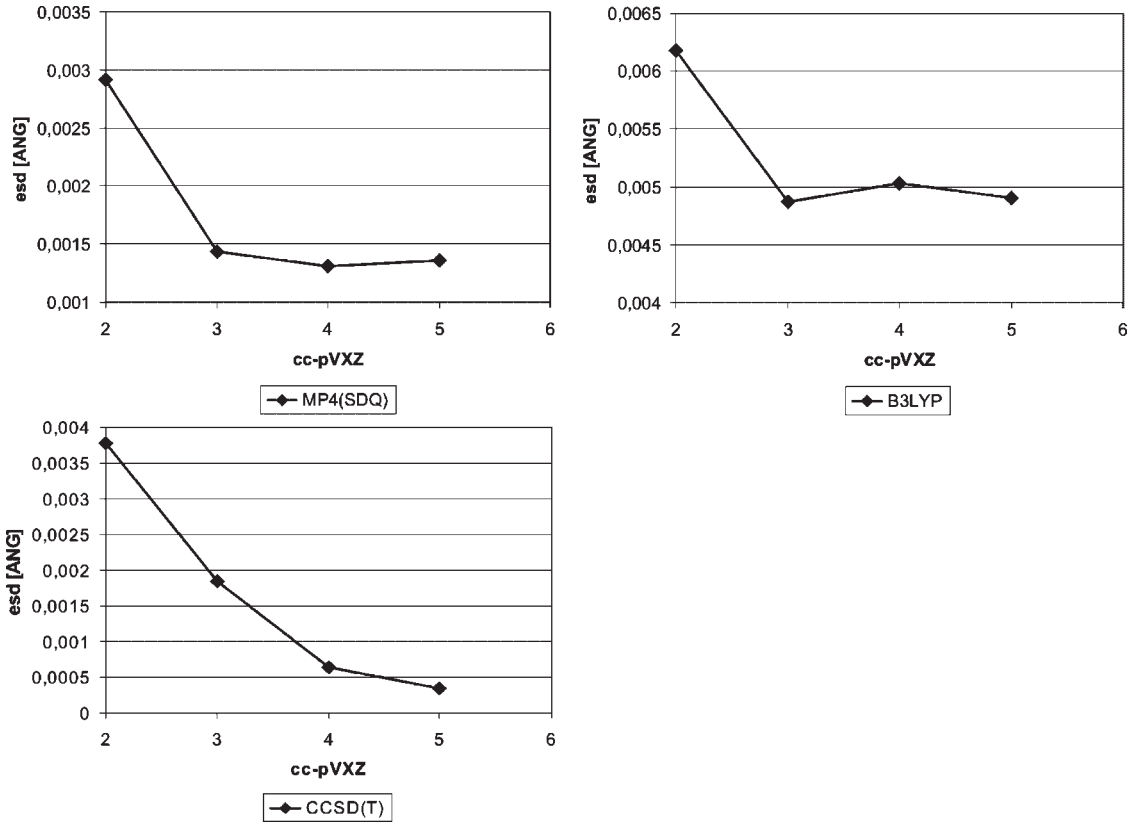

Figure 5. Behaviour of the estimated standard error (esd) for all investigated methods (MP4(SDQ), CCSD(T) and B3LYP) for the basis set expansion from cc-pVDZ to cc-pV5Z

well-formed reference methodology. The same trend is reflected by absolute average errors and standard deviations of average errors shown in Table 2.

The CCSD(T) esd values represent a decreasing trend with increasing basis set expansion. At 0.00034 Å they are the smallest for the cc-pV5Z basis set. However, the esd limit value of 0.00063 Å is slightly larger but both values indicate an extremely good precision at <0.001 Å.

The B3LYP esd values show an unsteady trend upon going from cc-pVDZ to cc-pV5Z basis set expansion.

Thus this method is rather unsuitable for predictions of unknown r_e distances.

Selection of a reference model for the prediction of r_e distances

In this section the results presented above will be used to determine the best reference model whose results fit best to experimentally determined equilibrium r_e distances using a selection scheme. The first criterion for this selection scheme states that the linear regression coefficient R of correlations of calculated against experimental values must be considerably high ($R > 0.999$). If $R < 0.999$ the linearity of the connection of experimental and calculated values is insufficient. Another criterion for linearity is presented by the estimated standard error (esd), which should be below 0.01 Å. These linearity criteria are fulfilled by all three investigated theoretical methods. The next important factor is the convergence of absolute average errors and standard deviations of absolute errors (σ) towards a limit comparable to experimentally determined r_e distances with successive expansion of the basis set size. This criterion is reached only by the CCSD(T) method. The systematic convergence of absolute average errors is necessary to make sure that distances extrapolated towards the basis set limit deliver the best results of the method. Another criterion is the consistent overestimation or underestimation of all calculated distances by the selected method. This behaviour permits an easy correction of distances towards experimentally available r_e distances. All mentioned criteria are fulfilled only by the CCSD(T) method in its basis set limit. Thus, this level of theory scaled by an absolute average error of -0.0018 Å from Table 2 is suggested for reference calculations for all kinds of considered bonds from CH, over CO, CN to CC with the same Eqn (2). The terms in parentheses in Eqn (2) denote the CCSD(T) method and its exponential extrapolation via Eqn (1) from triple to quintuple Dunning basis sets (TQ5-limit).

$$r_e^{exp}[\text{Å}] = \text{r(CCSD(T)/cc} - \text{pV}\infty\text{Z)} - 0.0018 \qquad (2)$$

The error range is estimated to be ± 0.0005 Å due to the standard deviation of absolute errors from Table 2.

Prediction of experimentally unknown r_e distances with high precision

Prediction of r_e distances of methanol (5), methylamine (8) and methylenimine (9). The aim of this study is to predict reliable r_e distances of organic molecules for which no experimental r_e bond lengths are available, therefore our selected reference model (Eqn 2) will be used to estimate r_e distances for **5**, **8** and **9** with derived values presented in Table 4. The predicted C=N r_e distance of **9** is 1.2729 Å, in comparison to 1.4666 Å for the C–N single bond in **8**. These distances should be the most precise values known for these systems today, with an error limit of ± 0.0005 Å.

Table 4. Prediction of r_e bond lengths of **7**, **8** and **9** for which no experimental equilibrium distances are available using the coupled cluster reference (Eqn 2)[a] distances are given in Å. Estimated uncertainties ± 0.0005 Å

Molecule	No.	Bond (Å)	$r_e^{pred.}$ via Eqn (2)
Methanol	**7**	C–O	[1.4156][b,c]
			1.416[d]
		H^1–C	1.0858
		H^2–C	1.0916
		H–O	0.9553
Methylamine	**8**	C–N	1.4666
		H^1–C	1.0912
		H^2–C	1.0854
		N–H	1.0087
Methylenimine	**9**	C=N	1.2729
		H^1–C	1.0824
		H^2–C	1.0865
		N–H	1.0168

[a]Eqn (2): r_e^{exp} [Å] = r (CCSD(T)/cc-pV∞Z) − 0.0018.
[b]Convergence of calculated C–O bond lengths failed (see Table 1).
[c]Suggested value: see text.
[d]Value determined in Ref. 12.

Unfortunately this method is unable to determine an approximated r_e value for the C–O single bond in **5**, because of the convergence failure of calculated CCSD(T)/cc-pVXZ C–O bond lengths (see Table 1). This behaviour was also observed by Koput.[12]

For determination of a reliable C–O bond length for **5** we used the observation that most of the calculated MP4(SDQ)/cc-pVQZ distances are too short and most of the calculated CCSD(T)/cc-pVQZ distances are too long compared with precise experimental determinations. Nine distances out of a set of twelve distances (= 75%) of molecules **1**–**4**, **6**, **7** and **10** for which experimental determinations are available obey this observation. Therefore it can be supposed that the C–O single bond distance in **5** is between 1.4147 Å (MP4(SDQ)/cc-pVQZ distance) and 1.4194 Å (CCSD(T)/cc-pVQZ distance). The core-correlated CCSD(T)(all electrons)/cc-pV∞Z level of theory delivers a C–O r_e limit distance of 1.4152 Å, which is within the range of the two indicated calculations. To estimate an explicit value for the C–O single bond length we suggest that the average from the CCSD(T)/cc-pVQZ calculated distance of 1.4194 Å reduced by the absolute average error (with regard to molecules **1**–**4**, **6**, **7** and **10**) is 0.0026 Å and the core-correlated calculated limit distance of 1.4152 Å reduced by the absolute average error (with regard to molecules **1**, **4**, **6** and **7**) is 0.0009 Å. This leads to a prediction for r_e

(C–O) of 1.4156 ± 0.002 Å with an estimated error of <0.002 Å. The value derived by Koput[12] (1.416 Å) is in full agreement with our result.

Prediction of r_e distances of larger molecules. Calculated equilibrium r_e distances obtained by the selected reference model (Eqn. 2) were used as replacements for highly exact experimental data, therefore extrapolated limit distances from Eqn (1) for the set of our ten training molecules **1**–**10**, which contain nearly all important bond lengths of organic compounds (C–H, C–C, C=C, C≡C, C–O, C=O, C≡O, O–H, C–N, C=N, C≡N, N–H), were corrected by use of Eqn (2) to obtain near r_e values for these molecules. Correlations of these with distances optimized by the computationally less demanding DFT B3LYP method using the cc-pVDZ basis set led to the linear regression Eqn (3) for all defined kinds of bonds, which allows the prediction of B3LYP/cc-pVDZ-based r_e distances of larger molecules containing the bond types mentioned above.

$$r_e^{exp}[\text{Å}] = 1.02732 \cdot r_e(\text{B3LYP/cc} - \text{pVDZ}) - 0.04222 \quad (3)$$

As a test example benzene (C_6H_6, **11**) is considered: experimental r_e distances[37] are 1.0802 ± 0.0020 Å and 1.3914 ± 0.0010 Å. Application of Eqn (3) leads to absolute deviations for predicted distances from the experimental determination of 0.0002 Å for CH and 0.0005 Å for CC bond lengths. As a second example, ketene ($H_2C{=}C{=}O$, **12**) is selected: experimental determinations[38] are 1.0758 ± 0.0001 Å, 1.1603 ± 0.0003 Å and 1.3121 ± 0.0003 Å for C–H, C=O and C=C bond lengths. The B3LYP/cc-pVDZ level of theory was used again for geometry optimization. Absolute deviations from experimental values for DFT predictions by use of Eqn (3) are 0.0011 Å for C–H, 0.0025 Å for C=O and 0.0003 Å for C=C distances.

This approach via Eqn (3) is, as a matter of course, less accurate than the approximation procedure postulated in Eqn (2) but it is a cost-effective way for estimations of r_e distances of larger organic molecules containing CC, CO, CN and CH bond types. We assume that our predictions are reliable and useful for accurate structure determinations of larger organic molecules.

CONCLUSIONS

We present a procedure for estimating known and unknown r_e bond lengths of organic molecules containing common bond types for CH, CC, CO and CN distances of organic compounds with high precision and by only one equation. For this, Eqn (2) as a 'reference model' was determined, which delivers reliable near-r_e distances. Using this reference model, predictions of equilibrium r_e distances of methanol (**5**), methylamine (**8**) and methylenimine (**9**), for which no experimental r_e distances are available, have been presented.

Another application of these reference distances is the prediction of near-r_e distances of larger molecules by correlating the computed 'experimental' distances against distances derived from a modest level of theory such as B3LYP/cc-pVDZ via Eqn (3). We could show for two larger molecules that this procedure leads to reliable predictions of known r_e distances and it may lead to valuable prognoses of unknown r_e bond lengths.

Supplementary material

Results of MP4(SDQ/cc-pVXZ, X = 2–5) optimizations (Table 1S) and B3LYP/cc-pVXZ, X = 2–5 optimizations (Table 2S) and bracketing behaviour of experimental r_e distances between MP4(SDQ)/cc-pVQZ and CCSD(T)/cc-pVQZ calculations (Fig. 1S) are available in Wiley Interscience.

REFERENCES

1. Helgaker A, Jørgensen P, Olsen J. *Molecular Electronic-Structure Theory.* Wiley: Chichester, 2000.
2. Domenicano A, Hargittai A. *Accurate Molecular Structures.* Oxford University Press: Oxford, 1992.
3. Vilkov LV, Mastryukov VS, Sadova NI. *Determination of the Geometrical Structure of Free Molecules.* Mir Publishers: Moscow, 1983.
4. Kuchitsu K. *MTP International Review of Science*, Phys. Chem. Series 1, vol. 2. Medical and Technical Publ. Co. and University Park Press: Baltimore, 1972; 203.
5. Häfelinger G, Regelmann CU, Krygowski TM, Wozniak K. *J. Comput. Chem.* 1989; **10**: 329–343.
6. Neugebauer A, Häfelinger G. *J. Mol. Struct. (Theochem)* 2002; **578**: 229–247.
7. Häfelinger G. In *Similarity Models in Organic Chemistry, Biochemistry and Related Fields*, Zalewski RI, Krygowski TM, Shorter J (eds). Elsevier: Amsterdam, 1991; 177–229.
8. Neugebauer A, Häfelinger G. *J. Mol. Struct. (Theochem)* 2002; **585**: 35–47.
9. Halkier A, Jørgensen P, Gauss J, Helgaker T. *Chem. Phys. Lett.* 1997; **274**: 235–241.
10. Neugebauer A. *Methoden- und Basisatzabhängigkeit von Dichtfunktional- und Post-Hartree-Fock ab initio Methoden zur Berechnung und Vorhersage genauer Gleichgewichtsabstände in organischen Molekülen sowie die Einführung eines energetischen Klassifizierungsschemas für organische Verbindungsklassen homologer Funktionalitäten auf quantenchemischer Grundlage.* Dissertation, University of Tübingen, 2002.
11. Martin JML. *Chem. Phys. Lett.* 1995; **242**: 343–350.
12. Koput J. *J. Phys. Chem.* 2000; **A 104**: 10017–10022.
13. Dunning TH Jr. *J. Chem. Phys.* 1989; **90**: 1007–1023.
14. Dunning TH Jr, Peterson KA, Woon DE. In *Encyclopedia of Computational Chemistry*, Ragué Schleyer Pv (ed.). Wiley: Chichester, 1998.
15. Martin JML, El-Yazal J, François J-P. *Mol. Phys.* 1995; **86**: 1437–1450.
16. Hohenberg P, Kohn W. *Phys. Rev.* 1964; **B 136**: 864–871.
17. Kohn W, Sham LJ. *Phys. Rev.* 1965; **A 140**: 1133–1138.
18. Becke AD. *J. Chem. Phys.* 1993; **98**: 5648–5652.
19. Lee C, Yang W, Parr RG. *Phys Rev.* 1988; **B 37**: 785–789.
20. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA, Stratmann RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN,

Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Petersson GA, Ayala PY, Cui Q, Morokuma K, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Baboul AG, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Gonzalez C, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Andres JL, Head-Gordon M, Replogle ES, Pople JA. *Gaussian 98, Revision A.7*. Gaussian: Pittsburgh PA, 1998.
21. Martin JML, Bauschlicher CW Jr, Ricca A. *Comput. Phys. Commun.* 2001; **133**: 189–201.
22. Kendall RA, Dunning TH Jr, Harrison RJ. *J. Chem. Phys.* 1992; **96**: 6796–6806.
23. Woon DE, Dunning TH Jr. *J. Chem. Phys.* 1994; **100**: 2975–2988.
24. Wilson AK, Dunning TH Jr. *J. Chem. Phys.* 1997; **106**: 8718–8728.
25. Almlöf J, Taylor PR. *J. Chem. Phys.* 1987; **86**: 4070–4077.
26. Almlöf J, Taylor PR. *Adv. Quantum Chem.* 1992; **22**: 301–373.
27. Helgaker T, Taylor PR. In *Modern Electronic Structure Theory*, Yarkony DR (ed.). World Scientific: Singapore, 1995.
28. Hehre WJ, Radom L, Ragué Schleyer Pv, Pople JA. *Ab initio Molecular Orbital Theory.* Wiley: New York, 1986.
29. Feller D. *J. Chem. Phys.* 1992; **96**: 6104–6114.
30. Sorkhabi O. Jackson WM, Daizadeh I. *J. Chem. Ed.* 1998; **75**: 238–240.
31. Carter S, Handy NC. *J. Mol. Spectrosc.* 1996; **179**: 65–72.
32. Gray DL, Robiette AG. *Mol. Phys.* 1979; **37**: 1901–1920.
33. Martin JML, Lee TJ, Taylor PR. *J. Chem. Phys.* 1998; **108**: 676–691.
34. Martin JML, Taylor PR. *Chem. Phys. Lett.* 1996; **248**: 336–344.
35. Harmony MD. *J. Chem. Phys.* 1990; **93**: 7522–7523.
36. Carter S, Mills IM. *J. Chem. Phys.* 1992; **97**: 1606–1607.
37. Gauss J, Stanton JF. *J. Phys. Chem.* 2000; **A104**: 2865–2868.
38. East ALL, Wesley AD, Klippenstein SJ. *J. Chem. Phys.* 1995; **102**: 8506–8532.