Basis set convergence of CCSD(T) equilibrium geometries using a large and diverse set of molecular structures

Peter R. Spackman, Dylan Jayatilaka, and Amir Karton[a)]
School of Chemistry and Biochemistry, The University of Western Australia, Perth, WA 6009, Australia

(Received 11 June 2016; accepted 20 August 2016; published online 12 September 2016)

We examine the basis set convergence of the CCSD(T) method for obtaining the structures of the 108 neutral first- and second-row species in the W4-11 database (with up to five non-hydrogen atoms). This set includes a total of 181 unique bonds: 75 H—X, 49 X—Y, 43 X=Y, and 14 X≡Y bonds (where X and Y are first- and second-row atoms). As reference values, geometries optimized at the CCSD(T)/aug′-cc-pV(6+d)Z level of theory are used. We consider the basis set convergence of the CCSD(T) method with the correlation consistent basis sets cc-pV(n+d)Z and aug′-cc-pV(n+d)Z (n = D, T, Q, 5) and the Weigend–Ahlrichs def2-nZVPP basis sets (n = T, Q). For each increase in the highest angular momentum present in the basis set, the root-mean-square deviation (RMSD) over the bond distances is decreased by a factor of ~4. For example, the following RMSDs are obtained for the cc-pV(n+d)Z basis sets 0.0196 (D), 0.0050 (T), 0.0015 (Q), and 0.0004 (5) Å. Similar results are obtained for the aug′-cc-pV(n+d)Z and def2-nZVPP basis sets. The double-zeta and triple-zeta quality basis sets systematically and significantly overestimate the bond distances. A simple and cost-effective way to improve the performance of these basis sets is to scale the bond distances by an empirical scaling factor of 0.9865 (cc-pV(D+d)Z) and 0.9969 (cc-pV(T+d)Z). This results in RMSDs of 0.0080 (scaled cc-pV(D+d)Z) and 0.0029 (scaled cc-pV(T+d)Z) Å. The basis set convergence of larger basis sets can be accelerated via standard basis-set extrapolations. In addition, the basis set convergence of explicitly correlated CCSD(T)-F12 calculations is investigated in conjunction with the cc-pVnZ-F12 basis sets (n = D, T). Typically, one "gains" two angular momenta in the explicitly correlated calculations. That is, the CCSD(T)-F12/cc-pVnZ-F12 level of theory shows similar performance to the CCSD(T)/cc-pV(n+2)Z level of theory. In particular, the following RMSDs are obtained for the cc-pVnZ-F12 basis sets 0.0019 (D) and 0.0006 (T) Å. Overall, the CCSD(T)-F12/cc-pVDZ-F12 level of theory offers a stellar price-performance ratio and we recommend using it when highly accurate reference geometries are needed (e.g., in composite *ab initio* theories such as W4 and HEAT). *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4962168]

I. INTRODUCTION

Coupled-cluster theory is one of the most cost-effective methods for approximating the exact solution for the nonrelativistic electronic Schrödinger equation.[1,2] Coupled-cluster theory entails a hierarchy of approximations that can be systematically improved towards the exact quantum mechanical solution, providing a roadmap for the determination of highly accurate and reliable chemical properties.[3–7] The CCSD(T) method (coupled-cluster with single, double, and quasiperturbative triple excitations) has been found to be a cost-effective approach for the calculation of highly accurate thermochemical and kinetic data[5,8–16] as well as molecular properties based on energy derivatives (e.g., equilibrium structures, vibrational frequencies, and electrical properties).[4,17–21] The CCSD(T) model is therefore often referred to as the "gold standard in quantum chemistry."[22] It should be stressed, however, that this expression can be misleading since in some cases the CCSD(T) shows poor performance (most notably, but not limited to,[6,23] multireference systems).[3–7]

The basis set convergence of the CCSD(T) method has been extensively studied for energetic properties.[3,4,6–8,13,15,17,18,24–28] There has been substantial work[6,18,19,21,29–38] exploring the potential accuracy of CCSD(T) molecular structures relative to experimental reference values, typically including other energetic contributions (e.g., post-CCSD(T), core-valence, and relativistic effects). Fewer studies have been dedicated to the basis set convergence of molecular geometries. Studies of the basis set effects on the molecular structures have been predominantly limited to small species (molecules with at most two non-hydrogen atoms) and pathologically multireference systems such as halogen oxides.[36,37]

Heckert *et al.*[39] have explored the basis-set convergence of CCSD(T) equilibrium structures for a set of 17 small first-row molecules, namely HF, H_2O, $CH_2(^1A_1)$, NH_3, CH_4, CO, N_2, F_2, HCN, HNC, C_2H_2, CO_2, OH, CN, NH_2, $CH_2(^3B_1)$, and NO, relative to CCSD(T)-R12 reference values at the CBS limit. The basis set convergence of these geometries was found to be smooth. The mean-absolute deviations (MADs) relative to the CBS reference values were found to be 0.0008 (cc-pVQZ), 0.00033 (cc-pV5Z), and 0.00021 (cc-pV6Z) Å. They further reported that

a)E-mail: amir.karton@uwa.edu.au

CCSD(T)/cc-pV{5,6}Z extrapolations are required for target accuracies of 0.0001 Å, and that extrapolations using small basis sets are not recommended. Knizia *et al.*[38] considered the basis set convergence of explicitly correlated CCSD(T)-F12b calculations for a set of 13 first-row diatomics relative to CCSD(T)/aug-cc-pV{5,6}Z reference values. They found that the CCSD(T)-F12b/aug-cc-pV*n*Z level of theory generates results comparable in quality to the CCSD(T)/aug-cc-pV(*n*+2)Z level of theory. More recently, Feller *et al.*[40] explored a set of somewhat larger hydrocarbons (of up to C_6H_{12}) and C_2, again finding a gain in accuracy of about two "zetas" in the basis set when using explicitly correlated CCSD(T)-F12b calculations relative to the conventional CCSD(T) calculations.

In the present work, we investigate the basis set convergence of the CCSD(T) method for the molecular structures in the W4-11 database.[5] Excluding pathologically multireference systems (e.g., O_3, C_2, and BN) for which the CCSD(T) approximation breaks down, the W4-11 database includes 122 species (results for the multireference systems are provided in Figures S1 and S2 of the supplementary material). These species cover a broad spectrum of bonding situations with a range of single and multiple bonds that involve varying degrees of covalent and ionic characters. As such the W4-11 database constitutes an excellent benchmark set for analysis of basis set effects on the molecular structures. For most of the systems in the W4-11 database, we were able to obtain reference structures at the CCSD(T)/aug′-cc-pV(6+d)Z level of theory, whilst for larger molecules (with low spatial symmetries), we use CCSD(T)/aug′-cc-pV(5+d)Z reference geometries. Using this large and diverse set of accurate reference geometries, we attempt to answer questions such as the following:

1. What is the accuracy of bond distances calculated with the CCSD(T) method in conjunction with the cc-pV(*n*+d)Z, aug′-cc-pV(*n*+d)Z, and def2-*n*ZVPP basis sets?
2. What is the accuracy of bond distances calculated with the CCSD(T)-F12 method in conjunction with the cc-pV(*n*+d)Z-F12 basis sets?
3. Do different bond types (e.g., single, double, and triple bonds) exhibit different rates of basis set convergence?
4. Can we accelerate the basis set convergence of CCSD(T) bond distances using basis set extrapolations or even simple scaling factors?
5. To what extent does the accuracy of the reference geometries affect the molecular energies calculated at the CCSD(T)/CBS level of theory?

II. COMPUTATIONAL METHODS

All calculations were carried out on the Linux cluster of the Karton group at the University of Western Australia. All calculations were carried out using the MOLPRO program suite.[41,42] Most of the CCSD(T) geometry optimizations and single-point energy calculations were carried out with the correlation-consistent basis sets of Dunning and co-workers.[43–45] The notation A′V*n*Z indicates the combination of the standard correlation-consistent cc-pV*n*Z basis sets on hydrogen,[43] the aug-cc-pV*n*Z basis sets on first-row elements,[44] and the aug-cc-pV(*n*+d)Z basis sets on second-row elements.[45] The notation V*n*Z indicates the combination of the cc-pV*n*Z basis sets on hydrogen and first-row elements and the cc-pV(*n*+d)Z basis sets on second-row elements. Geometry optimizations were also carried out with the Weigend–Ahlrichs def2-TZVPP and def2-QZVPP basis sets.[46] The explicitly correlated CCSD(T)-F12b calculations[38,47] were carried out in conjunction with the V*n*Z-F12 basis sets of Peterson *et al.*[48]

III. RESULTS AND DISCUSSION

A. Overview of the molecules in the W4-11 database and reference geometries

The W4-11 database contains 122 small first- and second-row molecules (excluding the 16 pathologically multireference and 3 beryllium-containing species which are not considered in the present work).[49] Table I lists the molecules in this set, which will be referred to as the W4-11-GEOM dataset. The systems in the W4-11-GEOM dataset include 85 closed shell, 21 radical, 9 singlet carbene, and 7 triplet species. In terms of elemental composition, the dataset includes 88 first-row species (containing H and B—F), 17 second-row species (containing H and Al—Cl), and 17 mixed first- and second-row species (containing H, B—F, and Al—Cl atoms). Overall, the W4-11-GEOM dataset includes hydrogen-containing (82), hydrogen-free (40), organic (63), and inorganic (59) compounds.

Table II gives an overview of the types of bonds in the W4-11-GEOM dataset. Overall, it includes 246 symmetry unique bonds. Of these, 182 are single bonds, 49 are double bonds, and 15 are triple bonds. The set of 182 single bonds includes 117 H—X and 65 X—Y bonds (where X and Y are non-hydrogen atoms from the first and second rows of the periodic table). For the complete list of bonds in the W4-11-GEOM dataset, see Table S1 of the supplementary material.

We were able to optimize the geometries for a subset of 108 molecules at the CCSD(T)/A′V6Z level of theory (hereinafter referred to as the GEOM-AV6Z dataset). These include molecules with up to five non-hydrogen atoms such as BF_3, CF_4, C_2N_2, C_2F_2, F_2CO, F_2O_2, AlF_3, SiF_4, SO_3, Cl_2O_2, P_4, S_4, and $AlCl_3$ (Table I lists the molecules in the GEOM-AV6Z subset). The entire W4-11-GEOM database includes 14 CCSD(T)/A′V5Z geometries of organic molecules in addition to the 108 CCSD(T)/A′V6Z geometries in the GEOM-AV6Z subset. In Sections III B 1–III B 4, we use the GEOM-AV6Z dataset to evaluate the performance of the CCSD(T)/A′V*n*Z (n = D − 5), CCSD(T)/V*n*Z (n = D − 5), CCSD(T)/def2-*n*ZVPP (n = T, Q), and CCSD(T)-F12/V*n*Z-F12 (n = D, T) levels of theory. In Section III B 5, we use the larger W4-11-GEOM dataset to evaluate the performance of the CCSD(T) and CCSD(T)-F12 methods in conjunction with basis sets of up to quadruple-zeta quality.

TABLE I. Overview of the 122 molecules in the W4-11-GEOM database for which we were able to obtain CCSD(T)/A′V6Z or CCSD(T)/A′V5Z reference geometries.[a]

CCSD(T)/A′V6Z reference geometries[b]			
AlCl	CH_2 ($^3B_1$)	HCl	NH_3
$AlCl_3$	CH_2C	HCN	NO
AlF	CH_2F_2	HCNH	NO_2
AlF_3	CH_2NH	HCNO	O_2
AlH	CH_3	HCO	OCS
AlH_3	CH_3F	HCOF	OH
Allene	CH_4	HF	Oxirane
B_2H_6	Cl_2	HNC	Oxirene
BF	ClCN	HNCO	P_2
BF_3	ClF	HN_3	P_4
BH	ClO	HNO	PH_3
BH_3	CN	HOCl	S_2
BHF_2	CO	HOCN	S_2O
BN ($^3\Pi$)	CO_2	HOF	Si_2H_6
cis-HCOH	CS	HONC	SiF
cis-HONO	CS_2	HOO	SiF_4
cis-N_2H_2	Dioxirane	HOOH	SiH
C_2H_2	F_2	HS	SiH_3F
C_2H_4	F_2CO	Ketene	SiH_4
C_2H_6	C_2F_2	N_2	SiO
CCH	Glyoxal	N_2H	SO
CCl_2	H_2	N_2H_4	SO_2
CF	H_2CN	N_2O	SO_3
CF_2	H_2CO	NCCN	S_2H
CF_4	H_2O	NH	*trans*-HCOH
CH	H_2S	NH_2	*trans*-HONO
CH_2 ($^1A_1$)	HCCF	NH_2Cl	*trans*-N_2H_2
CCSD(T)/A′V5Z reference geometries			
Acetaldehyde	CH_2CH	Ethanol	Propene
Acetic acid	CH_2NH_2	Formic acid	Propyne
C_2H_3F	CH_3NH	Methanol	
C_2H_5F	CH_3NH_2	Propane	

[a]The entire set of 122 molecules will be referred to as the W4-11-GEOM dataset.
[b]The subset of 108 molecules for which we were able to obtain CCSD(T)/A′V6Z reference geometries will be referred to as the GEOM-AV6Z dataset.

TABLE II. Overview of the bonds in the W4-11-GEOM and GEOM-AV6Z datasets.[a]

Bond type	Number[b]	Comment
W4-11-GEOM dataset (122 molecules, 246 unique bonds)		
H—X	117	X = H, B—F, Al—Cl
X—Y	65	X, Y = B—F, Al—Cl
X=Y	49	X, Y = C, N, O, Si, S, Cl
X≡Y	15	X, Y = B, C, N, P
GEOM-AV6Z dataset (108 molecules, 181 unique bonds)		
H—X	75	X = H, B—F, Al—Cl
X—Y	49	X, Y = B—F, Al—Cl
X=Y	43	X, Y = C, N, O, Si, S, Cl
X≡Y	14	X, Y = B, C, N, P

[a]For the complete list of bonds see Table S1 of the supplementary material.
[b]Number of unique bonds (i.e., not equivalent by symmetry, for example, CH_3Cl has two unique bonds).

B. Basis set convergence of bond distances

1. Basis set convergence of conventional CCSD(T) calculations against CCSD(T)/A′V6Z reference geometries

Table III gives the error statistics for the CCSD(T)/A′V*n*Z (n = D – 5), CCSD(T)/V*n*Z ($n = D - 5$), CCSD(T)/def2-*n*ZVPP (n = T, Q), and CCSD(T)-F12/V*n*Z-F12 (n = D, T) levels of theory relative to the CCSD(T)/A′V6Z bond distances for the GEOM-AV6Z dataset (which includes 108 molecules and 181 unique bonds). We begin by making three general observations:

- All the levels of theory tend to systematically overestimate the bond lengths as evident from MSD ≈ MAD. The extent of overestimation decreases with the size of the basis set.
- The V*n*Z basis sets show similar (or even slightly better) performance than the A′V*n*Z basis sets.
- The def2-*n*ZVPP basis sets show similar performance to the V*n*Z basis sets.
- The CCSD(T)-F12/V*n*Z-F12 level of theory shows similar performance to the CCSD(T)/A′V(n+2)Z level of theory at a significantly reduced computational cost. This is in agreement with previous studies,[38,40] albeit the current study includes a more diverse set of molecules consisting of both first- and second-row elements.

Let us first consider the basis set convergence with the orbital V*n*Z and A′V*n*Z basis sets in conventional CCSD(T) calculations. The VDZ basis set has been found to significantly overestimate bond lengths in MP2 and CCSD(T) geometry optimizations due to the lack of higher angular momentum polarization functions.[32,35,50–53] Our results confirm these observations over a very large and diverse set of chemical bonds. In particular, the CCSD(T)/VDZ level of theory systematically overestimates the CCSD(T)/A′V6Z bond lengths by significant amounts and results in a mean-signed deviation (MSD) of +0.0174 Å and a root-mean-square deviation (RMSD) of 0.0196 Å. The VDZ basis set shows particularly poor performance for bonds involving second-row atoms. For example, overestimations ranging between 0.04 and 0.06 Å are seen for the bonds involving second-row atoms in SiF_4, NH_2Cl, P_4, SiH_3F, Cl_2, SiF, HOCl, ClF, and ClO. This poor performance for the second-row systems is also reflected in the percentage errors. For instance, the largest percentage errors are obtained for SiF (2.8), Cl—O bond in HOCl (2.8), ClF (3.4), F_2 (3.4), and ClO (3.9%). We note, however, that removing the 35 second-row systems from the training set only results in a minor improvement in performance.[54] Namely, the RMSD for the 73 first-row systems with 138 unique bonds is 0.0167 Å (see Table S2 of the supplementary material for additional error statistics for the subset of first-row systems). Notably, the addition of diffuse functions does not bring succor and the A′VDZ basis set results in similar performance to the VDZ basis set (Table III).

Increasing the basis set size from a double-zeta to triple-zeta reduces the RMSD by a factor of ∼4. The VTZ

TABLE III. Overview of the basis set convergence of the CCSD(T) and CCSD(T)-F12 methods for the 181 unique bond lengths in the GEOM-AV6Z dataset (Å). The reference values are CCSD(T)/A′V6Z bond distances.[a]

Basis set	RMSD	MAD	MSD	LND	LPD
VDZ	0.0196	0.0174	0.0174	N/A	0.0611 (O═Cl)
VTZ	0.0050	0.0037	0.0037	−0.0026 (C—F)	0.0172 (Cl—Cl)
VQZ	0.0015	0.0010	0.0009	−0.0016 (C—F)	0.0060 (O═Cl)
V5Z	0.0004	0.0003	0.0001	−0.0005 (C—F)	0.0017 (Cl—Cl)
A′VDZ	0.0222	0.0200	0.0200	N/A	0.0525 (Cl—Cl)
A′VTZ	0.0059	0.0048	0.0048	N/A	0.0185 (Cl—Cl)
A′VQZ	0.0017	0.0013	0.0013	N/A	0.0060 (Cl—Cl)
A′V5Z	0.0005	0.0003	0.0003	−0.0001 (H—Cl)	0.0018 (Cl—Cl)
VDZ-F12	0.0019	0.0015	0.0014	−0.0030 (Al—Cl)	0.0081 (N—O)
VTZ-F12	0.0006	0.0005	0.0005	−0.0006 (Al—Cl)	0.0026 (F—F)
Def2-TZVPP	0.0047	0.0034	0.0034	−0.0023 (H—Cl)	0.0181 (Cl—Cl)
Def2-QZVPP	0.0015	0.0010	0.0009	−0.0006 (N—H)	0.0056 (Cl—Cl)

[a]RMSD = root-mean-square deviation, MAD = mean absolute deviation, MSD = mean signed deviation, LND = largest negative deviation, and LPD = largest positive deviation (the molecules associated with the LND and LPD are given in parentheses).

basis sets result in RMSDs of 0.0050 Å (cf., an RMSD of 0.02 Å for the VDZ basis set). The VTZ basis set still tends to systematically overestimate the bond lengths, where particularly large deviations of 0.015–0.017 Å are obtained for the bonds involving second-row atoms in S_2O, ClO, S_2H, and Cl_2. We note that upon removing the second-row systems from the training set, the RMSD is reduced to 0.0036 Å (Table S2 of the supplementary material). Similar to the performance of the VDZ and A′VDZ basis sets, the addition of diffuse functions does not lead to an improvement in performance, namely, the CCSD(T)/A′VTZ level of theory results in an RMSD and MAD of 0.0059 and 0.0048 Å, respectively, for the entire GEOM-AV6Z dataset (Table III).

The CCSD(T)/VQZ level of theory is used for optimizing geometries in highly accurate composite theories such as the W4[28,55,56] and HEAT[57–59] thermochemical protocols.[3] Here the performance of this level of theory is assessed against a large and diverse set of geometries. Relative to our CCSD(T)/A′V6Z reference values, the CCSD(T)/VQZ level of theory results in a respectable RMSD of 0.0015 Å and a MAD of 0.0010 Å. Again, the similarity of the MSD of 0.0009 Å to the MAD indicates that the bond lengths are still systematically overestimated. The largest overestimations of 0.004–0.006 Å are obtained for the bonds involving second-row atoms in CCl_2, P_4, S_2, NH_2Cl, HOCl, S_2O, S_2H, Cl_2, and ClO. Removing the 35 second-row systems from the training set has a noticeable effect on the performance, namely, the error statistics are reduced to RMSD = 0.0009, MAD = 0.0007, and MSD = 0.0005 Å (Table S2 of the supplementary material). Further, the largest overestimations are reduced to below 0.003 Å, namely, 0.0029 (O—O bond in dioxirane) and 0.0027 (F_2) Å.

The V5Z and A′V5Z basis sets both yield bond distances that are very close to the CCSD(T)/A′V6Z bond lengths. The V5Z basis set results in an RMSD of 0.0004 Å and a MAD of 0.0003 Å. An MSD of 0.0001 indicates that there is a very small bias towards overestimating the bond distances. The largest deviation, an overestimation of 0.0017 Å, is obtained for Cl_2. Upon eliminating the second-row systems, the RMSD and MAD are reduced to 0.0003 and 0.0002 Å, respectively.

Lastly, we note that the def2-TZVPP and def2-QZVPP basis sets result in overall error statistics that are very similar to those obtained for the VTZ and VQZ basis sets, respectively (Table III).

2. Basis set convergence of explicitly correlated CCSD(T)-F12 calculations

It is well established that inclusion of "geminal" terms that explicitly depend on the interelectronic distance into the wavefunction drastically accelerates the basis set convergence.[60–62] Experience with CCSD-F12 energy calculations has shown that typically the gain amounts to 1–2 angular momenta.[48,60–63] Table III gives an overview of the performance of the CCSD(T)-F12/V*n*Z-F12 level of theory (*n* = D, T) for the bond distances in the GEOM-AV6Z dataset.

The CCSD(T)-F12/VDZ-F12 level of theory results in RMSD = 0.0019, MAD = 0.0015, and MSD = 0.0014 Å. This performance is significantly better than that of the CCSD(T)/VTZ level of theory (RMSD = 0.0037) and is comparable to that of the CCSD(T)/VQZ level of theory (RMSD = 0.0015 Å, Table III). The CCSD(T)-F12/VDZ-F12 level of theory tends to systematically overestimate the bond distances. However, whilst the largest overestimations for the CCSD(T)/VTZ and CCSD(T)/VQZ levels of theory are obtained for bond distances involving second-row elements, the largest deviations for the CCSD(T)-F12/VDZ-F12 level of theory are obtained for first-row systems such as F_2 (0.007), the N—O bond in *trans*-HONO (0.008), and the O—O bond in dioxirane (0.008 Å). Consequently the error statistics for the CCSD(T)-F12/VDZ-F12 level of theory for the 73 first-row systems are very similar to those obtained for the entire GEOM-AV6Z dataset (Table S2 of the supplementary material).

We note that the CCSD(T)-F12/VDZ-F12 level of theory represents a significant saving in computational resources compared to the CCSD(T)/VQZ level of theory. Table IV gives relative central processing unit (CPU) times and disk space used by these levels of theory for two medium-sized systems

TABLE IV. Computational resources used for the CCSD(T)/V*n*Z and CCSD(T)-F12/V*n*Z-F12 single-point energy calculations for naphthalene and anthracene.[a]

	Naphthalene		Anthracene	
Basis set	Time[b]	Disk[c]	Time[b]	Disk[c]
VDZ	1	0.6	1	2
VTZ	25	10	22	31
VQZ	253	111	204	335
V5Z	1977	906	N/A	N/A
VDZ-F12	31	7	20	25
VTZ-F12	191	68	138	218

[a]All the calculations ran on 8 cores of otherwise idle dual Intel Xeon E5-2670v2 systems (3.1 GHz, 256 GB RAM).
[b]The numbers given are the ratios between the CPU time for the different levels of theory and that for the CCSD(T)/VDZ level of theory.
[c]Scratch disk usage in GB.

(naphthalene and anthracene). For anthracene ($C_{14}H_{10}$), a single-point energy CCSD(T)-F12/VDZ-F12 calculation which entails 510 basis functions in D_{2h} symmetry requires only 10% (!!) of the time required for the CCSD(T)/VQZ calculation which involves more than twice the number of basis functions (1070 basis functions). In terms of the disk usage, the CCSD(T)-F12/VDZ-F12 calculation uses 25 GB of scratch disk, whilst the CCSD(T)/VQZ calculation uses a total of 335 GB of disk space. Considering these very significant savings in terms of CPU time and disk usage for obtaining molecular geometries of similar quality, we recommend using the CCSD(T)-F12/VDZ-F12 level of theory rather than CCSD(T)/VQZ when high-quality geometries are needed (e.g., in thermochemical protocols such as W4lite, W4, and W4-F12).[55,56] Finally, we note that the CCSD(T)-F12/VDZ-F12 calculations have a similar computational cost to that of CCSD(T)/VTZ calculations (Table IV); however, the former level of theory leads to much more accurate bond distances (Table III).

The CCSD(T)-F12/VTZ-F12 level of theory results in an RMSD of 0.0006 Å, which is comparable to that of the CCSD(T)/V5Z level of theory (RMSD = 0.0004 Å, Table III). Similar to the VDZ-F12 basis set, the largest overestimations for the VTZ-F12 basis set are obtained for bonds involving first-row atoms, such as F_2 (0.003) and the O—O bond in dioxirane (0.002 Å). Thus, it seems like the CCSD(T)-F12 method shows a more balanced performance for first- and second-row systems than conventional CCSD(T) calculations in conjunction with double- and triple-zeta basis sets. Table IV illustrates the significant computational savings of the CCSD(T)-F12/VTZ-F12 level of theory compared to the CCSD(T)/V5Z level of theory. For a medium-sized system such as naphthalene, the CCSD(T)-F12/VTZ-F12 calculation uses about 10% of the CPU time and disk space as the CCSD(T)/V5Z calculation.

3. Accelerating the basis set convergence of CCSD(T) and CCSD(T)-F12 calculations

In Section III B 1, we have seen that the VDZ and A′VDZ basis sets systematically and severely overestimate the bond lengths. For *all* the 181 bonds in the GEOM-AV6Z database, these basis sets overestimate the CCSD(T)/A′V6Z bond distances by amounts ranging from 0.0017 (C—F bond in CH_3F) to 0.0611 (Cl═O) Å. For both the VDZ and A′VDZ basis sets, we find that there is a high statistical correlation with the CCSD(T)/A′V6Z bond distances. Namely, the squared correlation coefficient (R^2) is equal to 0.9991 (VDZ) and 0.9992 (A′VDZ). For the larger basis sets, we obtain R^2 values >0.99991, for additional details see Table S3 of the supplementary material. The high statistical correlation between the bond distances obtained at the CCSD(T)/A′V6Z and CCSD(T)/VDZ levels of theory is illustrated in Figure 1 and suggests that simple linear scaling of the bond distances may improve the accuracy.

Table V gives an overview of the performance of the scaled CCSD(T)/V*n*Z and CCSD(T)/A′V*n*Z bond distances in conjunction with basis sets of up to quadruple-zeta quality. Let us begin with the performance of the double-zeta quality basis sets in conventional CCSD(T) calculations. The CCSD(T)/VDZ and CCSD(T)/A′VDZ levels of theory result in a large RMSD of ~0.02 Å (Table III). Upon scaling of the bond distances by empirical factors of 0.9865 (VDZ) and 0.9842 (A′VDZ), the RMSDs are reduced by about 60% to 0.0080 (scaled VDZ) and 0.0076 (scaled A′VDZ) Å.

Moving to the triple-zeta quality basis sets, the RMSD for the unscaled bond lengths is 0.0050 (VTZ) and 0.0059 (A′VTZ) Å (Table III). These relatively large RMSDs are reduced by about 50% upon scaling. Namely, they are 0.0029 (scaled VTZ) and 0.0027 (scaled A′VTZ) Å. Again a near-zero MSD of −0.0003 Å obtained for both basis sets indicates that scaling eliminates the systematic bias towards overestimating the bond lengths. For comparison, the unscaled VTZ and A′VTZ results lead to MSDs which are more than one order of magnitude larger (namely, 0.004 and 0.005 Å, respectively).

The unscaled CCSD(T)/VQZ and CCSD(T)/A′VQZ methods lead to respectable RMSDs of 0.0015 and 0.0017 Å, respectively. Scaling of the VQZ bond distances leads to a small improvement in performance (RMSD = 0.0011 Å); however, scaling the A′VQZ bond distances reduces the RMSD by nearly 50% (RMSD = 0.0009 Å). Finally, we

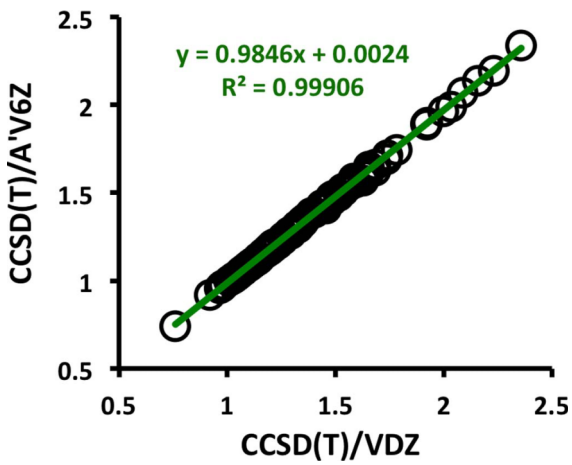

FIG. 1. Linear correlation between CCSD(T)/A′V6Z and CCSD(T)/VDZ bond distances (in Å) for the 181 bond lengths in the GEOM-AV6Z dataset.

TABLE V. Overview of the performance of scaled and extrapolated CCSD(T) and CCSD(T)-F12 bond distances for the 181 unique bonds in the GEOM-AV6Z dataset (Å). The reference values are CCSD(T)/A′V6Z bond distances.[a]

	Basis sets	α[b]	RMSD	MAD	MSD	LND	LPD
Scaled	VDZ	0.9865	0.0080	0.0056	−0.0001	−0.0197 (H—Al)	0.0390 (O═Cl)
	VTZ	0.9969	0.0029	0.0022	−0.0003	−0.0069 (C—F)	0.0109 (Cl—Cl)
	VQZ	0.9992	0.0011	0.0008	−0.0001	−0.0027 (C—F)	0.0048 (O═Cl)
	A′VDZ	0.9842	0.0076	0.0055	−0.0004	−0.0228 (H—Al)	0.0256 (O═Cl)
	A′VTZ	0.9960	0.0027	0.0020	−0.0003	−0.0050 (H—S)	0.0105 (Cl—Cl)
	A′VQZ	0.9989	0.0009	0.0007	−0.0001	−0.0016 (H—Al)	0.0038 (Cl—Cl)
	VDZ-F12	0.9989	0.0013	0.0008	0.0000	−0.0053 (Al—Cl)	0.0065 (N—O)
	Def2-TZVPP	0.9971	0.0028	0.0020	−0.0003	−0.0060 (H—Cl)	0.0123 (Cl—Cl)
	Def2-QZVPP	0.9992	0.0010	0.0007	−0.0001	−0.0015 (H—N)	0.0040 (O═Cl)
Extrap.[c]	V{D,T}Z	4.0	0.0033	0.0025	0.0003	−0.0060 (F—Si)	0.0120 (S—S)
	V{T,Q}Z	4.5	0.0007	0.0005	−0.0002	−0.0014 (H—F)	0.0024 (O═Cl)
	A′V{D,T}Z	3.5	0.0026	0.0020	0.0001	−0.0046 (H—H)	0.0091 (S—S)
	A′V{T,Q}Z	4.4	0.0004	0.0003	−0.0001	−0.0008 (H—F)	0.0015 (O═Cl)
	Def2-{T,Q}	4.3	0.0006	0.0004	−0.0001	−0.0015 (B—F)	0.0022 (H—C)

[a]Footnote a to Table III applies here.
[b]Scaling factors and extrapolation exponents used in the two-point extrapolations, these are optimized to minimize the RMSD over the set of 181 bond lengths in the GEOM-AV6Z dataset.
[c]Extrapolated using the $D_L = D_\infty + A/L^\alpha$ extrapolation formula (where D is the bond distance and L is the highest angular momentum represented in the basis set).

note that scaling the def2-TZVPP and def2-QZVPP bond distances results in overall error statistics that are very similar to those obtained for the scaled VTZ and VQZ basis sets (Table V).

What about extrapolating the bond lengths? This approach has been previously found to accelerate the basis set convergence in conjunction with sufficiently large basis sets.[6,30,31,33,35,39] Here we test this approach for a wider and more diverse set of bond lengths. We consider a two-point extrapolation formula of the form $D(L) = D_\infty + A/L^\alpha$ where D is the bond distance, L is the highest angular momentum present in the basis set, and α is an extrapolation exponent which is optimized to minimize the RMSD over the GEOM-AV6Z dataset. These results are presented in Table V. Extrapolating from the V{D,T}Z or A′V{D,T}Z basis set pairs results in RMSDs of 0.0033 and 0.0026 Å, respectively. This does not represent an improvement over simple scaling (Table V) and we do not recommend using basis set extrapolations that involve double-zeta quality basis sets. Extrapolating from the V{T,Q}Z or A′V{T,Q}Z basis set pairs results in RMSDs of 0.0007 and 0.0004 Å, respectively. This performance represents an improvement over simple linear scaling of the VQZ (RMSD = 0.0011) and A′VQZ results (RMSD = 0.0009 Å). Thus, extrapolations from triple-zeta and quadruple-zeta quality basis sets can be considered as a viable alternative to simple scaling. Extrapolations from the def2-TZVPP and def2-QZVPP basis sets result in overall error statistics that are similar to those obtained from the V{T,Q}Z extrapolations (Table V).

4. Basis set convergence for subsets of the GEOM-AV6Z dataset

In this subsection, we examine more closely the basis set convergence for the H—X, X—Y, X═Y, and X≡Y bonds in the GEOM-AV6Z dataset (Table II). Table VI gathers the RMSDs obtained for the H—X, X—Y, X═Y, and X≡Y subsets. For a comprehensive overview of the error statistics, see Table S4 of the supplementary material.

A few interesting features emerge from Table VI. First, the performance of all the considered basis sets is significantly better for the H—X bonds compared to the performance for the X—Y, X═Y, and X≡Y bonds. This is not surprising since bonds involving hydrogen are expected to converge faster to the basis set limit than the X—Y, X═Y, and X≡Y bonds. Second, all the considered basis sets show similar performance for the X═Y and X≡Y bonds. This result is somewhat counterintuitive, particularly for the double- and triple-zeta basis sets, and is attributed to the fact that the X═Y subset includes many bonds involving second-row elements, whilst the X≡Y subset includes mainly first-row elements. Table S5 of the supplementary material gives the RMSDs over the bonds involving only first-row elements. After removing the bonds involving second-row elements from both subsets, we

TABLE VI. Overview of the basis set convergence of the CCSD(T) and CCSD(T)-F12 methods for the 75 H—X, 49 X—Y, 43 X═Y, and 14 X≡Y bonds in the GEOM-AV6Z dataset (RMSDs, in Å).[a]

Basis set	H—X	X—Y	X═Y	X≡Y
VDZ	0.0148	0.0245	0.0199	0.0215
VTZ	0.0014	0.0071	0.0058	0.0057
VQZ	0.0004	0.0022	0.0017	0.0013
V5Z	0.0002	0.0006	0.0005	0.0003
A′VDZ	0.0137	0.0304	0.0229	0.0228
A′VTZ	0.0022	0.0086	0.0064	0.0058
A′VQZ	0.0005	0.0026	0.0020	0.0016
A′V5Z	0.0001	0.0007	0.0005	0.0004
VDZ-F12	0.0007	0.0029	0.0020	0.0019
VTZ-F12	0.0004	0.0009	0.0006	0.0005

[a]For a comprehensive overview of the error statistics obtained for the H—X, X—Y, X═Y, and X≡Y subsets, see Table S4 of the supplementary material.

obtain RMSDs of 0.0158 (X═Y bonds) and 0.0211 (X≡Y) for the VDZ basis set, and 0.0045 (X═Y) and 0.0050 (X≡Y) for the VTZ basis set. Thus, it appears that the triple bonds converge more slowly to the basis set limit compared to the double bonds. It should be noted that for basis sets of quadruple-zeta quality (and higher), the RMSDs for the X═Y and X≡Y bonds are practically the same (see Table S5 of the supplementary material). Similar trends are observed for the A′V*n*Z basis sets.

The subset of 49 X—Y bonds seems to be particularly challenging for the conventional CCSD(T) calculations involving double- and triple-zeta quality basis sets. The VDZ and A′VDZ basis sets result in large RMSDs of 0.025 and 0.030 Å, respectively. The VTZ and A′VTZ basis sets result in RMSDs larger than 0.007 Å. The VQZ, A′VQZ, and cost-effective VDZ-F12 basis sets result in RMSDs of 0.002–0.003 Å. The V5Z and A′V5Z basis sets result in small RMSDs of about 0.0006 Å (Table VI). These RMSDs are reduced after removing the bonds involving second-row elements (see Table S5 of the supplementary material).

5. Basis set convergence for the entire W4-11-GEOM dataset

The GEOM-AV6Z dataset contains systems with up to 5 non-hydrogen atoms. The largest systems in this database are highly symmetric, mostly inorganic systems, such as BF_3, CF_4, C_2N_2, C_2F_2, F_2CO, F_2O_2, AlF_3, SiF_4, SO_3, Cl_2O_2, P_4, S_4, and $AlCl_3$. The largest organic systems that are present in the GEOM-AV6Z dataset are allene, ketene, glyoxal, oxirene, oxirane, and dioxirane. It is therefore of interest to include larger organic systems in the dataset. For this purpose, we generate an additional test set (W4-11-GEOM) with the remaining systems from the W4-11 dataset (Table I). The additional 14 systems are relatively large organic systems involving first-row elements (namely, acetic acid, acetaldehyde, C_2H_3F, C_2H_5F, CH_2CH, CH_2NH_2, CH_3NH, CH_3NH_2, ethanol, formic acid, methanol, propane, propene, and propyne). We were able to optimize the geometries of these systems at the CCSD(T)/A′V5Z level of theory. As we saw in Section III B 1, the performance of the A′V5Z basis set for bond involving only first-row elements is very close to that of the A′V6Z basis set (RMSD = 0.0003 Å, Table S2 of the supplementary material). Due to the use of 14 CCSD(T)/A′V5Z reference geometries, we will only assess the performance of basis sets of up to quadruple-zeta quality against this dataset. Overall, the W4-11-GEOM datasets contain 122 molecules involving 246 unique bond distances. Table VII gathers the error statistics for the W4-11-GEOM dataset.

Generally, the error statistics obtained for the W4-11-GEOM database (Table VII) are similar to those obtained for the GEOM-AV6Z dataset (Tables III and V), and thus, our main conclusions in Sections III B 1–III B 4 remain largely unchanged. We note that inclusion of the 14 organic systems improves the performance for all the considered levels of theory. In particular, the RMSDs for the entire W4-11-GEOM database are ~0.02 (VDZ and A′VDZ), ~0.005 (VTZ and A′VTZ), and ~0.001 (VQZ and A′VQZ) Å. For the scaled bond distances, we obtain RMSDs of 0.008 (VDZ and A′VDZ), 0.003 (VTZ and A′VTZ), and ~0.001 (VQZ and A′VQZ) Å. Thus, again we see that scaling (in particular of the VDZ and A′VDZ distances) results in significant improvements in performance at no additional computational cost.

6. Energetic consequences of the level of theory used for optimizing the geometries

In Section III B, we have shown that the basis set incompleteness error can result in RMS deviations from CCSD(T)/A′V6Z bond distances ranging from ~0.02 (VDZ and A′VDZ) to ~0.002 (VQZ and A′VQZ), to ~0.0005 (V5Z and A′V5Z) Å. We note that errors in the bond distances (whether they are overestimations or underestimations) will always lead to overestimation of the molecular energies

TABLE VII. Overview of the performance of the CCSD(T) and CCSD(T)-F12 methods for the 246 unique bonds in the W4-11-GEOM dataset (Å). The reference values are 108 CCSD(T)/A′V6Z and 14 CCSD(T)/A′V5Z bond distances (see Table I).[a]

	Basis sets	RMSD	MAD	MSD	LND	LPD
	VDZ	0.0184	0.0165	0.0165	N/A	0.0611 (O═Cl)
	VTZ	0.0045	0.0033	0.0032	−0.0029 (C—F)	0.0172 (Cl—Cl)
	VQZ	0.0013	0.0008	0.0007	−0.0018 (C—F)	0.006 (O═Cl)
	A′VDZ	0.0207	0.0188	0.0188	N/A	0.0525 (Cl—Cl)
	A′VTZ	0.0053	0.0043	0.0043	N/A	0.0185 (Cl—Cl)
	A′VQZ	0.0015	0.0011	0.0011	N/A	0.006 (Cl—Cl)
	VDZ-F12	0.0017	0.0013	0.0013	−0.003 (Al—Cl)	0.0081 (N—O)
Scaled[b]	VDZ	0.0075	0.0051	−0.0006	−0.0197 (H—Al)	0.039 (O═Cl)
	VTZ	0.0027	0.0022	−0.0007	−0.0073 (C—F)	0.0109 (Cl—Cl)
	VQZ	0.0011	0.0008	−0.0003	−0.0029 (C—F)	0.0048 (O═Cl)
	A′VDZ	0.0070	0.0051	−0.0012	−0.0228 (H—Al)	0.0256 (O═Cl)
	A′VTZ	0.0026	0.0020	−0.0007	−0.005 (H—S)	0.0105 (Cl—Cl)
	A′VQZ	0.0009	0.0007	−0.0003	−0.0016 (H—Al)	0.0038 (Cl—Cl)
	VDZ-F12	0.0012	0.0007	−0.0001	−0.0053 (Al—Cl)	0.0065 (N—O)

[a]Footnote a to Table III applies here.
[b]The scaling factors are given in Table V.

(or underestimation of the atomization energies) relative to those obtained at the CCSD(T)/CBS equilibrium geometries. However, a number of important questions arise as follows: (i) by how much an RMSD of say 0.02, 0.002, or 0.0005 Å in the bond distances affects the CCSD(T)/CBS molecular energies? (ii) Is the geometry effect going to increase with the molecular size (e.g., when going from diatomics, to triatomics, and to tetra-atomics)? These questions are particularly relevant when deciding which reference geometry to use in highly accurate composite *ab initio* methods such as W*n*,[8,28,55,56] HEAT,[57–59] and Feller-Peterson-Dixon (FPD)[4,6,17,64] theories.[3] Here we will address these questions in the context of a large and diverse set of molecules.

We calculate the molecular energies at the CCSD(T)/CBS level of theory for the 108 molecules in the GEOM-AV6Z database using the following reference geometries: CCSD(T)/V*n*Z (n = D − 5); CCSD(T)/A′V*n*Z (n = D − 5); and CCSD(T)-F12/V*n*Z-F12 (n = D, T). The single-point CCSD(T)/CBS energies are calculated using W2-F12 theory.[8] An overview of these results is given in Table VIII, whilst individual errors for all the molecules are given in Table S6 of the supplementary material.

Calculating the W2-F12 energies using CCSD(T)/VDZ and CCSD(T)/A′VDZ reference geometries leads to very large deviations from W2-F12 energies calculated using CCSD(T)/A′V6Z geometries. In particular, the RMSDs are 2.4 (VDZ) and 3.0 (A′VDZ) kJ mol^{-1}, and the largest deviations exceed 10 (!!) kJ mol^{-1}. Such large errors far exceed the intrinsic accuracy of highly accurate composite *ab initio* methods,[3] and CCSD(T)/VDZ and CCSD(T)/A′VDZ geometries should not be used for these purposes.

Moving to the triple-zeta quality basis sets, we obtain an RMSD of ∼0.2 kJ mol^{-1} for the CCSD(T)/VTZ and CCSD(T)/A′VTZ geometries. These geometries should not be used in composite *ab initio* methods that attempt to approximate the full configuration interaction (FCI) CBS energy (e.g., HEAT, W4, and W4-F12), since these theories are capable of predicting atomization energies with 95% (2σ) confidence intervals narrower (or even significantly narrower) than 1 kJ mol^{-1} (see Table 2 of Ref. 3 for more details). On the other hand, for composite *ab initio* methods that approximate the CCSD(T)/CBS energy (e.g., W*n* and W*n*-F12, n = 1, 2) which are capable of 95% confidence intervals narrower than ∼1 kcal mol^{-1}, one can consider using VTZ and A′VTZ geometries. Nevertheless, it should be stressed that for a similar computational cost, one can obtain CCSD(T)-F12/VDZ-F12 geometries which lead to much better performance (*vide infra*).

The CCSD(T)/VQZ level of theory, which is used for optimizing geometries in the W4 and HEAT thermochemical protocols, leads to a near-zero RMSD of 0.02 kJ mol^{-1}. This confirms that this level of theory is adequate for optimizing the geometries in these highly accurate composite theories. The CCSD(T)/A′VQZ level of theory leads to a similar RMSD of 0.03 kJ mol^{-1}. In both cases the largest deviations (of ∼0.1 kJ mol^{-1}) are obtained for SO_3.

An important finding is that the CCSD(T)-F12/VDZ-F12 geometries lead to a similar performance to that of the CCSD(T)/VQZ and CCSD(T)/A′VQZ geometries, at a significantly reduced computational cost (Table IV). In particular, the CCSD(T)-F12/VDZ-F12 geometries result in an RMSD of 0.03 kJ mol^{-1} and a largest deviation of 0.1 kJ mol^{-1} (CF_4). Thus, we recommend using this economical level of theory for optimizing geometries in composite theories such as HEAT and W4.

Finally, we note that quintuple-zeta quality basis sets lead to RMSD of below 0.003 kJ mol^{-1} and largest deviations below 0.02 kJ mol^{-1}.

What about the dependence of the geometry effect on the size of the molecule? Inspection of Table S6 (supplementary material) reveals that, in general, there is an increase in the geometry effect with the number of non-hydrogen atoms in the system. For example, for the CCSD(T)/VDZ geometries, we obtain RMSDs of 0.9 (over the 21 systems with one non-hydrogen atom), 2.1 (over the 48 systems with two non-hydrogen atoms), 2.5 (over the 27 systems with three non-hydrogen atoms), and 3.3 (over the 9 systems with four non-hydrogen atoms). Similarly, for the CCSD(T)/VTZ geometries, we obtain RMSDs of 0.03 (one non-hydrogen atom), 0.1 (two non-hydrogen atoms), 0.2 (three non-hydrogen atoms), and 0.3 (four non-hydrogen atoms).

TABLE VIII. Effect of CCSD(T) and CCSD(T)-F12 reference geometry on molecular energies calculated at the CCSD(T)/CBS level of theory for the 108 molecules in the GEOM-AV6Z database (kJ mol^{-1}).[a,b]

Basis sets	RMSD	MAD	MSD	LPD
VDZ	2.38	1.90	1.90	10.39 (SiF_4)
VTZ	0.17	0.13	0.13	0.62 (P_4)
VQZ	0.02	0.01	0.01	0.07 (SO_3)
V5Z	0.002	0.001	0.001	0.008 (SO_3)
A′VDZ	3.01	2.42	2.42	10.42 (SO_3)
A′VTZ	0.22	0.17	0.17	0.92 (SO_3)
A′VQZ	0.03	0.02	0.02	0.13 (SO_3)
A′V5Z	0.003	0.002	0.002	0.017 (SO_3)
VDZ-F12	0.03	0.02	0.02	0.10 (CF_4)
VTZ-F12	0.004	0.003	0.003	0.015 (SO_3)

[a]Footnote a to Table III applies here.
[b]The reference values are non-relativistic, valence CCSD(T)/CBS values from W2-F12 theory calculated with CCSD(T)/A′V6Z reference geometries.

IV. CONCLUSIONS

We have optimized reference geometries for a diverse set of 108 molecules at the CCSD(T)/A′V6Z level of theory. This set includes inorganic species with up to five non-hydrogen atoms (e.g., BF_3, F_2O_2, AlF_3, SiF_4, SO_3, Cl_2O_2, P_4, S_4, and $AlCl_3$) as well as organic compounds of similar size (e.g., allene, ketene, glyoxal, oxirene, oxirane, dioxirane, CF_4, C_2N_2, C_2F_2, and F_2CO). Overall, the set includes a total of 181 unique bonds: 75 H—X, 49 X—Y, 43 X=Y, and 14 X≡Y bonds (where X and Y are first- and second-row atoms). We use these CCSD(T)/A′V6Z reference geometries to examine the basis set convergence of the CCSD(T) method with the V*n*Z and A′V*n*Z basis sets (n = D, T, Q, 5), def2-*n*ZVPP basis sets (n = T, Q), and the CCSD(T)-F12 method with the

V*n*Z-F12 basis sets (n = D, T). Our main conclusions can be summarized as follows:

- In CCSD(T) calculations, the RMSD over the bond distances is reduced by a factor of 3–4 with each additional 'zeta' in the basis set. For example, the RMSD for the CCSD(T)/V*n*Z levels of theory is 0.0196 (VDZ), 0.0050 (VTZ), 0.0015 (VQZ), and 0.0004 (V5Z) Å. A similar trend is obtained for the A′V*n*Z basis sets.
- An important finding is that the explicitly correlated CCSD(T)-F12/V*n*Z-F12 levels of theory show similar performance to the CCSD(T)/A′V(n+2)Z levels of theory. For instance, the CCSD(T)-F12/VDZ-F12 and CCSD(T)/VQZ levels of theory attain RMSDs of 0.0019 and 0.0015 Å, respectively. Similarly, the CCSD(T)-F12/VTZ-F12 and CCSD(T)/V5Z levels of theory attain RMSDs of 0.0006 and 0.0004 Å, respectively. In both cases, the CCSD(T)-F12 calculations use about 10% of the CPU time and disk space as the conventional CCSD(T) calculations. Given these findings it seems pointless to use conventional CCSD(T) calculations for geometry optimizations. We recommend replacing the CCSD(T)/VQZ geometries used in highly accurate composite *ab initio* theories (e.g., W4 and HEAT) with computationally more economical CCSD(T)-F12/VDZ-F12 geometries.
- Due to systematic errors in conventional CCSD(T) calculations, we find that simple scaling of the bond distances is an effective way to improve the performance at no additional computational cost. This is particularly true for the smaller double-zeta and triple-zeta quality basis sets. For example, the RMSD for the unscaled CCSD(T)/VDZ bond distances (0.0196 Å) is reduced by about 60% upon scaling (0.0080 Å). Thus, in cases where a CCSD(T)-F12/VDZ-F12 geometry optimization is computationally too expensive, we recommend using scaled CCSD(T)/VDZ bond distances.
- Finally, we note that the use of CCSD(T)/VDZ geometries in composite *ab initio* theories (such as W1-F12 and W2-F12) leads to an RMSD of over 2.4 kJ mol^{-1} relative to the use of CCSD(T)/A′V6Z geometries. This RMSD is reduced to 0.2 kJ mol^{-1} (for CCSD(T)/VTZ geometries) and merely 0.02 kJ mol^{-1} (for CCSD(T)/VQZ geometries). However, the computationally economical CCSD(T)-F12/VDZ-F12 geometries also result in a near-zero RMSD of 0.02 kJ mol^{-1}. We therefore recommend using the latter in highly accurate composite *ab initio* methods such as W4lite, W4, and W4-F12.
- The performance of the def2-TZVPP and def2-QZVPP basis sets is very similar to that of the VTZ and VQZ basis sets.

SUPPLEMENTARY MATERIAL

See supplementary material for an overview of the 246 bonds in the W4-11-GEOM dataset (Table S1); overview of the basis-set convergence of the CCSD(T) method for the first-row molecules in the GEOM-AV6Z dataset (Table S2); squared correlation coefficients between the bond distances obtained with the CCSD(T)/A′V6Z level of theory and the CCSD(T)/VnZ and CCSD(T)/A′VnZ levels of theory (Table S3); overview of the basis-set convergence of the CCSD(T) and CCSD(T)-F12 methods for the H—X, X—Y, X═Y, and X≡Y bonds in the GEOM-AV6Z dataset (Table S4); overview of the basis-set convergence of the CCSD(T) and CCSD(T)-F12 methods for the H—X, X—Y, X═Y, and X≡Y bonds involving only first-row elements in the GEOM-AV6Z dataset (Table S5); effect of CCSD(T) and CCSD(T)-F12 reference geometries on molecular energies calculated at the CCSD(T)/CBS level of theory for the molecules in the GEOM-AV6Z database (Table S6); overview of the basis-set convergence of the CCSD(T) method for pathologically multireference systems (Figures S1 and S2); geometries for all the optimized structures are available on the website of the Karton group http://www.chemtheorist.com.

ACKNOWLEDGMENTS

We gratefully acknowledge the system administration support provided by the Faculty of Science at UWA to the Linux cluster of the Karton group, the financial support of Danish National Research Foundation (Center for Materials Crystallography, DNRF-93) to P.R.S., and an Australian Research Council (ARC) Discovery Early Career Researcher Award to A.K. (Grant No. DE140100311).

[1]I. Shavitt and R. J. Bartlett, *Many-Body Methods in Chemistry and Physics: MBPT and Coupled-Cluster Theory*, Cambridge Molecular Science Series (Cambridge University Press, Cambridge, 2009).
[2]R. J. Bartlett and M. Musiał, Rev. Mod. Phys. **79**, 291 (2007).
[3]A. Karton, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **6**, 292 (2016).
[4]K. A. Peterson, D. Feller, and D. A. Dixon, Theor. Chem. Acc. **131**, 1079 (2012).
[5]A. Karton, S. Daon, and J. M. L. Martin, Chem. Phys. Lett. **510**, 165 (2011).
[6]D. Feller, K. A. Peterson, and D. A. Dixon, J. Chem. Phys. **129**, 204105 (2008).
[7]T. Helgaker, W. Klopper, and D. P. Tew, Mol. Phys. **106**, 2107 (2008).
[8]K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, Chem. Phys. Lett. **157**, 479 (1989).
[9]J. D. Watts, J. Gauss, and R. J. Bartlett, J. Chem. Phys. **98**, 8718 (1993).
[10]A. Karton and J. M. L. Martin, J. Chem. Phys. **136**, 124114 (2012).
[11]L. A. Curtiss, P. C. Redfern, and K. Raghavachari, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 810 (2011).
[12]N. DeYonker, T. R. Cundari, and A. K. Wilson, in *Advances in the Theory of Atomic Molecular Systems*, Progress in Theoretical Chemistry Physics Vol. 19, edited by P. Piecuch, J. Maruani, G. Delgado-Barrio, and S. Wilson (Springer, Netherlands, Dordrecht, 2009), pp. 197–224.
[13]J. M. L. Martin, "Computational thermochemistry: A brief overview of quantum mechanical approaches," Annu. Rep. Comput. Chem. **1**, 31 (2005).
[14]T. Helgaker, W. Klopper, K. L. Bak, A. Halkier, P. Jørgensen, and J. Olsen, "Highly accurate *ab initio* computation of thermochemical data," in *Quantum-Mechanical Prediction of Thermochemical Data*, Understanding Chemical Reactivity Vol. 22, edited by J. Cioslowski (Kluwer, Dordrecht, 2001), pp. 1–30.
[15]J. M. L. Martin and S. Parthiban, "W1 and W2 theory and their variants: Thermochemistry in the kJ/mol accuracy range," in *Quantum-Mechanical Prediction of Thermochemical Data*, Understanding Chemical Reactivity Vol. 22, edited by J. Cioslowski (Kluwer, Dordrecht, 2001), pp. 31–65.
[16]K. Raghavachari, Chem. Phys. Lett. **589**, 35 (2013).
[17]D. A. Dixon, D. Feller, and K. A. Peterson, Annu. Rep. Comput. Chem. **8**, 1 (2012).
[18]D. Feller, K. A. Peterson, and D. A. Dixon, Mol. Phys. **110**, 2381 (2012).
[19]A. Karton and J. M. L. Martin, J. Chem. Phys. **133**, 144102 (2010).

[20]J. M. L. Martin and M. K. Kesharwani, J. Chem. Theory Comput. **10**, 2085 (2014).
[21]P. R. Tentscher and J. S. Arey, J. Chem. Theory Comput. **8**, 2165 (2012).
[22]To the best of our knowledge, this expression was first coined by T. H. Dunning, Jr. in a lecture series in the late 1990s.
[23]A. Karton, Chem. Phys. Lett. **645**, 118 (2016).
[24]D. W. Schwenke, J. Chem. Phys. **122**, 014107 (2005).
[25]E. C. Barnes, G. A. Petersson, D. Feller, and K. A. Peterson, J. Chem. Phys. **129**, 194115 (2008).
[26]D. Feller, K. A. Peterson, and J. G. Hill, J. Chem. Phys. **135**, 044102 (2011).
[27]D. Feller, J. Chem. Phys. **138**, 074103 (2013).
[28]A. Karton, P. R. Taylor, and J. M. L. Martin, J. Chem. Phys. **127**, 064104 (2007).
[29]D. Feller, K. A. Peterson, and T. D. Crawford, J. Chem. Phys. **124**, 054107 (2006).
[30]M. Heckert, M. Kállay, and J. Gauss, Mol. Phys. **103**, 2109 (2005).
[31]D. Feller, K. A. Peterson, W. A. de Jong, and D. A. Dixon, J. Chem. Phys. **118**, 3510 (2003).
[32]K. L. Bak, J. Gauss, P. Jørgensen, J. Olsen, T. Helgaker, and J. F. Stanton, J. Chem. Phys. **114**, 6548–6549 (2001).
[33]D. Feller and K. A. Peterson, J. Chem. Phys. **108**, 154 (1998).
[34]T. Helgaker, J. Gauss, P. Jørgensen, and J. Olsen, J. Chem. Phys. **106**, 6430 (1997).
[35]D. Feller and K. A. Peterson, J. Chem. Phys. **126**, 114105 (2007).
[36]D. Feller, K. A. Peterson, and D. A. Dixon, J. Phys. Chem. A **114**, 613 (2010).
[37]A. Karton, S. Parthiban, and J. M. L. Martin, J. Phys. Chem. A **113**, 4802 (2009).
[38]G. Knizia, T. B. Adler, and H.-J. Werner, J. Chem. Phys. **130**, 054104 (2009).
[39]M. Heckert, M. Kállay, D. P. Tew, W. Klopper, and J. Gauss, J. Chem. Phys. **125**, 044108 (2006).
[40]D. Feller, K. A. Peterson, and J. G. Hill, J. Chem. Phys. **133**, 184102 (2010).
[41]H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut *et al.*, MOLPRO, version 2012.1, a package of *ab initio* programs, 2012, see http://www.molpro.net.
[42]H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Schütz, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **2**, 242 (2012).
[43]T. H. Dunning, J. Chem. Phys. **90**, 1007 (1989).
[44]R. A. Kendall, T. H. Dunning, Jr., and R. J. Harrison, J. Chem. Phys. **96**, 6796 (1992).
[45]T. H. Dunning, K. A. Peterson, and A. K. Wilson, J. Chem. Phys. **114**, 9244 (2001).
[46]F. Weigend and R. Ahlrichs, Phys. Chem. Chem. Phys. **7**, 3297 (2005).
[47]T. B. Adler, G. Knizia, and H.-J. Werner, J. Chem. Phys. **127**, 221106 (2007).
[48]K. A. Peterson, T. B. Adler, and H.-J. Werner, J. Chem. Phys. **128**, 084102 (2008).
[49]The 14 highly multireference systems, for which the %TAE[(T)] diagnostic is in excess of 10%, are Be_2, B_2, $C_2(^1\sum^+)$, $BN(^1\sum^+)$, OF, F_2O, FOO, FOOF, Cl_2O, ClOO, OClO, O_3, S_3, and S_4. Note that *cis*-HO_3 and *trans*-HO_3 for which %TAE[(T)] = 7.4 and 7.9, respectively, are also included in this subset, see Refs. 3 and 5 for further details. We also exclude two additional beryllium systems, for which not all the considered basis sets are available, and they are BeF_2 and $BeCl_2$.
[50]S. Wang and H. F. Schaefer III, J. Chem. Phys. **124**, 044303 (2006).
[51]K. B. Wiberg, J. Comput. Chem. **25**, 1342 (2003).
[52]J. M. L. Martin and P. R. Taylor, J. Phys. Chem. **100**, 6047 (1996).
[53]Y. Xie, G. E. Scuseria, B. F. Yates, Y. Yamaguchi, and H. F. Schaefer, J. Am. Chem. Soc. **111**, 5181 (1989).
[54]The 35 systems containing second-row elements are AlH, AlH_3, AlF, AlF_3, AlCl, $AlCl_3$, SiH, SiH_4, Si_2H_6, SiO, SiF, SiH_3F, SiF_4, PH_3, P_2, P_4, HS, H_2S, CS, CS_2, SO, SO_2, SO_3, OCS, S_2, S_2H, S_2O, HCl, CCl_2, ClCN, NH_2Cl, ClO, HOCl, ClF, and Cl_2.
[55]A. Karton, E. Rabinovich, J. M. L. Martin, and B. Ruscic, J. Chem. Phys. **125**, 144108 (2006).
[56]N. Sylvetsky, K. A. Peterson, A. Karton, and J. M. L. Martin, J. Chem. Phys. **144**, 214101 (2016).
[57]A. Tajti, P. G. Szalay, A. G. Császár, M. Kállay, J. Gauss, E. F. Valeev, B. A. Flowers, J. Vázquez, and J. F. Stanton, J. Chem. Phys. **121**, 11599 (2004).
[58]Y. J. Bomble, J. Vázquez, M. Kállay, C. Michauk, P. G. Szalay, A. G. Császár, J. Gauss, and J. F. Stanton, J. Chem. Phys. **125**, 064108 (2006).
[59]M. E. Harding, J. Vázquez, B. Ruscic, A. K. Wilson, J. Gauss, and J. F. Stanton, J. Chem. Phys. **128**, 114111 (2008).
[60]C. Hättig, W. Klopper, A. Köhn, and D. P. Tew, Chem. Rev. **112**, 4 (2012).
[61]L. Kong, F. A. Bischoff, and E. F. Valeev, Chem. Rev. **112**, 75 (2012).
[62]S. Ten-no and J. Noga, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **2**, 114 (2012).
[63]J. G. Hill, K. A. Peterson, G. Knizia, and H.-J. Werner, J. Chem. Phys. **131**, 194105 (2009).
[64]D. Feller, K. A. Peterson, and B. Ruscic, Theor. Chem. Acc. **133**, 1407 (2013).