Title
A larger basis set describes atomization energy core-valence correction better than a higher-order coupled-cluster method

Authors
Aleksandr A. Chamkin,* Elena S. Chamkina
A.N.Nesmeyanov Institute of Organoelement Compounds of Russian Academy of Sciences
Russia, 119334, Moscow, Vavilova St. 28, bld. 1, INEOS
* Corresponding author: Aleksandr A. Chamkin <chalex@ineos.ac.ru>

Abstract
The accuracy of coupled-cluster methods for the computation of core-valence correction to atomization energy was assessed. Truncation levels up to CCSDTQP were considered together with (aug-)cc-pwCVnZ (n = D, T, Q, 5) basis sets and three different extrapolation techniques (canonical and flexible Helgaker formula and Riemann zeta function extrapolation). With the exception of CCSD, a more accurate correction can be obtained from a larger basis set with a lower-level coupled-cluster method, and not vice versa. For the CCSD(T) level, it also implies faster computations with modern codes. We also discussed the importance of moving to higher-order or all-electron methods for geometry optimizations. The present study provides the general knowledge needed for the most accurate state-of-the-art computations.

1. Introduction
By now, computational chemistry has evolved from the theorists’ toy to a distinct method capable of explaining observed reactions and accurately estimating measurable thermochemical quantities. However, a single finite-basis-set calculation typically does not possess sufficient accuracy for that. Therefore, composite methods that combine the results of several different computations are used. Usually, these are ab initio high-
level approaches with a small basis set corrected by lower-level methods with larger basis sets, which capture the essential effects in an additive manner. Composite methods are suitable for the evaluation of atomization energies, enthalpies of formation, ionization energies, and electron affinities with so-called “chemical accuracy” (1 kcal mol\(^{-1}\) or better). A wide range of research groups now uses these methods, whether for a comprehensive benchmark or a specific state-of-the-art application.

In 1989, John Pople introduced Gaussian-1 (G1), the first composite approach. Its improvement led to a family of Gaussian-n methods (Gn, n = 1 – 4), in which the latter ones learn from the experience of the earlier ones. These methods typically consist of a set of energy evaluations on a DFT geometry and include empirical parametrizations. The fully non-empirical Weizmann-n approaches (Wn, n = 1 – 4) possess unprecedented sub-kJ mol\(^{-1}\) accuracy. However, the computational cost is significant; even the W1 theory becomes unfeasible for a system of nine non-hydrogen atoms. In contrast to explicitly defined Wn and Gn methods, the Feller-Peterson-Dixon (FPD) approach includes a flexible sequence of coupled-cluster calculations, providing additive corrections to the primary energy component. This method’s high accuracy relies on treating all contributions as accurately as possible. In combination with the recent development of DLPNO-CCSD(T) approximation, the FPD approach has become attractive for relatively large systems. Some other procedures have been described for achieving an accurate thermochemistry of organic species: T1, HEAT, and the correlation consistent composite approach (ccCA), which is essentially similar to the FPD method.

In its original formulation, the FPD approach is dedicated to the accurate computation of the total atomization energy (TAE) of a system. It generally expresses the TAE as the sum of several terms: TAE = \(\Delta E_{\text{elec}} + \Delta E_{\text{CV}} + \Delta E_{\text{SR}} + \Delta E_{\text{SO}} + \Delta E_{\text{HO}} + \Delta E_{ZPE}\). The primary term \(\Delta E_{\text{elec}}\) is represented by the single-referenced frozen core (FC) CCSD(T) energy extrapolated to the complete basis set (CBS) limit. The additive corrections account for the core-valence (CV) correlation (\(\Delta E_{\text{CV}}\)), relativistic effects (\(\Delta E_{\text{SR}}\)), spin-orbit coupling (\(\Delta E_{\text{SO}}\)), post-CCSD(T) contributions (\(\Delta E_{\text{HO}}\)), and zero-point vibrational energy (\(\Delta E_{ZPE}\)).
$\Delta E_{CV}$ is the largest correction, with a magnitude comparable to 1 kcal mol$^{-1}$;\cite{12, 14} thus, accounting for CV effects is necessary for chemical accuracy. For some classes of systems, like transition metal complexes\cite{36-40} or noncovalent bound clusters,\cite{41} subvalence correlation plays a non-negligible role. At the same time, the definition of the FC can be ambiguous and troublesome,\cite{37} giving rise to errors up to 100 kcal mol$^{-1}$ in pathological cases.\cite{42} The importance of such effects even resulted in the development of a semi-empirical density functional designed especially for describing the CV correlation energy.\cite{43}

However, typically, $\Delta E_{CV}$ is estimated using CCSD(T). These calculations are prohibitively expensive when a large number of electrons should be correlated.\cite{25} The idea of the FPD method is that the $\Delta E_{CV}$ can be estimated using a much smaller basis set than that used for $\Delta E_{elec}$. Fortunately, this approach performs well because $\Delta E_{CV}$ converges faster to the CBS limit than $\Delta E_{elec}$.

The reported CV calculations using higher-order (HO) coupled-cluster methods are scarce and unsystematic because they are extremely demanding.\cite{12, 14} The authors of the W1 theory noted that the CCSD underestimates up to 50% of the CCSD(T) CV correlation, and it is questionable whether this contribution actually converges at the CCSD(T) level.\cite{25} This topic was revisited during W4 development,\cite{15} and it was concluded that the post-CCSD(T) CV contributions are substantial only for strongly multireference systems. Later, it was shown that even the CCSDT(Q) CV corrections are significant in some cases, and the recommended basis sets were provided for their evaluation.\cite{44} Additionally, there were a few specific high-level studies of small molecules when the CV correlation was considered on the CCSDT level.\cite{45-48}

The present study addresses the question of which approach is better for $\Delta E_{CV}$ estimation in terms of the time-accuracy tradeoff: Is it better to use a given truncated coupled-cluster method with the largest possible basis set or to use a lower level of theory with an even larger basis set? This question is not limited by CCSD(T) as the highest truncation level. For instance, assume that the CCSDT/cc-pwCV5Z calculation is not feasible due to high computational demand, but the CCSDT/cc-pwCVQZ and CCSD(T)/cc-pwCV5Z levels are affordable. Then, which of them will be better? This study is motivated by the fact that an increase in the basis set size lowers the electronic energy more significantly than an increase in the coupled-cluster truncation level.
Therefore, the former value converges better with respect to the full configurational interaction (FCI) limit at CBS. It can be assumed that the TAE will be converged better too. However, the accuracy of the energy differences (e.g. TAE) is governed not only by the individual electronic energy errors of the involved species but also by their variance, so the abovementioned assumption should be taken with caution.

This paper is organized as follows. Section 2 describes the technical details of the methods used. The first part of the discussion considers TAE CV correction ($\Delta E_{CV}$) on the CCSD and CCSD(T) levels (Section 3.1). This was the most studied topic prior to the present study and was added for completeness. The next part is dedicated to the HO truncation levels up to CCSDT(Q) (Section 3.2). Section 3.3 generalizes the drawn conclusions up to the extremely demanding CCSDTQP level. The computational times are also discussed throughout the text of these Sections. The last part of the discussion considers the importance of the geometry optimization level for the evaluation of CV correction (Section 3.4). Finally, the general conclusions and recommendations are provided (Section 4).

2. Computational Methods

2.1. General Considerations

The total atomization energy (TAE) of a molecule is the difference between the sum of the energies of the constituent atoms (A, B) and the energy of the whole molecule (AB): $TAE = E(A) + E(B) - E(AB)$. The CV correction for a given individual system is taken as the difference in energy between the all-electron calculation and FC one: $E_{CV} = E_{AE} - E_{FC}$. When the energy of a system is represented as a valence electron correlation energy plus additive CV correction (e.g. like in the FPD approach), the CV correction to the atomization energy (TAE CV correction) can be defined similarly to TAE: $\Delta E_{CV} = E_{CV}(A) + E_{CV}(B) - E_{CV}(AB)$.

The main interest of the present study is the convergence of the TAE CV correction with the increase in the basis set size and truncation level of the coupled-cluster operator. The performed computations for the molecules and atoms (the full details are given in following Sections) resulted in a list of energies obtained at a given level of theory ("Raw Energies" Section of ESI). These energies were aggregated for each system and...
extrapolated to the CBS limit using hand-coded processing scripts. Then, the TAE CV corrections were constructed according to the elemental compositions of the molecules and analysed.

2.2. Single-Point Energies

The coupled-cluster computations for the present study were performed using the MRCC program of M. Kallay (version of 2022-03-18). The “maxit” variable was defined as 1000 in the source file MRCCCOMMON because some jobs exceeded the default hard-coded limit of 50 coupled-cluster iterations. The sources were compiled using the documented standard procedure with recommended compilers. A hand-coded wrapper script was used to perform batch computations and save output files. Default convergence tolerances were used. Density fitting approximation was turned off for both the SCF and correlated calculations. The coupled-cluster computations were performed on top of the conventional RHF/UHF determinant.

The single-point energy evaluations were performed on the geometries either extracted from the W4-11-GEOM database or optimized using CCSD(T) with the basis set and FC approximation type corresponding to the given energy evaluation (the full details are given in Section 2.3). The motivation of such an approach is provided in the corresponding discussion. The hydrogen molecule was initially excluded because it has no core electrons. The cluster operator was truncated at some excitation level, resulting in coupled-cluster computations ranging from CCSD up to CCSDTQP, including perturbative models. For conciseness, these levels were abbreviated as “SD”, “SD(T)”, ..., and “SDTQP”. The cc-pCVnZ and aug-cc-pCVnZ (cc-pVnZ and aug-cc-pVnZ for H and He) basis sets were used, which were abbreviated as “nZ” and “anZ”, respectively. These basis sets are adapted for describing the inner-shell correlation. For each given level of theory (truncation level and basis set), both FC and CV calculations were performed. Please note that CV abbreviation implies the correlation of all electrons; we do not divide the inner-shell correlation into core-core and core-valence contributions here, which allows us to capture all inner-shell correlation effects. Such a division belongs to a different type of research.
Generally, the performed computations were divided into batches, as described in Table 1. For each batch, the FC and CV coupled-cluster calculations were performed up to the given truncation level and basis set size. The maximum feasible number of systems was considered in each case (as indicated by the Max. number of basis functions in Table 1). In addition, the corresponding computations were performed on the isolated atoms taken in their respective ground states. The raw data are provided in Tables S16–S21. In total, 1458 single-point calculations were performed. The most exhaustive computations were CV-SDTQP/TZ for H₂O and BH₃. We were unable to move beyond this point because the available resources were completely exhausted.

<table>
<thead>
<tr>
<th>Batch</th>
<th>The highest coupled-cluster level</th>
<th>The largest basis set</th>
<th>Geometry</th>
<th>Max. number of basis functions</th>
<th>Total number of systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SD(T)</td>
<td>5Z</td>
<td>W4-11-GEOM</td>
<td>600</td>
<td>96</td>
</tr>
<tr>
<td>2</td>
<td>SDT(Q)</td>
<td>QZ</td>
<td>W4-11-GEOM</td>
<td>190</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>SDTQ(P)</td>
<td>TZ</td>
<td>W4-11-GEOM</td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>SDTQP</td>
<td>TZ</td>
<td>W4-11-GEOM</td>
<td>85</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>SDT(Q)</td>
<td>aQZ</td>
<td>W4-11-GEOM</td>
<td>205</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>SDT(Q)</td>
<td>QZ</td>
<td>SD(T)/nZ</td>
<td>190</td>
<td>25</td>
</tr>
</tbody>
</table>

2.3. Geometries

Most single-point energies were evaluated on the geometries extracted from the W4-11-GEOM database. Generally, this database consists of 122 small non-charged systems (94 singlets, 22 doublets, and 6 triplets), excluding pathologically multireference ones. The dataset consists of 88 first-row, 17 second-row, and 17 mixed first- and second-row species, including hydrogen-containing (82), hydrogen-free (40), organic (63), and inorganic (59) compounds. The total number of unique bonds is 246, including 182 single bonds (117 H–X and 65 X–Y), 49 double bonds, and 15 triple bonds with varying degrees of covalent and ionic bonding characters. Most geometries were optimized at SD(T)/aug-cc-pV(6+d)Z, except for the biggest ones, which were optimized at SD(T)/aug-cc-pV(5+d)Z. The inclusion of both the first- and second-row systems is beneficial for the present study because the contributions of core-core and core-valence effects to the inner-shell correlation are opposite for these rows. As a result, W4-11-GEOM provides an excellent benchmark set of nearly-converged geometries with a broad spectrum of bonding situations.
For the specific studies of the CV contribution dependency on the geometry optimization level (Section 3.4), some systems from W4-11-GEOM were reoptimized using other methods. FC-SD(T)/nZ and CV-SD(T)/nZ optimizations were performed using ORCA 5.0.4\textsuperscript{59,60} with very tight SCF and geometry convergence criteria, numerical gradients, and conventional integral handling. FC-SDT/nZ optimizations were performed using MRCC with the same parameters as the single-point energies.

2.4. Extrapolation Techniques

In addition to using finite nZ and anZ basis sets, we extrapolated the correlation energies to the CBS limit using three different methods (e1–e3) as described below.

**e1:** The correlation energy is assumed to converge as $E_X = E + AX^{-3}$ (widely known as the Helgaker formula\textsuperscript{61}), where $E_X$ stands for the correlation energy obtained with the basis set cardinal number $X$, $E$ is the CBS limit energy, and $A$ is a system-specific constant. The present formula was used for the two-point extrapolation of the total correlation energy for each considered coupled-cluster method.

**e2:** The following more flexible form was used: $E_X = E + AX^{-\beta}$, where $\beta$ is a constant specific for a given pair of cardinal numbers, $X$ and $X + 1$. Initially, this constant was derived for the two-point 2/3 ($\beta = 2.46$) and 3/4 ($\beta = 3.05$) extrapolations for the SD(T)/nZ level.\textsuperscript{62} We optimized this exponent for the 4/5 ($\beta = 3.22$) extrapolation by performing the 3/4/5 three-point extrapolations for all available SD(T)/nZ ($n = T, Q, 5$) data and taking the median value. The described exponents were used for extrapolation of the SD- and SD(T)-level correlation energies. The HO increments were extrapolated separately using the same formula but specifically derived exponents: 2.7174 (2/3, SDT–SD(T)), 2.9968 (2/3, SDT(Q)–SDT), 1.7139 (2/3, SDTQ–SD(T)), 2.4807 (3/4, SDT–SD(T)), 3.3831 (3/4, SDT(Q)–SDT), and 2.6072 (3/4, SDT(Q)–SDT).\textsuperscript{63} There are no exponents optimized for higher contributions (SDTQ(P)–SDTQ and SDTQP–SDTQ(P)), so we assumed that these increments were sufficiently converged at the TZ level. The final extrapolated energy was then taken as the sum of the extrapolated SD(T)-level energies, extrapolated SDT–SD(T), SDT(Q)–SDT, and SDTQ–SDT(Q) increments (each with a specific exponent), and the SDTQP/TZ – SDTQ/TZ quantity. The extrapolations for the
augmented basis sets were performed similarly, with the only difference in the 2/3 exponent ($\beta = 2.51$) for the SD(T)/anZ level.\textsuperscript{62}

\textbf{e3}: This is the recent extrapolation technique based on the Riemann zeta function.\textsuperscript{64} The key formula is $E = E_L + L^4(E_{L-1} - E_{L-1} - (1^{-4} + 2^{-4} + \ldots + L^{-4}))$, where $E_L$ is the correlation energy obtained with the basis set cardinal number $L$, and $E$ is the exact energy. This two-point extrapolation method does not require the determination of any additional L-specific constants. This technique was applied to the full correlation energies for each considered coupled-cluster method, as in the \textbf{e1} method.

For clarity, the abbreviation “eN{X,Y}Z” was used throughout the text, where eN represents the type of the two-point extrapolation (e1–e3), while XZ and YZ are two consecutive basis sets used for this extrapolation.

The extrapolated energies can be represented as a linear combination of finite basis set energies in all cases (e1–e3). Therefore, when the reaction energies are of interest (e.g. TAE), there is no difference in whether they are extrapolated directly or the energies for the individual systems are extrapolated first and then subtracted.

The extrapolation of the Hartree-Fock (HF) energy was performed similarly for all methods (e1–e3) by using $E_X = E + A \exp(-\alpha X^{1/2})$ formula with fitted values of $\alpha$ (4.42 for 2/3 and 5.46 for 3/4 extrapolation).\textsuperscript{62}

3. Results and Discussion

3.1. CCSD and CCSD(T) truncation levels

First, we analysed the TAE CV corrections on the SD and SD(T) levels using basis sets up to 5Z (Table 1, Batch 1). For this purpose, we considered the maximum number of systems taken from the W4-11-GEOM database, for which the single-point CV-SD(T)/5Z computations were feasible. This resulted in a total of 96 species (Table S1). The most accurate energies can be obtained from the SD(T)/eN{Q,5}Z ($N = 1 – 3$) levels of theory, i.e. from the CBS-extrapolation of QZ and 5Z values using some method. However, there is no generally accepted consensus about which one is universally superior. Following the FPD approach, we constructed the reference TAE CV corrections as the mean of all three SD(T)/eN{Q,5}Z ($N = 1 – 3$)
quantities. The resulting values are given in Table S1 and represent the most accurate estimation of TAE CV correction for a given system. Additionally, Table S1 shows that the discrepancies between SD(T)/eN{Q,5}Z (N = 1 – 3) levels are much lower than chemical accuracy (RMS of 0.0094 kcal mol⁻¹), which indicates that the performance of considered extrapolation methods is similar.

Another approach that could be chosen for the construction of accurate reference values is to use explicitly correlated F12 methods. However, the available cc-pCVnZ-F12 basis sets were designed for outer-core, not all-electron correlation, while the cc-pwCVnZ-F12 basis sets were not developed. It was shown that SD(T)-F12/cc-pCVnZ-F12 calculations are inferior to SD(T)/anZ for inner-shell correlation. On the contrary, SD(T)/e1{aT,aQ}Z was used for CV correlation contribution in highly accurate W4 theory, while further improvement can be expected from SD(T)/e1{aQ,a5}Z. For this reason, F12 methods do not provide a sensible alternative for computing accurate CV corrections.

With reference values in hand, we can benchmark the accuracy of the lower levels of theory. We performed the corresponding FC and CV calculations for both SD and SD(T) methods and nZ (n = D, T, Q, 5) basis sets, constructed the corresponding TAE CV corrections of interest for each level of theory, and investigated their absolute deviations from the reference values (Table S1). Notably, there is no HF component in CV corrections because the FC and CV counterparts were always evaluated on identical geometries. Thus, both calculations provide the same HF energy, which cancels after subtraction. For the same reason, there was no need for its extrapolation.
Figure 1. Absolute deviations from the mean of SD(T)/eN\{Q,5\}Z (N = 1 – 3) values. The median (solid), 25th, and 75th percentiles (dash) are shown.

Figure 1 shows the essence of the observed trends. The violin plots indicate the distributions of the absolute deviations from the reference. The highest level of a plot corresponds to the maximal error. The median values are also shown using solid lines. Additionally, RMS values can be seen in Figure S2.

It is clear from Figure 1 that SD(T) generally outperforms SD even when the latter is combined with a larger basis set. The only exception is the pair of SD(T)/DZ and SD/TZ levels, for which the former is less accurate. The reason is that using the DZ-level basis set is meaningless for SD or SD(T) coupled-cluster methods (please note that this does not hold for the HO truncations). On the contrary, taking advantage of the CBS extrapolation significantly decreases the errors. Moreover, the 2/3 extrapolation of SD(T) energies provides sensible results, at least for the property of interest. It is also notable that the accuracy of SD is saturated at the error of ca. 0.25 kcal mol\(^{-1}\) even for the highest 5Z and eN\{Q,5\}Z levels.

Table S2 provides more detailed data. The positive values (highlighted in red) indicate that for the given levels of theory, the TAE CV correction computed on the SD level with a larger basis set (SD/(n+1)Z) is more accurate than on the SD(T) level with a smaller one (SD(T)/nZ). Except for the abovementioned SD(T)/DZ – SD/TZ difference, most of the values in Table S2 are negative, which highlights the superiority of SD(T). If
considering only the extrapolated data, there are 31 positive values out of 576 (5%). The provided in Table S2 statistical values support this finding.

The analysis of the computational times for SD and SD(T) is quite unambiguous. Figure S3 shows that the computation of perturbative triples takes about 5 – 30% of the time needed for the parent SD calculation. The SD(T) is always feasible when the SD is feasible. Thus, the TAE CV corrections should be evaluated using SD(T) with the largest possible basis set. Downgrading to SD is meaningless in terms of both accuracy and computational complexity, no matter which basis set is affordable with it. As a final remark, if we chose any individual of the SD(T)/eN{Q,5}Z (N = 1 – 3) levels as a reference or perform a separate analysis for each multiplicity, that will lead to the same conclusions.

3.2. CCSDT and CCSDT(Q) truncation levels

In the next part of the study, we considered TAE CV corrections estimated using HO coupled-cluster levels up to SDT(Q)/QZ (Table 1, Batch 2). The analysis is similar to that described in Section 3.1. We performed a set of SDT/nZ and SDT(Q)/nZ (n = D, T, Q) calculations for 25 systems from the W4-11-GEOM database (Table S3). The SD level was excluded because its analysis belongs to Section 3.1. It is important to note that the HO coupled-cluster theories converge faster to the CBS limit than SD or SD(T) methods.\textsuperscript{12, 44, 66} For this reason, firstly, the QZ basis set provides a decent accuracy for benchmark studies, and secondly, the DZ basis set provides at least meaningful results, although the accuracy of non-extrapolated DZ energies is expected to be low. The reference values in this part of the study were computed as the mean of SDT(Q)/eN{T,Q}Z (N = 1 – 3) energies, which still have low discrepancies (RMS of 0.0114 kcal mol\textsuperscript{-1}).

Figure 2 visualizes the deviations of lower levels of theory from the reference (RMS values are shown in Figure S5). This Figure actually conveys the main conclusion of this paper because the observed trends are strikingly opposite to those for SD and SD(T) methods. For now, the basis set size is more important than the truncation level of the coupled-cluster operator. For instance, SD(T)/eN{T,Q}Z (yellow plots in panels with

https://doi.org/10.26434/chemrxiv-2023-n7xbt-v2 ORCID: https://orcid.org/0000-0001-6060-8307 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0
extrapolated data) outperforms both SDT/eN[D,T]Z and even SDT(Q)/eN[D,T]Z. Similarly, SD(T)/QZ is more accurate than SD(T)/TZ and SD(T)/QZ.

Figure 2. Absolute deviations from the mean of SDT(Q)/eN[T,Q]Z (N = 1 – 3) values. The median (solid), 25th, and 75th percentiles (dash) are shown.

We can also evaluate the results directly. A positive value (highlighted in red) in Table S4 implies that the estimation of TAE CV correction becomes less accurate when one sacrifices a unit of basis set cardinal number for moving to the next coupled-cluster truncation level. It is clearly shown that most (84 %) of the values are positive. Quite interestingly, the SDT(Q)/eN[D,T]Z – SDT/eN[T,Q]Z differences are approximately two times higher than SDT/eN[D,T]Z – SD(T)/eN[T,Q]Z ones, suggesting that the importance of sticking with largest feasible basis set is increasing in a row of coupled-cluster methods.

To strengthen the performed analysis, we also considered the augmented basis sets (anZ) instead of their standard counterparts, nZ (Table 1, Batch 5). Diffuse functions are typically necessary for anions, weakly bound systems, and evaluation of some properties (polarizations, electron affinities, and dispersion forces), which is generally not the case in our study. However, the augmented basis sets are more saturated in the radial part. Thus, they are often used for extrapolation because CBS convergence is at least faster.64

We reperformed the abovementioned investigation using anZ sets (Tables S9 and S10). The reference values converged twice as better for anZ (RMS value of 0.0048 for anZ vs 0.0114 kcal mol\(^{-1}\) for nZ).
(S11) shows the results. A comparison of Figures 2 and 3 shows that the deviations from the corresponding reference are 2–3 times lower when augmented basis sets are employed. The qualitative conclusion is the same: The estimation of TAE CV correction is more accurate when a larger basis set is used. Using the SDT(Q)/eN{T,Q}Z or SDT(Q)/eN{aT,aQ}Z (N = 1 – 3) values separately as a reference or performing a separate analysis for each multiplicity leads to same conclusions for both Batches 2 and 5.

![Graph showing deviations from the mean of SDT(Q)/eN{aT,aQ}Z (N = 1 – 3) values.]

Figure 3. Absolute deviations from the mean of SDT(Q)/eN{aT,aQ}Z (N = 1 – 3) values. The median (solid), 25th, and 75th percentiles (dash) are shown.

Figure 4 shows the computational times for the discussed levels of theory. For rigor, the comparison of the SD(T) level with HO theories was performed for the same set of systems (i.e. with SDT(Q) data available). One important thing to note is that HO computations can be performed using only the general string-based CC algorithm “mrcc” of the MRCC program, while SD(T) computations can be performed using either the “mrcc” or optimized “ccsd” program (Table S11). The results of both types of calculations are reflected in Figure 4 (the separate plots are provided in Figures S12 and S13).
Figure 4. Computational times for different levels of theory. The SD(T) data were obtained using the “ccsd” or “mrcc” (transparent plots) programs. The median (solid), 25\text{th}, and 75\text{th} percentiles (dash) are shown.

The computational complexity of these coupled-cluster levels possesses a remarkable feature. For instance, consider the SDT(Q)/aDZ, SDT/aTZ, and SD(T)/aQZ methods. Accuracy is increased in this sequence (Figure 3). SDT(Q)/aDZ is the fastest and least accurate. Moving towards the two other methods requires more computational time. However, SD(T)/aQZ is better than SDT/aTZ in terms of both accuracy and computational efficiency when an optimized algorithm is used. Furthermore, SD(T)/aQZ also outperforms SDT(Q)/aTZ. Using the general string-based “mrcc” program did not support this trend. In that case, higher accuracy always implies higher computational times.

In conclusion, one should choose the largest feasible basis set and use extrapolation techniques when higher accuracy is desired. This can be done at the cost of lowering the coupled-cluster level to SD(T). Surprisingly, this can also lower the computational cost, depending on the implementation code. The same is valid for non-augmented basis sets (Figure S6).

3.3. CCSDTQ, CCSDTQ(P), and CCSDTQP truncation levels

The analysis of SDTQ(P) and SDTQP levels was the most demanding part of the present study (Table 1, Batches 3 and 4). These calculations were limited by the TZ basis set. The references were taken as the mean of
SDTQ(P)/eN{D,T}Z and SDTQP/eN{D,T}Z (N = 1 – 3) values, respectively (Tables S5 and S7). The lowest considered truncation level here was SDT(Q) because the lower methods were analysed in previous Sections.

The accuracy trends are aggregated in Figure 5 (S8). Two panels contain deviations for different sets of systems (Batches 3 and 4). As well as for the other HO levels, the usage of the larger basis set (TZ) resulted in much lower deviations from the references. The 2/3 extrapolation was readily available after a TZ calculation was performed. Clearly, it enhanced the convergence of TAE CV correction even more, with the e1 method being slightly less accurate. The direct analysis shows that the larger basis set is beneficial in all studied cases (Tables S6 and S8). The separate utilisation of extrapolated SDTQ(P)/eN{D,T}Z (SDTQP/eN{D,T}Z) or non-extrapolated SDTQ(P)/TZ (SDTQP/TZ) levels as the references lead to the same conclusions.

Figure 5. Absolute deviations from the mean of SDTQ(P) (left panel) or SDTQP (right panel) eN{D,T}Z (N = 1 – 3) values. The median (solid), 25th, and 75th percentiles (dash) are shown.

Please note that the deviations depicted in two panels of Figure 5 are computed versus two different references. Thus, it cannot be concluded that SDTQ(P) level provides an absolute accuracy similar to that of SDTQ and SDT(Q) by the direct comparison of two panels. In the first place, Figure 5 shows that SDT(Q)/TZ is superior to both SDTQ/DZ and SDTQ(P)/DZ, while SDTQ(P)/TZ is superior to SDTQP/DZ.

The computational times for the discussed truncation levels have an extremely steep increase as the size of the system is enlarged. For instance, SDTQP scales formally as $M^{12}$, where $M$ is the number of contracted
basis functions. Even when the corresponding calculations were feasible, the differences in computational costs were distinctive for the DZ and TZ basis sets (Figure 6). Moving from SDT(Q)/TZ to SDTQ/TZ and then to SDTQP/TZ requires about a tenfold increase in computational costs for each step.

![Figure 6. Computational times for different levels of theory.](image)

Figure 6. Computational times for different levels of theory. The median (solid), 25\textsuperscript{th}, and 75\textsuperscript{th} percentiles (dash) are shown.

As shown above, calculations with TZ basis set significantly increase both the accuracy and computational cost. Thus, when possible, a larger basis set should be preferred for TAE CV estimation. On the contrary, moving to the next truncation level with a DZ basis is pointless. Notably, SDT(Q)/eN{D,T}Z and SDTQ/eN{D,T}Z levels possessed similar deviations from the corresponding reference (Figure 5); however, their computational times differed by a factor of 20 (Figure 6). At least for the studied set of systems, the increase in accuracy does not worth the computational cost of the SDTQ method.

3.4. Geometry dependency

In the previous Sections 3.1 – 3.3, we discussed the TAE CV corrections computed in a single-point fashion on the geometries taken from the W4-11-GEOM database.\textsuperscript{51, 58} That means both CV and FC energy evaluations were performed on a single geometry, which was optimized using another (third) level of theory. In W4-11-GEOM, these geometries were optimized at decent FC-SD(T)/aug-cc-pV(n+d)Z (n = 5, 6) level and can be considered nearly-converged. It is common for a composite method to perform a set of single-point evaluations.
on a single geometry. Usually, this geometry comes from DFT. However, the original definition of the CV correction in the FPD approach states that FC and CV components should be evaluated at their respective optimized geometries.\textsuperscript{12} Hence, we will address the following points: 1) The importance of moving from SD(T) to HO methods for geometry optimization. 2) Basis set convergence of the geometries. 3) The importance of moving from FC to CV methods for geometry optimization. 4) Finally, we will evaluate TAE CV corrections following the original FPD approach definition and compare the results with those obtained earlier for SDT(Q) level.

3.4.1. Higher-order coupled-cluster methods for geometry optimization

First of all, we investigated the HO truncation levels for geometry optimization. Considering that some single-point evaluations took about a week to be finished, we are quite limited in our arsenal of methods and basis sets for optimizations. We chose a representative set of ten molecules from W4-11-GEOM database, which include first- and second-row elements, different multireference characters, multiplicity, and bond orders (Table 2). Then we performed the geometry optimizations using FC-SD(T) and FC-SDT methods (Table S14). The chosen basis sets (nZ) are in agreement with the other parts of this study. Moreover, they include "+d" basis functions by definition and thus are not inferior to those used in the W4-11-GEOM database.

Table 2. Bond length absolute differences for the selected systems optimized using FC-SD(T)/nZ and FC-SDT/nZ methods, in Å. Maximum (MAX), mean (MAD), and root mean square (RMS) values are provided.

<table>
<thead>
<tr>
<th>System</th>
<th>DZ</th>
<th>TZ</th>
<th>QZ</th>
<th>5Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>hs</td>
<td>0.00025</td>
<td>0.00019</td>
<td>0.00019</td>
<td>0.00018</td>
</tr>
<tr>
<td>sih</td>
<td>0.00047</td>
<td>0.00037</td>
<td>0.00034</td>
<td>0.00033</td>
</tr>
<tr>
<td>nh</td>
<td>0.00027</td>
<td>0.00022</td>
<td>0.00023</td>
<td>0.00021</td>
</tr>
<tr>
<td>alh</td>
<td>0.00057</td>
<td>0.00045</td>
<td>0.00023</td>
<td>0.00021</td>
</tr>
<tr>
<td>co</td>
<td>0.00015</td>
<td>0.00013</td>
<td>0.00019</td>
<td>0.00019</td>
</tr>
<tr>
<td>co</td>
<td>0.00012</td>
<td>0.00003</td>
<td>0.00003</td>
<td>0.00003</td>
</tr>
<tr>
<td>hcl</td>
<td>0.00103</td>
<td>0.00061</td>
<td>0.00051</td>
<td>0.00051</td>
</tr>
<tr>
<td>no</td>
<td>0.00472</td>
<td>0.00323</td>
<td>0.00285</td>
<td>0.00285</td>
</tr>
<tr>
<td>bn3pi</td>
<td>0.00322</td>
<td>0.00221</td>
<td>0.00193</td>
<td>0.00193</td>
</tr>
<tr>
<td>MAX</td>
<td>0.00472</td>
<td>0.00332</td>
<td>0.00285</td>
<td>0.00033</td>
</tr>
<tr>
<td>MAD</td>
<td>0.00108</td>
<td>0.00078</td>
<td>0.00067</td>
<td>0.00024</td>
</tr>
<tr>
<td>RMS</td>
<td>0.00186</td>
<td>0.00130</td>
<td>0.00112</td>
<td>0.00025</td>
</tr>
</tbody>
</table>
Table 2 summarizes the absolute differences in bond lengths between FC-SD(T) and FC-SDT levels. It shows how the geometry changes when one moves from SD(T) to SDT level. Even for DZ, the RMS deviation is about 0.00186 Å. According to the conclusions of the study on W4-11-GEOM geometries, such small deviation corresponds to the error in energy of ca. 0.0048 kcal mol$^{-1}$ for the state-of-the-art ab initio methods, which allows to obtain FCI-quality results (see Section 3.4.2). For larger basis sets, the deviation is even lower. For most systems (excluding difficult cases of cf, no, and bn3pi), the deviation is always lower than 0.001 Å. Thus, the consideration of HO methods is impractical because it requires tremendous computational effort and results in tiny (negligible in most cases) improvements.

### 3.4.2. Basis set convergence of the geometries at CCSD(T) level

Since the HO methods are impractical, we can now concentrate only on SD(T) truncation level. The basis set convergence at the FC-SD(T) method was comprehensively investigated for the W4-11-GEOM database. The main conclusions relevant to the present study are aggregated in Table 3. The second column of Table 3 provides RMS bond length deviations for systems optimized using a given basis set. The geometries of the original W4-11-GEOM database were chosen as a reference. Then the authors performed highly accurate W2-F12 energy evaluations on the poorly converged geometries corresponding to a given basis set. The third column of Table 3 shows RMS deviations of obtained energies from reference values evaluated on the W4-11-GEOM geometries.

Table 3. RMS deviations of bond lengths for a given basis set and of W2-F12 energies evaluated at the corresponding geometries.

<table>
<thead>
<tr>
<th>Basis set</th>
<th>RMSD of bond length, Å</th>
<th>RMSD of energy, kcal mol$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>cc-pV(D+d)Z</td>
<td>0.0196</td>
<td>0.5685</td>
</tr>
<tr>
<td>cc-pV(T+d)Z</td>
<td>0.0050</td>
<td>0.0406</td>
</tr>
<tr>
<td>cc-pV(Q+d)Z</td>
<td>0.0015</td>
<td>0.0048</td>
</tr>
<tr>
<td>cc-pV(5+d)Z</td>
<td>0.0004</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

https://doi.org/10.26434/chemrxiv-2023-n7xbt-v2 [ORCID: https://orcid.org/0000-0001-6060-8307] Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0
In other words, Table 3 shows how a bond length deviation is translated into an energy error. The authors conclude that cc-pV(D+d)Z should not be used for geometry optimization, cc-pV(T+d)Z is suitable for obtaining the CBS-quality results (error of <1 kcal mol\(^{-1}\)), and cc-pV(Q+d)Z will provide FCI-quality energies (error of <1 kJ mol\(^{-1}\)). With these data at hand, we can reiterate that it is beneficial to use SD(T) level with the largest feasible basis set instead of moving to HO theories for geometry optimization.

### 3.4.3. Core-valence methods for geometry optimization

The next question we address concerns the usage of CV-SD(T) for geometry optimization. For this purpose, we optimized all 25 systems from Batch 2 (Table 1) at FC-SD(T)/nZ and CV-SD(T)/nZ levels (n = D, T, Q). The optimized geometries are explicitly given in ESI, Section "Cartesian Coordinates of Optimized Structures". For now, let us consider only the diatomic systems (Table S15). Table 4 shows the absolute differences in their bond lengths optimized using FC and CV methods.

#### Table 4. Bond length absolute differences for the Batch 2 diatomic systems optimized using FC-SD(T)/nZ and CV-SD(T)/nZ methods, in Å. Maximum (MAX), mean (MAD), and root mean square (RMS) values are provided.

<table>
<thead>
<tr>
<th>System</th>
<th>DZ</th>
<th>TZ</th>
<th>QZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{hf})</td>
<td>0.00034</td>
<td>0.00050</td>
<td>0.00057</td>
</tr>
<tr>
<td>(\text{oh})</td>
<td>0.00049</td>
<td>0.00072</td>
<td>0.00083</td>
</tr>
<tr>
<td>(\text{nh})</td>
<td>0.00065</td>
<td>0.00107</td>
<td>0.00122</td>
</tr>
<tr>
<td>(\text{f2})</td>
<td>0.00080</td>
<td>0.00122</td>
<td>0.00138</td>
</tr>
<tr>
<td>(\text{no})</td>
<td>0.00062</td>
<td>0.00139</td>
<td>0.00164</td>
</tr>
<tr>
<td>(\text{hcl})</td>
<td>0.00037</td>
<td>0.00143</td>
<td>0.00180</td>
</tr>
<tr>
<td>(\text{ch})</td>
<td>0.00074</td>
<td>0.00164</td>
<td>0.00189</td>
</tr>
<tr>
<td>(\text{o2})</td>
<td>0.00090</td>
<td>0.00170</td>
<td>0.00191</td>
</tr>
<tr>
<td>(\text{n2})</td>
<td>0.00060</td>
<td>0.00170</td>
<td>0.00202</td>
</tr>
<tr>
<td>(\text{hs})</td>
<td>0.00017</td>
<td>0.00178</td>
<td>0.00222</td>
</tr>
<tr>
<td>(\text{co})</td>
<td>0.00073</td>
<td>0.00197</td>
<td>0.00234</td>
</tr>
<tr>
<td>(\text{cn})</td>
<td>0.00077</td>
<td>0.00221</td>
<td>0.00259</td>
</tr>
<tr>
<td>(\text{cf})</td>
<td>0.00113</td>
<td>0.00246</td>
<td>0.00291</td>
</tr>
<tr>
<td>(\text{bh})</td>
<td>0.00073</td>
<td>0.00276</td>
<td>0.00317</td>
</tr>
<tr>
<td>(\text{sih})</td>
<td>0.00085</td>
<td>0.00280</td>
<td>0.00346</td>
</tr>
<tr>
<td>(\text{bf})</td>
<td>0.00121</td>
<td>0.00317</td>
<td>0.00386</td>
</tr>
<tr>
<td>(\text{alh})</td>
<td>0.00156</td>
<td>0.00366</td>
<td>0.00460</td>
</tr>
<tr>
<td>MAX</td>
<td>0.00156</td>
<td>0.00366</td>
<td>0.00460</td>
</tr>
<tr>
<td>MAD</td>
<td>0.00074</td>
<td>0.00189</td>
<td>0.00226</td>
</tr>
<tr>
<td>RMS</td>
<td>0.00081</td>
<td>0.00207</td>
<td>0.00249</td>
</tr>
</tbody>
</table>

https://doi.org/10.26434/chemrxiv-2023-n7xbt-v2 ORCID: https://orcid.org/0000-0001-6060-8307 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0
Surprisingly, the discrepancies are higher for a larger basis set. The RMS differences are about 0.002 – 0.0025 Å for the sensible TZ and QZ levels. Taking into account the basis set convergence of the geometries for FC approximation, it can be speculated that such differences should result in energy deviations of 0.005 – 0.04 kcal mol\(^{-1}\) (see Table 3 and the corresponding discussion).\(^{58}\) However, in contrast to the usage of HO theories for optimization (Section 3.4.1), we can test this assumption to a greater extent, which will be done in the next Section.

3.4.4. Frozen core and core-valence energies evaluated at their respective optimized geometries

Finally, we can evaluate the convergence of TAE CV corrections estimated using HO methods in a manner similar to the FPD approach.\(^{12}\) Following the original definition, we should perform FC and CV computations for a given system at the corresponding optimized geometries. For instance, we should perform FC-SDT(Q)/nZ//FC-SDT(Q)/nZ and CV-SDT(Q)/nZ//CV-SDT(Q)/nZ calculations in order to obtain TAE CV correction on the SDT(Q)/nZ level. However, HO geometry optimizations are impractical and, in most cases, unfeasible at all (Section 3.4.1). A sensible alternative is to use only the SD(T) level combined with the corresponding correlation type (FC or CV). In this case, we should perform FC-SDT(Q)/nZ//FC-SD(T)/nZ and CV-SDT(Q)/nZ//CV-SD(T)/nZ for the example given above.

In this part of the study, we took the geometries of 25 systems from Batch 2 optimized at FC-SD(T)/nZ and CV-SD(T)/nZ (n = D, T, Q) levels and evaluated the corresponding FC-SDT(Q)/nZ and CV-SDT(Q)/nZ components needed to construct the TAE CV corrections (Batch 6). The performed analysis is similar to that in Section 3.2. The only difference is that the HF energy should now be considered because the FC and CV calculations for a given system were performed on different geometries and provided different HF energies. Again, the reference values were computed as the mean of SDT(Q)/eN{T,Q}Z (N = 1 – 3). We constructed the deviations of lower levels from them (Table S12) and visualized them in Figure 7 (S15).
Figure 7. Absolute deviations from the mean of SDT(Q)/eN{T,Q}Z (N = 1 – 3) values. The geometries were optimized at the corresponding FC- and CV-SD(T)/nZ levels. The median (solid), 25th, and 75th percentiles (dash) are shown.

The comparison of Figures 7 and 2 shows that the title conclusion holds. Again, SD(T) combined with a larger basis set outperforms both SDT and SDT(Q) levels with a smaller one for both extrapolated and non-extrapolated approaches. In 88% of considered cases, it's better to enlarge a basis set by sacrificing a coupled-cluster truncation level (Table S13). Notably, Figures 7 and 2 are visually almost indistinguishable. As shown in Section 3.4.3, FC- and CV-optimized geometries are slightly different, but this has little effect on the final results. The RMS difference between the references of Batch 2 and Batch 6 is tiny and equals 0.008 kcal mol\(^{-1}\).

4. Conclusions

In the present study, we considered the accuracy of coupled-cluster methods for evaluating CV correction for energy differences, namely, for TAE. The following conclusions can be drawn.

1. For non-HO levels, we reiterate that SD(T) is the first meaningful method. It should be combined with the largest feasible basis set and CBS extrapolation techniques. DZ basis set should be avoided. However, 2/3 extrapolation improves the accuracy compared to the non-extrapolated TZ basis set. The usage of SD always
results in a persistent error of 0.25 kcal mol$^{-1}$. There is no single reason to use SD in terms of both accuracy and efficiency.

2. The conclusions for the HO theories are more striking. The usage of a larger basis set is more sensible than moving to a higher truncation level. The SD(T)/(n+1)Z method is more accurate than both SDT/nZ and SDT(Q)/nZ. The lower variability of electronic energies increases the probability of compensating for errors in their difference. This holds for both extrapolated and non-extrapolated results (in any case, extrapolation is always advisable). The utilization of augmented basis sets additionally increases the accuracy and CBS convergence rate.

3. Furthermore, SD(T)/(n+1)Z calculations are typically faster than SDT/nZ and SDT(Q)/nZ because the former is a ubiquitous method that is optimized better in most implementations. This is even more important for practice because moving down the periodic table steeply increases the FC size compared to the number of valence electrons and particularly hinders CV HO calculations.

4. The accuracy trend holds beyond the SDT(Q) level. A larger basis set and a lower-level coupled-cluster method should be chosen when possible, not the other way round. For a given basis set, the CV correction becomes near converged on such high truncation levels. The further increase in accuracy may not be worth the high computational costs associated with the next coupled-cluster method, especially when other approximations are used.

5. Regarding the CBS extrapolation method, the best one still cannot be suggested. $\text{e}_1$ was the best one for SD(T) truncation level, $\text{e}_3$ – for SDT(Q), and $\text{e}_2$ – for SDTQP. Considering that $\text{e}_1$ performs badly for valence computations and the recent $\text{e}_3$ needs additional benchmarks, we can suggest using $\text{e}_2$ or averaged mixture of techniques.

6. Moving from SD(T) to HO levels for geometry optimization requires high computational effort but results in low improvements. Again, it is practical to use SD(T) level with the largest feasible basis set. Basis set convergence of the geometries is smoother than the convergence with respect to the coupled-cluster truncation level.
7. FC and CV converged geometries are different. The typical bond length differences between them are about 0.002 – 0.0025 Å for the SD(T)/TZ and SD(T)/QZ levels. The corresponding electronic energy differences can be estimated at 0.005 – 0.04 kcal mol\(^{-1}\).

8. TAE CV can be evaluated on the geometry optimized with the same method or taken from the other source (DFT, XRD). Indeed, the absolute accuracies of the different approaches are not the same. However, the title conclusion generally holds – a larger basis set describes TAE CV correction better than a higher-order coupled-cluster method.

**Conflicts of interest**

There are no conflicts to declare.

**Acknowledgements**

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Contract No. 075-03-2023-642).

**References**


https://doi.org/10.26434/chemrxiv-2023-n7xbt-v2 ORCID: https://orcid.org/0000-0001-6060-8307 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0
