JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

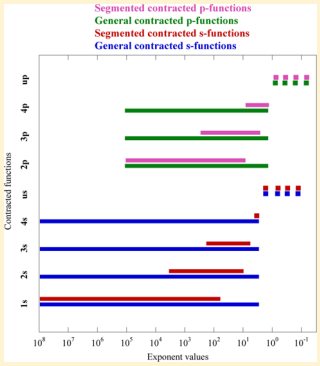
Unifying General and Segmented Contracted Basis Sets. Segmented Polarization Consistent Basis Sets

Frank Jensen*

Department of Chemistry, Aarhus University, DK-8000 Aarhus, Denmark

Supporting Information

ABSTRACT: We propose a method, denoted P-orthogonalization, for converting a general contracted basis set to a computationally more efficient segmented contracted basis set, while inheriting the full accuracy of the general contracted basis set. The procedure can be used for any general contracted basis set to remove the redundancies between general contracted functions in terms of primitive functions. The P-orthogonalization procedure is used to construct a segmented contracted version of the polarization consistent basis sets, which are optimized for density functional theory calculations. Benchmark calculations show that the new pcs-*n* basis sets provide uniform error control of the basis set incompleteness for molecular systems composed of atoms from the first three rows in the periodic table (H−Kr) and for different exchange-correlation functionals. The basis set errors at a given zeta quality level are lower than other existing basis sets, and the pcs-*n* basis sets are furthermore shown to be among the computationally most efficient. The pcs-*n* basis sets are available in qualities ranging from (unpolarized) double-zeta to pentuple zeta quality and should therefore be well suited for both routine and benchmark calculations using density functional theory methods in general.



INTRODUCTION

Gaussian basis sets are at the root of all application oriented electronic structure methods in computational chemistry. Many different basis sets have been proposed over the years, and two recent reviews can be consulted for an overview.[1,2] For routine applications, the basis set choice is often done implicitly as part of a specific computational model, the ubiquitous B3LYP/6-31G* combination being a vivid example. Such fixed method and basis set combinations often benefit from error cancellations between method and basis set errors and/or have been used so extensively that the expected performance can be estimated from previous studies. In order to separate the method and basis set errors, and quantify the latter, it is necessary to perform calculations with at least two different basis sets of well-defined quality. Basis set quality can conveniently be classified in terms of the highest angular momentum function included, and this gives rise to the familiar Double/Triple/Quadruple etc. Zeta notation.[3]

The role of the basis set is to provide the best representation for the orbitals/density for as low a computational cost as possible. Since different methods have different basis set requirements, and different computer implementations have different performance characteristics, it is not possible to have a single optimum basis set of a given zeta quality. Optimum basis sets for independent particle methods, like Hartree−Fock (HF) and density functional theory (DFT), have different compositions in terms of angular momentum functions than optimum basis sets for correlated methods. Older basis sets, such as the Pople style *n*-*klm*G basis sets,[4−15] were designed as being of general purpose, while modern basis sets usually are optimized for specific methods. The Dunning cc-pVXZ,[16−20] Petersson nZaP,[21−23] Sapporo,[24−27] Roos ANO-XZP,[28−32] and Jorge XZP[33−36] basis sets are aimed at correlated wave function methods, while the pc-*n* basis sets[37−42] have been optimized for DFT methods. The Karlsruhe Def2 family of basis sets[43] is provided in two flavors, depending on whether they are used for correlated or independent particle (DFT/HF) methods. We note that the Pople *n*-*klm*G designation used in the Gaussian-09 program[44] includes basis sets with a different number of primitive and contracted functions than indicated by the 6-31G and 6-311G notations for e.g. second row[11] and transition metals.[14,15] More recently, basis sets optimized for F12-correlated methods[45,46] have started to appear,[47−49] and they have requirements intermediate between HF and traditional electron correlation methods.

Basis sets differ in terms of number and types of primitive functions and how these are grouped into contracted functions. Basis set contraction is a key component for improving the computational efficiency, since it reduces the number of independent variational parameters. Two types of contraction schemes have been devised: a general contraction,[50] where all primitive functions are allowed to contribute to all contracted functions, and a segmented contraction,[51] where each primitive function is allowed to contribute to only one contracted function. These two contraction types are in practice rarely used in their purest forms. General contracted basis sets usually allow the outer valence functions to be uncontracted, which can be viewed as a partial segmentation. Segmented contracted basis sets, on the other hand, often allow duplication of some of

Received: November 26, 2013
Published: February 13, 2014

   dx.doi.org/10.1021/ct401026a | *J. Chem. Theory Comput.* 2014, 10, 1074−1085

the primitive functions, which can be viewed as a limiting form for general contraction. The cc-pVXZ,[16−20] nZaP,[21−23] ANO-XZP,[28−32] and pc-*n*[37−42] basis sets are of the general contracted type, while the Pople,[4−15] Sapporo,[24−27] Jorge,[33−36] and Karlsruhe[43] basis sets are of the segmented type. All of these basis sets have been optimized by minimizing the atomic (in some cases also molecular) energy with respect to the exponents of the primitive functions and possibly also the contraction coefficients. Parameterized or nonoptimized basis sets tend to be less efficient, since they require more functions for the same accuracy, and they are consequently used less frequently.[2]

Contraction is always a compromise between reducing the flexibility of the basis set and improving the computational efficiency. The optimum contraction can be defined as the strongest contraction where the contraction error is smaller than, but comparable to, the inherent error in the underlying (uncontracted) primitive basis set relative to the basis set limit. This principle is straightforward to apply for a general contraction scheme: the primitive set of functions is initially contracted to a minimum set of functions, and the valence primitive functions are then sequentially uncontracted until the contraction error drops below the target criterion. The key issue is that the same set of (optimized) primitive functions is used in a sequence of contracted basis functions of increased flexibility, and the results therefore smoothly connect to the results for the completely uncontracted set of functions. In a segmented contraction, on the other hand, the contraction coefficients and exponents of the primitive functions are optimized simultaneously, and each different contraction thus employs a different set of primitive functions. This often leads to irregular changes between different choices of contraction, and thus difficulties in defining the actual contraction error, and in defining the best contraction scheme. Selecting the number of primitive functions (including possible duplications of some of these) and selecting how these are partitioned between contracted functions is a combinatorial problem that potentially leads to many possible choices, especially for large basis sets and/or for elements beyond Ar.[25,52] For a specific combination of primitive functions and partitioning, the combined exponent and contraction coefficient optimization is a highly nonlinear problem with multiple local minima,[53] which further adds to the ambiguity of selecting the optimum combination. Indeed, the final choice of number of primitive functions, their segmentation into contracted functions, and selection of the proper local minimum in the parameter space is often described as being the result of substantial benchmark calculations.[34,54]

Segmented contraction can be considered as a special case of general contraction, and computer programs that are designed for general contracted basis sets can therefore also handle segmented basis sets without loss in efficiency. Programs designed for segmented contracted basis sets, on the other hand, can only handle general contracted basis sets by explicitly duplicating the number of primitive functions as many times as there are contracted functions, which significantly increases the size of the underlying (primitive) set of functions. Integral screening is commonly used for reducing the computational effort for large systems,[55−57] and since the screening is performed for contracted functions, this works more efficiently for segmented contraction than for general contraction. For highly correlated methods, like CCSD, the computational effort is dominated by solving nonlinear equations depending on the number of molecular orbitals, and integral evaluation usually only accounts for a minor fraction of the total computational time. The choice of general vs segmented contracted basis set is therefore not critical. HF and DFT calculations, on the other hand, strongly depend on the size and contraction of the basis set, and segmented contraction here has a clear computational advantage. Recent developments in electron correlation methods, like reduced scaling techniques,[58,59] F12-correlated,[45,46] and integral direct methods,[60] may refocus the computational effort toward efficiency in the integral evaluation.[61]

In the present work we show how a general contracted basis set in a systematic procedure can be converted into segmented basis sets of different sizes and associated qualities by varying a single parameter. By requiring that the total energies of the general and segmented basis sets are identical (to within a predefined threshold), this allow selecting a single segmentation scheme from the (combinatorial) many possibilities of combining primitive functions into contracted functions and uniquely defines how many primitive functions that should be duplicated. It furthermore provides a tool for deciding which of the many local minima in the combined exponent and coefficient parameter space that should be selected. The procedure is based on the observation by Davidson[62] that general contracted basis sets contain redundancies, and the unique segmentation scheme can be considered as eliminating the redundancy in the general contraction. The redundancy elimination allows generation of computationally more efficient segmented basis sets while inheriting the full accuracy of the general contracted basis sets. The principle is used to design segmented versions of the pc-*n* basis sets for the first 36 elements in the periodic table, and benchmark calculations show that they perform equivalently to the general contracted versions, while being significantly more efficient computationally. For DFT calculations they display the lowest basis set errors at a given zeta quality level, and at the same time are among the computationally most efficient basis sets.

THEORY

Consider K sets of atomic orbital (AO) coefficients (e.g., 1s, 2s, 3s, ...) defined by the vectors $\mathbf{c}_k$ with elements $c_{k\alpha}$ in a set of primitive basis functions χ_α of size N_{basis}. The primitive basis functions are nonorthogonal with overlap elements $S_{\alpha\beta}$, and the $\mathbf{c}_k$ vectors are orthonormal as defined in eq 1.

$$\langle \mathbf{c}_i | \mathbf{c}_j \rangle = \sum_{\alpha\beta}^{N_{basis}} c_{i\alpha} c_{j\beta} \langle \chi_\alpha | \chi_\beta \rangle = \sum_{\alpha\beta}^{N_{basis}} c_{i\alpha} c_{j\beta} S_{\alpha\beta} = \delta_{ij} \tag{1}$$

We will assume that the primitive basis functions are arranged according to the magnitude of their exponents and refer to them as innermost (large exponents χ_1, χ_2, χ_3,...) and outermost (small exponents ..., $\chi_{N-2}, \chi_{N-1}, \chi_N$).

The AO coefficients can be used to define K general contracted basis functions $\mathbf{a}_k$, such that the M innermost functions are contracted while the N_{basis}-M outermost functions are left uncontracted, and the contracted basis set therefore has a total of K+N_{basis}-M functions. The $a_{k\alpha}$ elements are thus equal to $c_{k\alpha}$ for $\alpha \leq M$ and zero for $\alpha > M$. The contracted functions are no longer orthogonal, since the sum now is only over M primitive functions.

$$\langle \mathbf{a}_i | \mathbf{a}_j \rangle = \sum_{\alpha\beta}^{M} a_{i\alpha} a_{j\beta} S_{\alpha\beta} \neq \delta_{ij} \tag{2}$$

dx.doi.org/10.1021/ct401026a | J. Chem. Theory Comput. 2014, 10, 1074−1085

Table 1. Contraction Coefficients for the Primitive s-Functions for the Kr pc-1 Basis Set[a]

	Original coefficients				Davidson purified				P-orthogonalized ($C_{ortho} = 10^{-4}$)			
Exponent	1s	2s	3s	4s	1s	2s	3s	4s	1s	2s	3s	4s
1.47E+05	3.1E-03	-5.7E-04	1.9E-04	8.9E-05	3.1E-03	0.0	0.0	0.0	3.1E-03	4.8E-05	1.1E-06	2.3E-08
2.21E+04	2.4E-02	-4.4E-03	1.4E-03	6.9E-04	2.4E-02	1.0E-06	0.0	0.0	2.4E-02	3.7E-04	9.4E-06	4.8E-07
5.02E+03	1.2E-01	-2.3E-02	7.5E-03	3.6E-03	1.2E-01	-7.1E-04	-5.0E-04	0.0	1.2E-01	1.1E-03	2.1E-05	-1.1E-06
1.42E+03	4.3E-01	-8.7E-02	2.9E-02	1.4E-02	4.2E-01	-7.7E-03	-5.5E-03	-5.3E-06	4.2E-01	-1.2E-03	-2.3E-05	7.4E-07
4.59E+02	9.9E-01	-2.5E-01	8.5E-02	4.1E-02	9.8E-01	-6.9E-02	-4.9E-02	-1.8E-04	9.8E-01	-5.4E-02	-1.1E-03	-1.0E-04
1.61E+02	1.0E+00	-3.6E-01	1.3E-01	6.4E-02	1.0E+00	-1.9E-01	-1.4E-01	-1.9E-03	1.0E+00	-1.7E-01	1.0E-03	9.3E-05
4.30E+01	1.6E-01	4.3E-01	-2.2E-01	-1.1E-01	9.9E-02	4.7E-01	3.8E-01	1.8E-02	9.9E-02	4.7E-01	-4.0E-02	-2.1E-03
1.87E+01	-5.5E-02	1.0E+00	-7.0E-01	-3.8E-01	-1.9E-01	1.0E+00	1.0E+00	1.2E-01	-1.9E-01	1.0E+00	-2.9E-01	1.8E-03
6.04E+00	2.2E-02	1.8E-01	3.3E-01	2.4E-01	0.0	2.0E-01	-2.2E-01	-1.8E-01	0.0	2.0E-01	3.9E-01	-2.9E-02
2.54E+00	-1.0E-02	-3.2E-02	1.0E+00	1.0E+00	0.0	0.0	-9.1E-01	-7.9E-01	0.0	0.0	1.0E+00	-4.1E-01
5.47E-01	3.2E-03	7.7E-03	9.2E-02	-8.9E-01	0.0	0.0	0.0	1.0E+00	0.0	0.0	0.0	1.0E+00

[a]Yellow indicate coefficient values less than ∼0.002. All three sets of contraction coefficients span the same vector space and produce identical results.

The $\boldsymbol{a}_k$ set of contracted functions contains redundancies, as pointed out by Davidson.[62] One of the $a_{k\alpha}$ ($k \neq 1$) coefficients can be made equal to zero by subtracting a constant x times the $\mathbf{a}_1$ vector from the $\mathbf{a}_k$ vector.

$$\mathbf{a}'_k = \mathbf{a}_k - x\mathbf{a}_1 \tag{3}$$

If the x factor is chosen as the a_{k1}/a_{11} ratio, then the coefficient for the χ_1 primitive function becomes zero for the $\boldsymbol{a}'_k$ contracted function. This procedure can be applied recursively to remove the χ_2 primitive function from the contracted functions $\mathbf{a}'_{k>2}$ by subtracting a constant times the $\mathbf{a}'_2$ vector etc., such that the k'th contracted function has (k-1) zero coefficients for the (k-1)'th innermost primitive functions. The elimination procedure can subsequently be applied in the reverse direction by subtracting a fraction of the $\boldsymbol{a}'_K$ function from $\boldsymbol{a}'_{K-1}$ etc., such that in the final set of functions, the k'th contracted function has K-k zero coefficients for the (K-k)'th outermost primitive functions. This inside-out purification removes (K-1) coefficients from each contracted function, as shown in Table 1 labeled *Davidson purified*. The purified set of $\mathbf{a}'_k$ contracted functions span the same vector space as the original set of functions, and thus produce identical results (within the usual numerical accuracies). As also pointed out by Davidson,[62] it is possible to further reduce the number of primitive functions contributing to each contracted function by removing functions with coefficients smaller than a suitable threshold, but this represents a genuine reduction of the vector space, which of course can be controlled by the magnitude of the cutoff parameter. A coefficient cutoff parameter of $\sim 10^{-5}$ will for most practical applications produce negligible differences in the final results, and larger cutoff may also be acceptable, depending on the application.

The x factor (eq 3) in the above purification procedure was chosen to remove the innermost primitive functions from the outermost contracted functions and the outermost primitive functions from the innermost contracted functions, but this is by no means a unique choice. Any of the a_{kn} coefficients ($k \neq 1$) can be eliminated by choosing the x factor as the ratio a_{kn}/a_{1n}, but the inside-out procedure leads to the computationally most efficient contracted basis functions for use with integral screening techniques. If the elimination procedure is used in connection with neglect of small coefficients, however, it becomes attractive to find elimination factors which lead to the largest number of coefficients in the final set of contracted functions below the coefficient cutoff threshold. By using suitable choices for the elimination factor and coefficient cutoff, we anticipated that it could be possible in a systematic fashion to generate a sequence of (partly) segmented contracted basis functions from a general contracted set of functions, such that an optimum segmentation could be generated for a predefined maximum contraction error. As illustrated in the following, this is indeed possible but requires avoiding a couple of potential pitfalls.

The $\mathbf{a}_k$ functions are nonorthogonal (eq 2), and the purified set of $\mathbf{a}'_k$ functions is also nonorthogonal. The purification process may either reduce or increase the degree of nonorthogonality, depending on the x factor, but a large x factor will always lead to final contracted functions which are near-linear dependent. This can be seen from eqs 4 and 5, where the $\mathbf{a}_j$ and $\mathbf{a}_i$ functions are assumed to be normalized.

$$\mathbf{a}'_j = \mathbf{a}_j - x\mathbf{a}_i \tag{4}$$

$$\frac{\langle \mathbf{a}'_j | \mathbf{a}_i \rangle}{\sqrt{\langle \mathbf{a}'_j | \mathbf{a}'_j \rangle}} = \frac{\langle \mathbf{a}_j | \mathbf{a}_i \rangle - x}{\sqrt{1 + x^2 - 2x\langle \mathbf{a}_j | \mathbf{a}_i \rangle}} \tag{5}$$

The normalized overlap between the purified $\mathbf{a}'_j$ function and the $\mathbf{a}_i$ function approaches ±1 as the x factor becomes large. The $\langle a_j | a_i \rangle$ overlap is numerically less than one, and even a modest x factor of ∼5 thus results in near-linear dependence of the final set of contracted functions, which clearly is undesirable.

The elimination of inner primitive functions from the outer contracted functions by recursive subtraction of fractions of the inner contracted functions is prone to numerical instabilities. This becomes especially problematic as the size of the primitive basis set increases, and as the number of contracted functions (occupied AOs) increases. For a large basis set like pc-4 for a third-row element, the 1s, 2s, 3s, and 4s AO coefficients for the innermost primitive functions are numerically small and already near-linear dependent before any purification. After the coefficients for the χ_1 and χ_2 primitive functions have been removed from $\mathbf{a}_3$ and $\mathbf{a}_4$ by subtraction of fractions of $\mathbf{a}_1$ and $\mathbf{a}_2$, the remaining $\mathbf{a}'_3$ and $\mathbf{a}'_4$ coefficients for the inner primitive functions often suffer from near-loss of all important information. The final elimination of the χ_3 primitive function from $\mathbf{a}'_4$ by subtraction of a fraction of $\mathbf{a}'_3$ consequently often leads to large and unpredictable x factors and consequently near-linear dependence of the final set of contracted functions. This numerical instability is similar to the problem in Gauss-Jordan elimination when solving systems of linear equations, and here the numerical stability can be improved by pivoting,

i.e. using the numerical largest elements for elimination. This, however, is not a desirable option for segmenting basis sets, as it will eliminate primitive functions in the middle range of the contracted functions. Enforcing a minimum numerical coefficient value for calculating the x factor in principle solves the numerical issues, but it was found that the x factors, and thus the final set of contracted functions, were very sensitive to the minimum coefficient value used and again often lead to near-linear dependency of the final set of contracted functions. This sensitivity was observed regardless of whether the x factors were calculated as simple ratios of coefficients or as (possibly weighted) averages of coefficient ratios.

Given the objective of eliminating as many primitive functions from the contracted functions as possible for a given accuracy, we also investigated whether suitable x factors for function purification could be obtained by maximizing the number of coefficients below a given threshold in the purified functions. Although this gave better numerical stability, it occasionally also produced near-linear dependent set of contracted functions.

The observation that the inner parts of different AOs are significantly correlated lead us to propose the following strategy for eliminating redundancy between sets of general contracted functions. The normalized overlap O^P_{ij} between two functions $\mathbf{a}_i$ and $\mathbf{a}_j$ up to the P'th primitive function is defined in eqs 6 and 7.

$$O^P_{ij} = \frac{|\langle \mathbf{a}_i|\mathbf{a}_j\rangle^P|}{\sqrt{\langle \mathbf{a}_i|\mathbf{a}_i\rangle^P \langle \mathbf{a}_j|\mathbf{a}_j\rangle^P}} \tag{6}$$

$$\langle \mathbf{a}_i|\mathbf{a}_j\rangle^P = \sum_{\alpha\beta}^{P} a_{i\alpha} a_{j\beta} S_{\alpha\beta} \tag{7}$$

The O^P_{ij} quantity is 1 (exactly) for $P = 1$ and gradually becomes smaller as P increases toward M (and would become 0 for $P = N_{basis}$). We introduce an orthogonalization parameter C_{ortho}, such that two functions $\mathbf{a}_i$ and $\mathbf{a}_j$ that have 1-O^P_{ij} less than C_{ortho} are considered to be linear dependent up to the primitive function P. For two such functions the $\mathbf{a}_j$ function is P-orthogonalized against $\mathbf{a}_i$ for the full set of M primitive functions according to eq 4, where the x factor is calculated as in eq 8.

$$x^P_{ij} = \frac{\langle \mathbf{a}_i|\mathbf{a}_j\rangle^P}{\langle \mathbf{a}_i|\mathbf{a}_i\rangle^P} \tag{8}$$

The P-orthogonalization will make $\mathbf{a}'_j$ orthogonal to $\mathbf{a}_i$ for the first P primitive functions. To avoid numerical instability, we initially set $P = M$ and decrement P until 1-$O^P_{ij} \leq C_{ortho}$. The corresponding P_c value is the maximum number of primitive functions for which the $\mathbf{a}_i$ and $\mathbf{a}_j$ contracted functions are considered linear dependent within the given threshold. All $\mathbf{a}_{j>1}$ functions are initially P-orthogonalized against $\mathbf{a}_1$ up to level P_{c1}, where P_{c1} is the smallest common factor for all 1-j pairs of functions. Analogous to the Davidson purification procedure, the P-orthogonalization is used recursively, i.e. all $\mathbf{a}'_{j>2}$ functions are subsequently P-orthogonalized against $\mathbf{a}'_2$ up to level P_{c2}, where P_{c2} is the smallest common factor for all 2-j pairs of functions, and so forth up to level (K-1). It is important that the same P_{ci} value is used for all P-orthogonalizations against a given $\mathbf{a}_i$, as the orthogonality otherwise is lost when the procedure is used recursively. In the limit of C_{ortho} being very small, the P-orthogonalization procedure reduces to the Davidson inside-out purification, as the x^P_{ij} factor (eq 8, P = 1) simply becomes the ratio of the a_{j1}/a_{i1} coefficients.

The outside-in elimination of the coefficient for the outermost primitive functions from the innermost contracted functions do not suffer from the above numerical and linear dependency problems, and the original method of using x factors as simple coefficient ratios is unproblematic.

Once the near-linear dependencies between contracted functions are eliminated by P-orthogonalization, it is convenient to use an intermediate normalization (largest coefficient = 1.0) for the contracted functions. Table 1 shows the effect of P-orthogonalization for the s-functions of the pc-1 basis for the Kr atom for a C_{ortho} value of 10^{-4}, which leads to orthogonalization levels of 4, 6 and 8, respectively, for the 2s, 3s, and 4s functions. It is important to stress that the three sets of contracted functions in Table 1 produce identical results (within the usual numerical considerations). The P-orthogonalized functions, however, have significantly more numerical small coefficients than the original set of coefficients. The premise of the proposed method is that all primitive functions that are part of the P-orthogonalization are neglected (indicated in yellow in Table 1), and this typically corresponds to neglecting coefficients smaller than ∼10 times C_{ortho}. Table 1 shows that neglect of coefficients smaller than ∼0.002 only removes 5 primitive functions in the original set of functions, while the Davidson purification removes 18, and the P-orthogonalization removes 24 primitive functions.

The P-orthogonalization employs a user-defined parameter C_{ortho} for reducing redundancy between the general contracted functions, which implicitly defines a coefficient threshold for neglecting numerically small coefficients in the purified functions. The neglect of primitive functions in the contracted functions means that the exponents/coefficients are no longer optimal in an energetic sense, and they are in a subsequent step optimized by minimizing the atomic energy. A series of segmented contracted functions with increasingly larger contraction errors can thus be produced from the same set of general contracted functions by gradually increasing the C_{ortho} value. This procedure ensures that there is a smooth relationship between C_{ortho} and the contraction error and thus provides a method for identifying the importance of a given primitive function in a given contracted function. This information is not visible from the coefficient in the original general contracted functions (Table 1), as they contain significant redundancies. C_{ortho} values in the 10^{-5}–10^{-3} range in practice serve to generate a sequence of segmented contracted functions with a decreasing number of primitive functions.

The P-orthogonalization procedure has the advantage that it allows a selection of balanced segmented basis sets, where the number of primitive functions in each contracted function is defined by a single parameter (C_{ortho}). This parameter furthermore ensures that there is a balance between the number of primitive functions used in different types (s-, p-, d-, ..) of contracted functions. There is no guarantee that the same segmentation is produced by a given C_{ortho} value for different elements, and selection of a consistent segmentation for elements in the same row or block in the periodic system must be done manually. This is in practice unproblematic, as only C_{ortho} values near a switch between different segmentation schemes leads to different segmentations for different elements.

dx.doi.org/10.1021/ct401026a | J. Chem. Theory Comput. 2014, 10, 1074−1085

Table 2. Basis Set Composition As a Function of C_{ortho} and the Associated BLYP Energy (Hartree) for the Ar pc-2 Basis Set[a]

		primitive functions						
	C_{ortho}	1s	2s	3s	2p	3p	energy	ΔE
uncontracted							−527.54717	0
general		9	9	9	7	7	−527.54717	0.00000
segmented	1×10^{-5}	9	6	3	7	4	−527.54796	−0.00079
	2×10^{-5}	9	5	3	7	4	−527.54796	−0.00079
	5×10^{-5}	9	5	3	7	3	−527.54779	−0.00063
	2×10^{-4}	**9**	**4**	**3**	**7**	**2**	**−527.54752**	**−0.00036**
	1×10^{-3}	9	3	3	7	2	−527.54678	0.00039
	4×10^{-3}	9	1	1	7	2	−527.15266	0.39451
	6×10^{-3}	9	1	1	7	1	−527.15049	0.39667

[a]**Bold** indicates the largest C_{ortho} value where the energy still is lower than the energy for the uncontracted basis set.

A given C_{ortho} value leads to a specific segmented contracted basis set which, after combined exponent and coefficient optimization, allows a quantification of the resulting contraction error relative to the results using the original set of primitive functions. How large a contraction error that is deemed acceptable is a user decision, and it should be recognized that different properties have different sensitivity to contraction. As the *P*-orthogonalization primary focus is to eliminate near-redundancy without reducing the total function space, we suggest that the total energy and exponent range of the primitive functions should be used as guidelines for deciding the maximum allowed contraction error. The switch in segmentation level as a function of C_{ortho} often removes several primitive functions at the same time, and it is recommended that a limited search is carried out before the final decision of segmentation level, as illustrated in the next section.

The simultaneous optimization of exponents and contraction coefficients is a highly nonlinear optimization problem with multiple local minimum.[53] Careful comparison of both exponents and coefficients with those from the P-orthogonalization procedure is necessary to ensure that the desired minimum has been located. Although the contraction coefficients from the P-orthogonalization provide a good starting point for the optimization, it is in many cases necessary to employ a sequence of constrained optimizations to guide the optimization, especially for the larger basis sets where the energy is a very flat function of the parameters. A change in sign for the coefficients indicates a node in the AO, and the optimized exponents will display a gap at the nodes. It is important that the distribution of optimized exponents for the primitive functions and the sign of the corresponding coefficients is the same as for the nonoptimized set of P-orthogonalized parameters. It is furthermore essential to compare exponents and coefficients for all elements in a row in the periodic table to ensure that same minimum has been obtained.

ILLUSTRATION OF THE SEGMENTATION PROCEDURE

The recommended procedure for defining the proper segmentation level is illustrated below for the Ar pc-2 basis set (valence TZP quality). The general contraction of the 13s10p set of primitive functions to 5s4p can be written as (11)(11)(11)11/8811, and the original Davidson purification procedure can be applied to eliminate two s- and one p-function from each of the contracted functions, leading to a 99911/7711 composition in terms of primitive functions. Table 2 shows how the number of primitive functions in each contracted function is reduced as the C_{ortho} value is increased, the associated BLYP energy for the corresponding exponent/coefficient optimized basis set, and the energy difference relative to the uncontracted basis set.

A C_{ortho} value of 2×10^{-4} leads to a 943/72 composition (compared to 999/77), which after exponent/coefficient optimization has a lower energy than the uncontracted 13s10p, while a C_{ortho} value of 1×10^{-3} leads to a 933/72 composition, which has a slightly higher energy. A further C_{ortho} increase leads to a 911/72 composition with an unacceptable energy penalty. Figures 1 and 2 show the optimized exponents

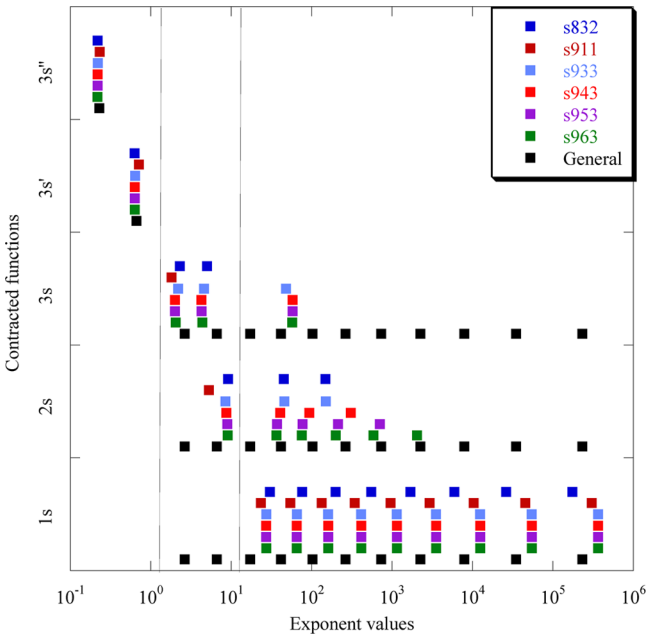


Figure 1. Optimized s-exponents for different segmented contractions of the Ar pc-2 basis set. Vertical lines indicate orbital nodes.

for the primitive basis functions at each segmentation level where the vertical lines indicate a change in sign for the AO-coefficients and thus the existence of an orbital node. It is evident that the optimization of the segmented basis sets opens an exponent gap near the nodes, and it is important to emphasize that of the many local minima in the combined exponent/coefficient space for each segmentation, only the one which has the same number of functions between each node as defined by the P-orthogonalization should be selected. It should also be noted that the total energy is not a suitable criterion for

dx.doi.org/10.1021/ct401026a | J. Chem. Theory Comput. 2014, 10, 1074−1085

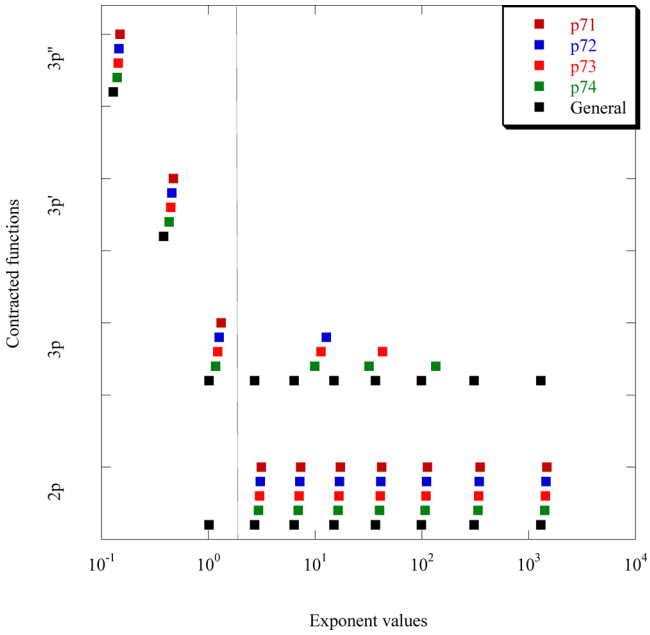


Figure 2. Optimized p-exponents for different segmented contractions of the Ar pc-2 basis set. The vertical line indicates the 2p-orbital node.

selecting the proper local minimum, as several other minima with lower energies often can be found at each segmentation level. These lower energy minima correspond to primitive basis functions with large(r) exponents that serve to improve the core region, which is energetically important, rather than describing the chemically important valence region.

The 943/72 segmentation is the smallest C_{ortho} selected composition that has a lower energy than the uncontracted pc-2 basis set, and it is used as a reference point for sequentially removing the energetically least important function, which after reoptimizing of the exponents and coefficients provides the energies shown in Table 3. The uncontracted pc-2 basis set has an energy 0.00387 hartree above the basis set limit (estimated from extrapolation of the pc-2, -3, -4 results), which defines the maximum allowed contraction error. The entries in Table 3 are arranged according to their energy contributions, and the energetically least important primitive function is thus the innermost s-function in the 3s-contracted function, followed by the innermost s-function in the 2s-contracted function, the outermost s-function in the 1s-contracted function, etc. Table 3 shows not only that removal of one primitive function from each of the three s-contractions leads to an acceptable contraction error but also that the 72 segmentation for the p-functions cannot be further reduced. The final selected segmented pc-2 basis set is thus 83211/7211 in composition, and Figures 1 and 2 show the exponent values corresponding to the proper local exponent/coefficient minimum.

Table 3. Basis Set Composition and Associated BLYP Energy (Hartree) for the Ar pc-2 Basis Set As a Function of Number of Primitive Functions in Each Contracted Function[a]

	primitive functions						
	1s	2s	3s	2p	3p	energy	ΔE
basis set limit						−527.55104	−0.00387
uncontracted						−527.54717	0
segmented	9	4	3	7	2	−527.54752	−0.00036
	9	4	2	7	2	−527.54749	−0.00032
	9	3	2	7	2	−527.54675	0.00042
	8	**3**	**2**	**7**	**2**	**−527.54510**	**0.00206**
	8	3	2	7	1	−527.54295	0.00421
	8	3	2	6	1	−527.54083	0.00634
	8	3	1	6	1	−527.53614	0.01103

[a]**Bold** indicates the smallest number of primitive functions where the energy difference relative to the value for the uncontracted basis set is smaller than the difference between the uncontracted and basis set limit.

SEGMENTED POLARIZATION CONSISTENT BASIS SETS

The above procedure illustrated for the Ar pc-2 basis set has been applied for all pc-*n* (n = 0,1,2,3,4) basis sets for the first 36 elements in the periodic table. In the analysis for selecting the final segmentation levels, we have considered it important that the same segmentation is employed for all elements sharing the same occupied orbitals, in order to achieve a good basis set balance. The same composition for the s-functions is thus employed for the s- and p-block elements in the first (Li−Ne) and second (Na−Ar) row elements and for the s-, p-, and d-block elements (K−Kr) in the third row. Similarly, the same composition for the p-functions is employed for the s- and d-block elements (K−Zn) in the third row. This choice in some cases lead to minor complications for the first element in each block, where only a single electron occupies the valence orbital. The exponent/coefficient optimization in these cases often leads to parameter values that differ from the expected values based on extrapolation from the subsequent elements. For the smaller basis sets (pc-0,-1,-2) the problem arises because the optimization attempts to improve the energetically important core orbital(s) by using the valence functions, at the expense of the energetically relatively unimportant single valence electron. The larger basis sets (pc-3,-4) approach completeness, and the energy as a function of the basis set parameters that describe the single valence electron is a very flat surface, and the optimum parameter values are consequently poorly defined. We have resolved these issues by typically fixing a single contraction coefficient or exponent at the value extrapolated from the next two elements, such that all basis set parameters display a smooth variation with the atomic number.

The pc-*n* basis sets have been developed over a decade in time, and the present reoptimization for all the first 36 elements gave the opportunity to reconsider the previous choice for the polarization functions for the first row s-block elements (Li, Be). These where originally assigned based on explicit optimized exponents for a collection of molecular systems.[40] For the third row s-block elements (K, Ca), however, we found that the d-type polarization functions could be obtained by extrapolation of the (occupied) d-functions for Sc and Ti.[41] We have employed the same idea in the present case, by assigning p-type polarization functions for Li and Be by a linear extrapolating of optimized basis sets for C and B containing the same number of p-functions. This means that the polarization contraction coefficients are now defined by extrapolation, rather than taken from excited state calculations. This choice provides polarization functions in good agreement with the previous assignments and may serve as a guideline for future extensions of the pc-*n* basis sets to other elements or for constructing new basis sets. First level polarization functions for elements where core orbitals of the same type is present (p-

Table 4. Comparison of TZP Quality Basis Sets for Ca[a]

				contraction				
basis set	contracted	primitive	dupl	s	p	$E_{contracted}$	$E_{uncontracted}$	ΔE
pcg-2	6s4p	16s11p		14-14-14-14-1-1	9-9-1-1	−677.59264	−677.59304	0.00040
pcs-2	6s4p	19s14p	3s3p	10-4-2-1-1-1	8-4-1-1	−677.59145	−677.59369	0.00224
Karlsruhe	6s5p	17s12p	3s1p	8-4-2-1-1-1	6-3-1-1-1	−677.57621	−677.58450	0.00829
Sapporo	8s6p	18s13p	3s1p	7-4-2-1-1-1-1-1	7-2-1-1-1-1	−677.57467	−677.58063	0.00596
Jorge	9s6p	16s13p	none	7-1-1-1-2-1-1-1-1	7-1-1-2-1-1	−677.57922	−677.59294	0.01372
Pople	8s7p	14s11p	none	6-2-1-1-1-1-1-1	3-3-1-1-1-1-1	−677.57597	−677.58413	0.00816

[a]Dupl indicates the number of duplicated primitive functions. $E_{(un)contracted}$ is the B3LYP energy for the contracted and uncontracted basis sets, and ΔE is the contraction error (Hartrees).

Table 5. Comparison of QZP Quality Basis Sets for Kr[a]

				contraction					
basis set	contracted	primitive	dupl	s	p	d	$E_{contracted}$	$E_{uncontracted}$	ΔE
pcg-3	7s6p4d	20s16p10d		17-17-17-17-1-1-1	13-13-13-1-1-1	8-1-1-1	−2753.85045	−2753.85110	0.00065
pcs-3	7s6p4d	28s20p10d	8s4p	12-7-4-2-1-1-1	10-5-2-1-1-1	8-1-1-1	−2753.85021	−2753.85102	0.00081
Karlsruhe	11s7p4d	24s20p10d	4s4p	11-4-1-1-1-1-1-1-1-1-1	10-5-1-1-1-1-1	7-1-1-1	−2753.84392	−2753.85078	0.00686
Sapporo	7s6p4d	19s14p10d	4s2p	7-4-3-2-1-1-1	7-3-1-1-1-1	7-1-1-1	−2753.78329	−2753.80394	0.02065
Jorge	10s7p4d	17s14p7d	none	8-1-1-1-1-1-1-1-1-1	7-1-2-1-1-1-1	4-1-1-1	−2753.73030	−2753.84400	0.11370

[a]Dupl indicates the number of duplicated primitive functions. $E_{(un)contracted}$ is the B3LYP energy for the contracted and uncontracted basis sets, and ΔE is the contraction error (Hartrees).

type polarization for Na, Mg and K−Zn, and d-type polarization for Ga−Kr) have been taken from the original pc-*n* basis sets (assigned by scaling the outermost exponent), and higher order polarization functions have also been taken from the original pc-*n* basis sets.

For the first row p-block elements we originally employed a 5Z contraction for the pc-3 and a 7Z contraction for the pc-4 basis sets, which gave (very) low contraction errors.[37] We have in the present case adjusted the contraction level for these basis sets to conform with the second and third row elements, such that pc-3 now uniformly employs a QZP valence, and pc-4 employs a 5ZP valence set of functions. This slightly increases the contraction error for the first row p-block elements, but brings the errors in line with those for the other elements.

The unpolarized pc-0 basis set was intended as a computationally cheap basis set which could be useful for preliminary qualitative investigations and for use in multilevel calculations for the outer-level atoms. The number of primitive functions was based on extrapolation from the pc-1 and pc-2 basis sets and general contracted to valence DZ. While the number of contracted functions qualified the pc-0 as DZ, the basis set exponents in some cases suggest that the valence space is of only single-zeta quality. We have in the present work redefined the pcs-0 basis set such that it represents the smallest basis set which can be optimized where the basis set exponents are of valence DZ quality. This significantly lowered the basis set errors for atomization energies, and since polarization functions for the s-block elements are assigned based on error balancing, this consequently lead to minor adjustments of the polarization functions. For the first row s-block elements (Li and Be) the polarization space was extended from a single p-function to two primitive p-functions contracted to one. Similarly, the pcs-0 basis set for third row s-block elements (K and Ca) now includes two primitive d-functions contracted to one.

Tables S1−S3 (Supporting Information) show the composition of all the present basis sets and compare with other commonly used basis sets. Tables 4 and 5 show illustrative examples for TZP and QZP quality basis sets for Ca and Kr, respectively, in terms of primitive and contracted functions for the general and segmented contracted versions of pc-3 (pcg-3 and pcs-3), and comparison to the corresponding Karlsruhe, Sapporo, Jorge, and Pople segmented basis sets. The advantages of the P-orthogonalization procedure for generating segmented basis sets increase with the size of the basis and atomic number. In order to ensure TZP quality results for Ca, the Sapporo, Jorge, and Pople basis sets are actually valence 5−6Z for the s- and p-functions, rather than valence TZ (Table 4). Similarly, Table 5 shows that the Karlsruhe, Sapporo, and Jorge Kr basis sets are valence 6−8Z for the s-functions and 5−6Z for the p-functions, rather than valence QZ. These expanded valence spaces are results of analyzing different contractions schemes for the underlying primitive functions and illustrate the problem of identifying the optimum contraction from the large combinatorial space of partitioning primitive into contracted functions.[53] In order to quantify the contraction errors we have employed a different functional (B3LYP[63,64]) than used in the construction of the basis sets (BLYP). The difference between the energies obtained with the contracted and uncontracted basis sets shows that the pcs-*n* basis set has small contractions errors, while small contraction errors for the other basis sets are achieved by uncontracting a larger part of the valence primitive functions. The Jorge basis sets display significant contractions errors, despite that they also have more uncontracted valence functions than the pcs-*n* basis set. The Kr pcs-3 basis set contains 12 (8s4p) primitive functions more than pcg-3 (Table 5), corresponding to functions that need to be duplicated between different contracted functions. This increase should be compared to the need for duplicating 17 s-functions three times and 13 p-functions twice, corresponding to 77 (51s26p) extra functions, when using the pcg-3 basis set in programs designed for segmented basis sets.

COMPUTATIONAL DETAILS

All basis set optimizations have been done using a pseudo-Newton−Raphson algorithm in connection with gradients of the BLYP[65,66] energy with respect to basis set exponents and

contraction coefficients generated by numerical differentiation. Third row transition metals have been optimized in their s^2d^n electronic configuration.[42] Benchmark calculations have been done using the Gaussian-09,[44] Gamess-US,[67] and Dalton2011[68] program packages. The Gaussian implementation of the Davidson purification procedure[62] has been employed for the general contracted pcg-*n* and cc-pVXZ basis sets. The cc-pVXZ basis sets for second row elements include the additional tight d-function.[17] Basis set limiting values for atomization energies have been generated by exponential extrapolation[69,70] of the uncontracted pc-2,-3,-4 results and are estimated to be accurate to within ∼0.1 kJ/mol.

BENCHMARKS

There are already a number of different basis set families in common use, and new basis sets should only be introduced if they provide lower basis set errors for a similar computational cost, or similar basis set errors for a lower computational cost, than existing alternatives. We have in previous work shown that the general contracted pc-*n* basis sets provide lower basis set errors with DFT methods for a given number of contracted basis functions, compared to other basis sets.[39−42,71] In the present section, we compare the performance of the new segmented pc-*n* basis sets (denoted pcs-*n*) to the general contracted pc-*n* (denoted pcg-*n*) and to other commonly used basis sets and provide relative computational times for three representative systems and three widely used program packages. The performance of a given basis set depends both on the number of primitive and contracted functions for the (occupied) atomic s-, p-, and possibly d-functions as well as the polarization functions, and these differ for each basis set. The polarization functions for the pcg-*n* and pcs-*n* basis sets are nearly identical, and these results can thus be compared directly.

Figures 3 and 4 show the mean absolute deviations (MAD) compared to the basis set limit for BLYP atomization energies relative to diatomic reference systems (AE2) for the molecules in Table 6 as a function of the average number of contracted and primitive functions, respectively. The AE2 MAD values are representative for typical applications where relative stabilities of different molecular systems are investigated. The pcg-*n* and pcs-*n* basis sets have almost the same number of contracted functions (pcs-3 and pcs-4 have slightly less for first row elements), and Figure 3 shows that they have virtually the same AE2 MAD and that this measure of the basis set error is lower than other basis sets for the same number of functions. The pcs-0 basis set has significantly smaller errors than pcg-0, due to the redesign discussed in the "segmented polarization consistent basis set" section. Figure 4 shows the same AE2 MAD displayed as a function of the average number of primitive functions. The pcs-*n* graph is displaced toward the left compared to the pcg-*n* graph, which reflects the reduced number of primitive basis functions. The Karlsruhe Def2 basis sets[43] display comparable performance to the pcs-*n* (but have more contracted functions, Figure 3 and Tables 4 and 5), while the other basis sets have significantly larger errors relative to the basis set limit. Tables S4−S13 (Supporting Information) show a breakdown of the AE2 MAD values over the s-, p-, and d-blocks for the first, second, and third row compounds as well as the corresponding maximum absolute deviations. The Pople style 6-31G(d) and 6-311(2df) basis sets[4−15] perform particularly poor for the transition metal and third row s-block compounds, which also have relatively large errors with

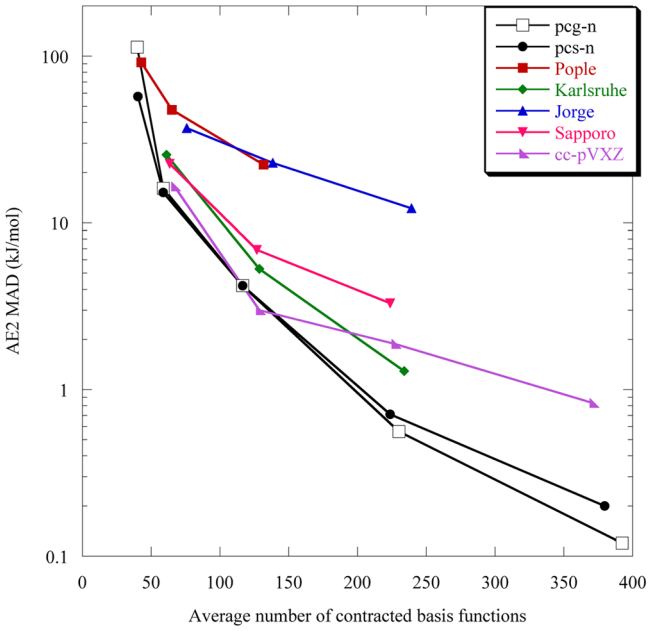


Figure 3. Mean absolute deviation relative to the basis set limit for atomization energies relative to diatomic reference energies as a function of the average number of contracted function for the systems in Table 6.

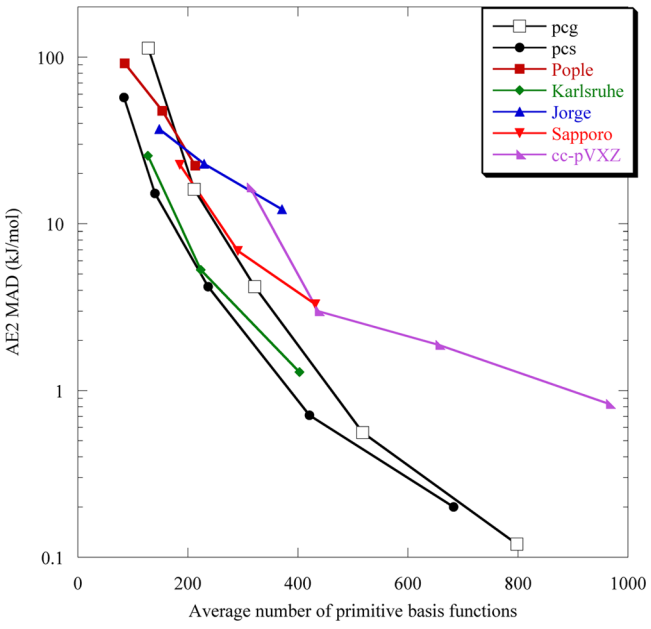


Figure 4. Mean absolute deviation relative to the basis set limit for atomization energies relative to diatomic reference energies as a function of the average number of primitive function for the systems in Table 6.

the Jorge DZP and TZP basis sets.[33−36] The Sapporo basis sets[24−26] have significantly larger errors for the s-block compound than for the p- and d-block systems. Tables S14−S15 show the corresponding performance for the B3LYP[63,64] and wB97X[72] exchange-correlation functionals, and they are very similar to the BLYP results. This indicates that pcs-*n* basis sets should be applicable to DFT methods in general. Table S16 shows that the basis set convergence of the HOMO energy follows the trend for atomization energies, and the pcs-*n* basis

dx.doi.org/10.1021/ct401026a | J. Chem. Theory Comput. 2014, 10, 1074−1085

Table 6. Benchmark Molecular Systems

Allene, ethanamid, aniline, aziridine, B_2H_6, B_2O_3, $B_3N_3H_6$, $B(NH_2)_3$, benzene, BF_3, BH_3, bicyclobutane, $B(CH_3)_3$, $B(OH)_3$, butadiene, butane, butyne, C_2F_4, C_2H_2, C_2H_4, C_2H_6, C_2N_2, C_3H_4, toluen, cyclobutane, cyclobutene, CF_2, CF_3CN, CF_4, CH_2F_2, CH_3CHO, CH_3CN, CH_3F, CH_3NH_2, CH_3NO_2, CH_3OF, CH_3OH, CH_3ONO, CH_4, CHF_3, CO_2, CO, cyclopropane, cyclopropene, ethylamin, ethanol, F_2CO, F_2O, furan, glyoxal, H_2BNH_2, H_2CCO, H_2CO, H_2O_2, H_2O, HCN, HCOOH, $HCOOCH_3$, HF, HNO, HOF, iso-propanol, iso-butane, iso-butene, $(CH_3)_2CO$, $(CH_3)_2NH$, $(CH_3)_2O$, $(CH_3)_3N$, CH_3CFO, CH_3COOH, methylcyclopropan, ethyl-methylether, N_2H_2, N_2H_4, N_2O, NF_3, NH_3, O_3, Oxirane, phenol, propane, propene, pyridine, pyrrole, spiro-pentane, C_2H_3CN, C_2H_3F, Be_2H_4, BeC_2, $BeCl_2$, BeF_2, BeH_2, $Be(CH_3)_2$, BeO, $BeOH_2$, BeS, $Be(SH)_2$, $BeSi_2$, Li_2C_2, Li_2NH, Li_2O_2, Li_2O, Li_2PH, Li_2S_2, Li_2Si_2, LiCl, LiF, LiH, $LiCH_3$, $LiNH_2$, LiOH, $LiPH_2$, LiSH, CH_3COCl, Al_2Cl_6, Al_2H_6, Al_2O_3, Al_2S_3, $Al_3(PH_2)_3$, $AlCl_3$, AlF_3, AlH_3, AlP_3, $Al(SH)_3$, $AlSi_3$, C_2Cl_4, CCl_4, CF_3Cl, CH_2Cl_2, CH_3Cl, CH_3PH_2, CH_3SH, CH_3SiH_3, $CHCl_3$, Cl_2SO_2, ClF_3, ClF, ClNO, CS_2, CSO, CS, $(CH_3)_2SO_2$, $(CH_3)_2SO$, $(CH_3)S$, H_2AlPH_2, H_2CSS, H_2S, HCl, HOCl, P_4, PCl_3, PCl_5, PF_3, PF_5, PH_3, $POCl_3$, S_2Cl_2, SCl_2, SF_6, Si_2H_2, Si_2H_4, Si_2H_6, $SiCl_2$, $SiCl_4$, SiF_4, SiH_4, SiO, SiS, SO_2, SO_3, thiirane, thiophene, C_2H_3Cl, $C_2H_3PH_2$, C_2H_3SH, $C_2H_3SiH_3$, CH_3MgCl, Mg_2H_4, MgC_2, $MgCl_2$, MgF_2, MgH_2, $Mg(CH_3)_2$, MgO, $Mg(OH)_2$, MgS, $Mg(SH)_2$, Na_2C_2, Na_2NH, Na_2PH, Na_2O_2, Na_2O, Na_2S, NaCl, NaF, NaH, $NaCH_3$, $NaNH_2$, $NaPH_2$, NaOH, NAsH, GaH_3, GeH_4, AsH_3, SeH_2, BrH, $GaLi_3$, $GeLi_4$, $AsLi_3$, $SeLi_2$, BrLi, GaF_3, GeF_4, AsF_3, SeF_2, BrF, $GaCl_3$, $GeCl_4$, $AScl_3$, $SeCl_2$, BrCl, GaGe, GaAs, GASe, GaBr, GeAs, GeSe, GeBr, AsSe, AsBr, SeBr, $GaBr_3$, $GeBr_4$, $AsBr_3$, $SeBr_2$, $AsBr_5$, As_4, $GeSe_2$, Se_2Br_2, KH, KLi, KCH_3, K_2O, K_2O_2, KF, K_2S, K_2S_2, KCl, CaH_2, $CaLi_2$, $Ca(CH_3)_2$, CaO, $Ca(OH)_2$, CaF_2, CaS, $Ca(SH)_2$, $CaCl_2$, GaK_3, GeK_4, AsK_3, K_2Se, K_2Se_2, KBr, CaK_2, CaSe, $Ca(SeH)_2$, $CaBr_2$, CuH, ScH_2, TiH_2, VH_2, CrH_2, MnH_2, FeH_2, CoH_2, NiH_2, CuH_2, ZnH_2, ScH_3, VH_3, CrH_3, MnH_3, FeH_3, CoH_3, TiH_4, VH_4, VH_5, CuCl, $ScCl_2$, $TiCl_2$, VCl_2, $CrCl_2$, $MnCl_2$, $FeCl_2$, $CoCl_2$, $NiCl_2$, $CuCl_2$, $ZnCl_2$, $ScCl_3$, VCl_3, $CrCl_3$, $MnCl_3$, $FeCl_3$, $CoCl_3$, $TiCl_4$, VCl_4, VCl_5, CuBr, $ScBr_2$, $TiBr_2$, VBr_2, $CrBr_2$, $MnBr_2$, $FeBr_2$, $CoBr_2$, $NiBr_2$, $CuBr_2$, $ZnBr_2$, $ScBr_3$, VBr_3, $CrBr_3$, $MnBr_3$, $FeBr_3$, $CoBr_3$, $TiBr_4$, VBr_4, VBr_5, CuLi, $TiLi_2$, $CrLi_2$, $FeLi_2$, $CoLi_2$, $NiLi_2$, $CuLi_2$, $ZnLi_2$, Cu_2O, ScO, TiO, VO, CrO, MnO, FeO, CoO, NiO, CuO, ZnO, TiO_2, MnO_2, CrO_3, Fe_2O_3, V_2O_5, Mn_2O_7

sets again have lower errors than other types of basis sets of similar quality.

Table 7 (extract of Tables S5 and S6) shows that the pcs-*n* basis sets display a uniform performance for first, second, and third row compounds across the s, p-, and d-blocks elements but that hypervalent second row molecules have significantly larger basis set errors at the pcs-1 and pcs-2 levels than the remaining molecules in the test set, and this slower basis set convergence has also been noted by Martin.[73] Testing indicated that hypervalent atoms require a more extensive set of polarization functions than provided by the basis set at the DZP and TZP levels, while it is adequate at the QZP level. We have been unable to find a simple general fix akin to the addition of a single d-function with a large exponent, as for the cc-pVXZ basis sets for second row atoms.[17] A heuristic fix is to simply employ a basis set of one higher quality for the hypervalent atom, as shown by the values in parentheses where the pcs-(n+1) basis set has been used on the hypervalent atom. Table 7 shows that this mixed basis set approach brings the error for the hypervalent molecules into line with the other systems for a given basis set quality. A similar and related problem occurs for molecules composed of elements with very different electronegativity (as for example MgF_2, $Mg(OH)_2$, NaF, NaOH, AlF_3, Al_2O_3), where the basis set error at the pcs-1 and pcs-2 levels is significantly larger than for less ionic compounds. The mixed basis set approach, where one higher quality basis set is employed for the electropositive atom(s), can also be employed for such systems to reduce the basis set error to the expected value at the given basis set quality. The higher errors for the second row s-block elements in Table 7 is due to the presence of relatively many ionic Na and Mg compounds in the test set (Table 6).

Table 7. Mean Absolute Deviations of Atomization Energies Relative to Diatomic Reference Systems Relative to the Basis Set Limit (kJ/mol) for the Molecules in Table 6 Grouped According to Rows and Periods[a]

		pcs-1	pcs-2	pcs-3
1. row	s-block	9.3	3.5	0.37
	p-block	8.8	2.7	0.79
2. row	s-block	25.5	7.0	0.86
	p-block	13.3	5.0	0.78
	p-block, hypervalent	108.6 (8.9)	19.7 (5.0)	0.43
3. row	s-block	7.4	2.8	0.86
	d-block	15.4	3.9	0.80
	p-block	10.1	3.2	0.35
all		15.2	4.2	0.56

[a]Values in parentheses employ the pcs-(n+1) basis set for the hypervalent atom and pcs-*n* for the other atoms.

Table 8 shows the ratio between the computational time for a B3LYP[63,64] energy calculation with different basis sets relative to pcs-*n* for the O_{12}, S_{12}, and Se_{12} molecules in a crown conformation using three popular computational packages: Gaussian-09,[44] Gamess-US,[67] and Dalton2011.[68] Note that the timing comparison is only between different basis sets with the same program package and molecular system and not between computational packages or molecular systems. The calculations have been run on an 8-core server and all employ the direct SCF option with integral screening[55−57] and same convergence criterion. Timing ratios include a normalization for the (possible) difference in the number of SCF cycles. Table 8 shows that the pcs-*n* basis sets are computationally more efficient than the pcg-*n*, as expected since they have a reduced number of primitive functions. The improvement, however, depends significantly on the molecule, the basis set quality, and the program package and varies between a few percent and up to an order of magnitude. Most of the other basis sets are computationally less efficient, and we only comment on a few cases with ratios below 0.80 (corresponding to 20+% faster computational time). For the first row example (O_{12}) the Pople style 6-31G(d) basis set is more efficient than pcs-1 with the Gamess program package (but not with Gaussian-09 or Dalton), while the 6-311G(2df) basis set is more efficient than pcs-2 with all three program packages. This is primarily a consequence of the Pople style basis sets having a larger exponent for the outermost p-function, which leads to more efficient integral screening. Figure 3 shows that this improved computational efficiency comes at a price, as the lack of sufficiently diffuse p-functions leads to larger basis sets errors. At the DZP level, the pcs-1 AE2 MAD is 8.8 kJ/mol for first row p-block molecules compared to 14.7 kJ/mol with the 6-31G(d) basis set (Table S5). The corresponding values at the TZP level (pcs-2 and 6-311G(2df)) are 2.7 and 7.1 kJ/mol, respectively (Table S6). While the 6-311G(2df) basis set is computationally faster than pcs-2 by ∼30−40%, it actually only deliver results of the same quality as the pcs-1, and the pcs-1 is computationally almost an order of magnitude faster than 6-311G(2df). The same effect is present for the Jorge and cc-pVXZ basis sets at the QZP level, where Table 8 shows a better computational efficiency compared to pcs-3 but at the same time leads to larger basis set errors (Table S7, AE2 MAD values of 12.2 and 1.9 kJ/mol, respectively, compared to 0.7 kJ/mol for pcs-3). There are fewer basis sets that are computationally more efficient than pcs-*n* for the second and third row examples (S_{12} and Se_{12}), and these are again related to the exponent value

Table 8. Timing Ratios for Three Different Molecules (O_{12}, S_{12}, Se_{12}) with Three Different Program Packages and Three Basis Set Qualities

	basis set								
	DZP			TZP			QZP		
	G09[a]	GMS[a]	DLT[a]	G09[a]	GMS[a]	DLT[a]	G09[a]	GMS[a]	DLT[a]
O_{12}									
pcg-*n*	1.10	1.03	1.50	1.04	1.03	1.55	1.10	1.16	1.80
pcs-*n*	1	1	1	1	1	1	1	1	1
Karlsruhe	0.96	0.91	0.93	1.07	0.92	0.95	0.66	0.64	0.60
Pople	1.15	0.79	1.13	0.89	0.55	0.77			
Jorge	1.13	1.17	1.19	1.04	0.83	0.88	0.52	0.51	0.47
Sapporo	1.09	1.35	1.47	1.24	1.74	1.84	1.50	1.32	1.35
cc-pVXZ	1.01	1.24	1.74	0.94	0.88	1.30	0.53	0.55	0.71
S_{12}									
pcg-*n*	1.24	1.47	4.50	1.07	1.69	4.75	1.36	1.54	3.94
pcs-*n*	1	1	1	1	1	1	1	1	1
Karlsruhe	0.96	0.65	0.60	1.78	1.52	1.27	1.37	1.43	1.55
Pople	1.09	0.62	1.52	1.59	1.27	2.23			
Jorge	2.06	1.34	1.15	1.88	1.47	1.26	0.93	0.95	0.76
Sapporo	1.21	1.55	2.02	1.47	3.48	2.76	1.84	2.47	1.92
cc-pVXZ	2.13	3.30	7.97	2.85	3.66	6.96	1.39	1.76	2.91
Se_{12}									
pcg-*n*	1.42	3.65	10.59	2.53	3.16	6.19	1.73	2.51	3.61
pcs-*n*	1	1	1	1	1	1	1	1	1
Karlsruhe	1.57	1.17	0.68	1.04	1.12	0.72	0.97	0.97	0.74
Pople	1.28	0.98	0.68	1.17	1.22	3.86			
Jorge	2.23	1.37	0.67	1.15	0.81	0.34	1.15	0.67	0.42
Sapporo	1.29	1.83	6.23	2.02	2.67	3.74	1.52	1.68	0.88
cc-pVXZ	1.68	4.52	12.32	3.49	4.63	7.63	2.05	3.05	4.50

[a]The programs are as follows: G09 = Gaussian-09, GMS = Gamess-US, DLT = Dalton.

of the most diffuse function and the resulting difference in integral screening efficiency. Table 8 in connection with Figure 3 show that the pcs-*n* basis sets are among the computationally most efficient of the commonly used basis sets and at the same time provide the lowest basis set errors. In combination this suggests that the pcs-*n* basis sets provide a favorable performance/cost combination for DFT calculations in general.

SUMMARY

The present work shows that a general contracted basis set in a unique way can be converted into a computationally more efficient segmented contracted basis set, while inheriting the full accuracy of the general contracted basis set. The proposed P-orthogonalization procedure can be considered as removing the redundancy between primitive functions in the general contraction, and it provides a method for progressing from an uncontracted basis set, to a general contraction, and to a segmented contraction with full control of the contraction error. It furthermore uniquely identifies the proper local minimum in the combined exponent/coefficient parameter space, which previously has been an unsolved problem.

The P-orthogonalization procedure is used to convert the previously proposed general contracted polarization consistent basis sets optimized for density functional theory methods into segmented contracted versions. Benchmark calculations show that the new segmented polarization consistent basis sets (pcs-*n*) not only display the lowest basis set errors at a given zeta quality level but also that they are among the computationally most efficient. They provide a uniform error control of the basis set incompleteness for molecules composed of atoms from the first three rows in the periodic table (H−Kr) and for different exchange-correlation functionals. The pcs-*n* basis sets are available[74] in qualities ranging from (unpolarized) double-zeta to pentuple zeta quality and should therefore be well suited for both routine and benchmark calculations using density functional theory methods in general.

ASSOCIATED CONTENT

Supporting Information

Tables comparing basis set composition for first, second, and third row elements (Tables S1−S3), mean and maximum absolute deviations for the atomization energies for the molecules in Table 6 divided according to rows and block in the periodic table for the basis sets shown in Figures 1 and 2 (Tables S4−S13), mean and maximum absolute deviations for the atomization energies for the molecules in Table 6 for the pcs-*n* basis sets with the BLYP, B3LYP, and wB97X functionals (Tables S14−S15), mean absolute deviations for the HOMO energies for the molecules in Table 6 (Table S16). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: frj@chem.au.dk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by grants from the Danish Center for Scientific Computation and the Danish Natural Science Research Council.

■ REFERENCES

(1) Hill, J. G. *Int. J. Quantum Chem.* **2013**, *113*, 21−34.
(2) Jensen, F. *WIREs* **2013**, *3*, 273−295.
(3) Weigend, F.; Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *119*, 12753−12762.
(4) Hehre, W. J.; Ditchfie, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257−2261.
(5) Dill, J. D.; Pople, J. A. *J. Chem. Phys.* **1975**, *62*, 2921−2923.
(6) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; Defrees, D. J.; Pople, J. A. *J. Chem. Phys.* **1982**, *77*, 3654−3665.
(7) Rassolov, V. A.; Pople, J. A.; Ratner, M. A.; Windus, T. L. *J. Chem. Phys.* **1998**, *109*, 1223−1229.
(8) Rassolov, V. A.; Ratner, M. A.; Pople, J. A.; Redfern, P. C.; Curtiss, L. A. *J. Comput. Chem.* **2001**, *22*, 976−984.
(9) Binning, R. C.; Curtiss, L. A. *J. Comput. Chem.* **1990**, *11*, 1206−1216.
(10) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650−654.
(11) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639−5648.
(12) Curtiss, L. A.; McGrath, M. P.; Blaudeau, J. P.; Davis, N. E.; Binning, R. C.; Radom, L. *J. Chem. Phys.* **1995**, *103*, 6104−6113.
(13) Blaudeau, J. P.; McGrath, M. P.; Curtiss, L. A.; Radom, L. *J. Chem. Phys.* **1997**, *107*, 5016−5021.
(14) Wachters, A. J. *J. Chem. Phys.* **1970**, *52*, 1033−1036.
(15) Hay, P. J. *J. Chem. Phys.* **1977**, *66*, 4377−4384.
(16) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(17) Dunning, T. H.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244−9253.
(18) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; Dunning, T. H. *J. Chem. Phys.* **1999**, *110*, 7667−7676.
(19) Prascher, B. P.; Woon, D. E.; Peterson, K. A.; Dunning, T. H., Jr.; Wilson, A. K. *Theor. Chem. Acc.* **2011**, *128*, 69−82.
(20) Balabanov, N. B.; Peterson, K. A. *J. Chem. Phys.* **2005**, *123*, 064107.
(21) Zhong, S.; Barnes, E. C.; Petersson, G. A. *J. Chem. Phys.* **2008**, *129*, 184116.
(22) Barnes, E. C.; Petersson, G. A.; Feller, D.; Peterson, K. A. *J. Chem. Phys.* **2008**, *129*, 194115.
(23) Barnes, E. C.; Petersson, G. A. *J. Chem. Phys.* **2010**, *132*, 114111.
(24) Thakkar, A. J.; Koga, T.; Saito, M.; Hoffmeyer, R. E. *Int. J. Quantum Chem.* **1993**, 343−354.
(25) Tatewaki, H.; Koga, T.; Takashima, H. *Theor. Chem. Acc.* **1997**, *96*, 243−247.
(26) Koga, T.; Tatewaki, H.; Matsuyama, H.; Satoh, Y. *Theor. Chem. Acc.* **1999**, *102*, 105−111.
(27) Noro, T.; Sekiya, M.; Koga, T. *Theor. Chem. Acc.* **2003**, *109*, 85−90.
(28) Widmark, P. O.; Malmqvist, P. A.; Roos, B. O. *Theor. Chem. Acc.* **1990**, *77*, 291−306.
(29) Widmark, P. O.; Joakim, B.; Persson; Roos, B. O. *Theor. Chem. Acc.* **1991**, *79*, 419−432.
(30) Pouamerigo, R.; Merchan, M.; Nebotgil, I.; Widmark, P. O.; Roos, B. O. *Theor. Chem. Acc.* **1995**, *92*, 149−181.
(31) Pierloot, K.; Dumez, B.; Widmark, P. O.; Roos, B. O. *Theor. Chem. Acc.* **1995**, *90*, 87−114.
(32) Roos, B. O.; Lindh, R.; Malmqvist, P. A.; Veryazov, V.; Widmark, P. O. *J. Phys. Chem. A* **2004**, *108*, 2851−2858.
(33) Jorge, F. E.; Sagrillo, P. S.; de Oliveira, A. R. *Chem. Phys. Lett.* **2006**, *432*, 558−563.
(34) Barbieri, P. L.; Fantin, P. A.; Jorge, F. E. *Mol. Phys.* **2006**, *104*, 2945−2954.
(35) Machado, S. F.; Camiletti, G. G.; Canal Neto, A.; Jorge, F. E.; Jorge, R. S. *Mol. Phys.* **2009**, *107*, 1713−1727.
(36) Neto, A. C.; Muniz, E. P.; Centoducatte, R.; Jorge, F. E. *THEOCHEM* **2005**, *718*, 219−224.
(37) Jensen, F. *J. Chem. Phys.* **2001**, *115*, 9113−9125.
(38) Jensen, F. *J. Chem. Phys.* **2002**, *116*, 3502−3502.
(39) Jensen, F.; Helgaker, T. *J. Chem. Phys.* **2004**, *121*, 3463−3470.
(40) Jensen, F. *J. Phys. Chem. A* **2007**, *111*, 11198−11204.
(41) Jensen, F. *J. Chem. Phys.* **2012**, *136*, 114107.
(42) Jensen, F. *J. Chem. Phys.* **2013**, *138*, 014107.
(43) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−3305.
(44) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian-09*; 2009.
(45) Haettig, C.; Klopper, W.; Koehn, A.; Tew, D. P. *Chem. Rev.* **2012**, *112*, 4−74.
(46) Kong, L.; Bischoff, F. A.; Valeev, E. F. *Chem. Rev.* **2012**, *112*, 75−107.
(47) Peterson, K. A.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 084102.
(48) Hill, J. G.; Mazumder, S.; Peterson, K. A. *J. Chem. Phys.* **2010**, *132*, 054108.
(49) Hill, J. G.; Peterson, K. A. *Phys. Chem. Chem. Phys.* **2010**, *12*, 10460−10468.
(50) Raffenetti, R. C. *J. Chem. Phys.* **1973**, *58*, 4452−4458.
(51) Dunning, T. H. *J. Chem. Phys.* **1970**, *53*, 2823−2833.
(52) Campos, C. T.; Ceolin, G. A.; Canal Neto, A.; Jorge, F. E.; Pansini, F. N. N. *Chem. Phys. Lett.* **2011**, *516*, 125−130.
(53) Jensen, F. *J. Chem. Phys.* **2005**, *122*, 074111.
(54) Ceolin, G. A.; de Berredo, R. C.; Jorge, F. E. *Theor. Chem. Acc.* **2013**, *132*, 1339.
(55) Almlof, J.; Faegri, K.; Korsell, K. *J. Comput. Chem.* **1982**, *3*, 385−399.
(56) Haser, M.; Ahlrichs, R. *J. Comput. Chem.* **1989**, *10*, 104−111.
(57) Lambrecht, D. S.; Ochsenfeld, C. *J. Chem. Phys.* **2005**, *123*, 184101.
(58) Werner, H.-J.; Schuetz, M. *J. Chem. Phys.* **2011**, *135*, 144116 DOI: http://dx.doi.org/10.1063/1.3641642.
(59) Kristensen, K.; Hoyvik, I.-M.; Jansik, B.; Jorgensen, P.; Kjaergaard, T.; Reine, S.; Jakowski, J. *Phys. Chem. Chem. Phys.* **2012**, *14*, 15706−15714.
(60) Schutz, M.; Lindh, R.; Werner, H. J. *Mol. Phys.* **1999**, *96*, 719−733.
(61) Karton, A.; Martin, J. M. L. *J. Chem. Phys.* **2011**, *135*, 144119.
(62) Davidson, E. R. *Chem. Phys. Lett.* **1996**, *260*, 514−518.
(63) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(64) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623−11627.
(65) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100.
(66) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785−789.
(67) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., Jr. *J. Comput. Chem.* **1993**, *14*, 1347−1363.

(68) Angeli, C.; Bak, K. L.; Bakken, V. Christiansen, O.; Cimiraglia, R.; Coriani, S.; Dahle, P.; Dalskov, E. K.; Enevoldsen, T.; Fernandez, B.; Ferrighi, L.; Frediani, L. Hattig, C.; Hald, K.; Halkier, A.; Heiberg, H.; Helgaker, T.; Hettema, H.; Jansik, B.; Jensen, H. J. Aa.; Jonsson, D.; Jørgensen, P.; Kirpekar, S.; Klopper, W.; Knecht, S.; Kobayashi, R.; Kongsted, J.; Koch. H.; Ligabue, A.; Lutnæs, O. B.; Mikkelsen, K. V.; Nielsen, C. B.; Norman, P.; Olsen, J.; Osted, A.; Packer, M. J.; Pedersen, T. B.; Rinkevicius, Z.; Rudberg, E.; Ruden, T. A.; Ruud, K.; Salek, P.; Samson, C. C. M.; Sanchez de Meras, A.; Saue, T.; Sauer, S. P. A.; Schimmelpfennig, B.; Steindal, A. H.; Sylvester-Hvid, K. O.; Taylor, P. R.; Vahtras, O.; Wilson, D. J.; Ågren, H. *DALTON*.
(69) Jensen, F. *Theor. Chem. Acc.* **2005**, *113*, 267−273.
(70) Karton, A.; Martin, J. M. L. *Theor. Chem. Acc.* **2006**, *115*, 330−333.
(71) Jensen, F. *J. Chem. Phys.* **2002**, *116*, 7372−7379.
(72) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.
(73) Martin, J. M. L. *J. Mol. Struct.: THEOCHEM* **2006**, *771*, 19−26.
(74) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045−1052.