The Journal of Chemical Physics

Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases

Cite as: J. Chem. Phys. **127**, 114105 (2007); https://doi.org/10.1063/1.2770708
Submitted: 23 April 2007 • Accepted: 18 July 2007 • Published Online: 18 September 2007

Joost VandeVondele and Jürg Hutter

AIP Publishing

Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases

Joost VandeVondele and Jürg Hutter
Physical Chemistry Institute, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

(Received 23 April 2007; accepted 18 July 2007; published online 18 September 2007)

We present a library of Gaussian basis sets that has been specifically optimized to perform accurate molecular calculations based on density functional theory. It targets a wide range of chemical environments, including the gas phase, interfaces, and the condensed phase. These generally contracted basis sets, which include diffuse primitives, are obtained minimizing a linear combination of the total energy and the condition number of the overlap matrix for a set of molecules with respect to the exponents and contraction coefficients of the full basis. Typically, for a given accuracy in the total energy, significantly fewer basis functions are needed in this scheme than in the usual split valence scheme, leading to a speedup for systems where the computational cost is dominated by diagonalization. More importantly, binding energies of hydrogen bonded complexes are of similar quality as the ones obtained with augmented basis sets, i.e., have a small (down to 0.2 kcal/mol) basis set superposition error, and the monomers have dipoles within 0.1 D of the basis set limit. However, contrary to typical augmented basis sets, there are no near linear dependencies in the basis, so that the overlap matrix is always well conditioned, also, in the condensed phase. The basis can therefore be used in first principles molecular dynamics simulations and is well suited for linear scaling calculations. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2770708]

I. INTRODUCTION

A significant share of molecular calculations in quantum chemistry is nowadays based on density functional theory (DFT). The success of DFT can be attributed to its comparably low computational cost while maintaining in many cases a satisfactory accuracy. DFT is routinely used to describe the chemistry of molecular systems containing hundreds to thousands of atoms.[1–3] These models can include heavy elements, large ligands, and explicit solvent. Furthermore, dynamic and finite temperature effects can be investigated using molecular dynamics[4,5] or Monte Carlo simulations[6,7] based on DFT. For a given system, both fundamental as well as technical aspects limit the accuracy. The approximate nature of currently employed density functionals is a fundamental limitation that cannot be ignored in actual studies. The focus of the present work, however, is on a major technical aspect of calculations, namely, the basis set used to solve the Kohn-Sham equations. There exists a wide variety of used functional forms, including Gaussian functions,[8] Slater functions,[9] plane waves,[10,11] wavelets,[12] numerical basis functions,[13] and many others. Especially the literature on Gaussian basis sets is enormous, including reviews[14] and books,[15] thereby emphasizing their importance. For each of these functional forms, there are schemes to increase the size, and hence typically the accuracy of the basis. Often, there is no perfect recipe to do so, and different schemes will be used to obtain, for example, good total energies, geometries, interaction energies, or special electronic properties. Ultimately, all of these approaches try to balance computational cost and accuracy.

In the present work we are concerned with basis sets of contracted Gaussian functions. We are looking for Gaussian basis sets that are adapted for the use in large scale simulations of molecular systems, including gas phase systems, interfaces, and liquids, with good performance for total energies, geometries, and hydrogen bonding energies. The latter property, in particular, is sensitive to basis set superposition errors (BSSEs). The basis set superposition error is intrinsic to atom centered basis sets, but its magnitude can typically be reduced significantly for weakly bonded systems by augmenting the basis set with diffuse primitives. For total energies, excellent basis sets exist in the context of DFT. The polarization consistent basis sets by Jensen[16,17] are particularly noteworthy as these provide a sequence of basis sets that can be conveniently used to obtain results within a few μhartrees/at. of the basis set limit. Our focus on large systems imposes additional requirements on basis set size and on the condition number (ratio of largest to smallest eigenvalue) of the overlap matrix. At the same time the need to keep the number of primitive functions at a minimum, especially at high angular momentum, can be relaxed. Indeed, with the advent of highly efficient and linear scaling methods for the calculation of the Kohn-Sham matrix,[2,18–20] the computational cost of DFT calculations of large systems is dominated by the diagonalization of the Kohn-Sham matrix or, more generally, the density matrix update procedure. At fixed system size, all these update methods contain terms that scale at least quadratically, but typically cubically or even higher with number of basis functions,[18,21] hence the need for optimally small basis sets. The importance of the condition number of the overlap matrix is more indirect. The standard technique to deal with a near singularity of the overlap matrix is the removal of the eigenvectors corresponding to the small eigenvalues by diagonalization, and such an approach can thus not be adopted in a linear scaling context. Furthermore, such a scheme is problematic in the context of molecular dynamics or geometry optimization, since the number of eigenvalues below a given threshold might depend on the configuration. The above technique will then result in small discontinuities in the total energy as the system moves. Even if the condition number of the overlap matrix is sufficiently small to perform stable optimizations in a traditional approach, linear scaling methods benefit from a further reduction of the condition number because the sparsity of the inverse overlap matrix, and thus the density matrix, is directly related to this condition number. Linear scaling approaches that avoid the computation of the inverse typically rely on iterative methods to compute the required operators, and these exhibit slower convergence as the condition number increases.

In summary, an optimal basis set should fulfill the following requirements. High accuracy for smaller basis sets and a route for systematic improvements. One and the same basis set should perform in various environments from isolated molecules to condensed phase systems. Basis sets should lead for all systems to well conditioned overlap matrices and be therefore well suited for linear scaling algorithms. In order to fulfill all the above requirements, generally contracted[22] basis sets with shared exponents, the so-called family basis sets, for all angular momentum states are proposed in this work. In particular, a full contraction over all primitive functions is used for both valence and polarization functions. The set of primitive functions includes diffuse functions with small exponents. However, in contrast to the practice used in augmented basis sets, these primitive functions are always part of a contraction with tighter functions. The inclusion of diffuse functions is mandatory for the description of weak interactions, e.g., hydrogen bonding. Anticipating results and in line with earlier work on polarized atomic orbitals,[23–25] it is found that these highly contracted basis sets always lead to well conditioned overlap matrices. In order to derive optimal Gaussian exponents for the primitives and contraction coefficients for the basis sets, parameters are fully optimized based on molecular calculations.

II. BASIS SET CONSTRUCTION

The basis sets in this work have been derived for use with the analytical dual-space pseudopotentials proposed by Goedecker, Teter, and Hutter (GTH)[26] and as parametrized for various density functionals.[27,28] These accurate pseudopotentials can be used in conjunction with the Gaussian and plane wave (GPW) scheme,[19] as implemented in the CP2K/QUICKSTEP program.[2,29] In this work, all calculations have beenperformed with the density functional of Perdew-Burke-Ernzerhof.[30] The GPW method allows for accurate and fast density functional calculations of large systems. In particular, the Kohn-Sham matrix is evaluated in linear scaling time,[2] and for large systems, diagonalization or wave function optimization based on the orbital transformation method[1] dominate the computational cost. Obviously, basis sets constructed for use with these pseudopotentials cannot be used in all-electron calculations. Nevertheless, it is expected that the ideas presented here apply equally well to the all-electron case. An advantage of pseudopotentials is that only the valence electrons need to be described, so that the number of electrons, basis functions, and primitive Gaussians per atom are smaller than in the corresponding all-electron case. This can reduce significantly the cost of large scale calculations. In this work, basis sets for the elements H, C, N, O, F, Si, P, S, and Cl are presented. The library contains basis sets denoted m-SZV $(1s1p/1s)$, m-DZVP $(2s2p1d/2s1p)$, m-TZVP $(3s3p1d/3s1p)$, m-TZV2P $(3s3p2d/s2p)$, and m-TZV2PX $(3s3p2d1f/3s2p1d)$ for each of the above elements and is available as supplementary material[31] or can be downloaded from the CP2K website.[29] The "m-" prefix refers to the fact that the basis sets have been optimized in molecular calculations. To further clarify the notation, the number of radial functions for a given angular momentum is shown between parentheses for the nonhydrogen/hydrogen case. Spherical Gaussians are used throughout, i.e., 1, 3, 5, and 7 functions for s, p, d, and f symmetries, respectively. The basis is a generally contracted family basis set, meaning that all radial functions of all angular momenta are contractions of Gaussian functions with the same exponents and the contractions span all Gaussians primitives. The algorithms used in CP2K (Ref. 1) are suited to handle this kind of basis set efficiently. A further choice made for the library is that the larger basis sets extend the smaller ones identically. For example, by removing the f functions from the m-TZV2PX basis for O, the m-TZV2P basis for that element is obtained. Finally, whereas the Gaussian exponents and the contraction coefficients will be determined by optimization, the number of Gaussian primitive functions used for each atom needs to be fixed in advance. Based on previous experience[32] and additional tests, seven Gaussians for hydrogen and first row elements (H, C, N, O, F) have been used, while six Gaussians have been used for the second row elements (Si, P, S, Cl). This choice of the number of Gaussians leads in atomic calculations to total energies within $\approx 50 \cdot 10^{-6}$ hartree of the basis set limit. The choice is also consistent with the fact that the pseudized valence orbitals of the second row elements are smoother than the corresponding functions of first row elements.

All the Gaussian exponents (α_i) and contraction coefficients (c_i) are determined by optimizations based on molecular calculations. The objective function that is minimized $(\Omega(\{\alpha_i, c_i\}))$ is a linear combination of the total energy (E_{tot}) of a set of molecules M, as well as the logarithm of the condition number (κ) of the overlap matrix for these molecules as obtained with the different basis sets (B) in the library,

$$\Omega(\{\alpha_i, c_i\}) = \sum_B \sum_M \left(E_{\text{tot}}^{B,M}(\{\alpha_i, c_i\}) + \gamma \ln \kappa^{B,M}(\{\alpha_i, c_i\})\right). \tag{1}$$

The optimization of all basis sets concurrently avoids a bias of the Gaussian exponents, and to a lesser extend the contraction coefficients, toward optimal values for the smallest

TABLE I. Shown are the geometric mean errors in the total energies (in hartree per nonhydrogen atom) with respect to an estimate of the basis set limit, as obtained from fully uncontracted basis sets, including all exponents up to *g* functions. For each basis set, results for (A) molecules in the test set (H_2O, CH_4, CH_3CN, H_2S, SF_6, C_2H_4, CCl_2O), molecules not in the test set (B) at equilibrium geometries (CO_2H_2, CH_3NH_2, PH_3, $CHONH_2$, CO), and (C) transition states (H_2SiH_3, $ClCH_3Cl$, HCN) are provided separately.

	SZV	DZVP	TZVP	TZV2P	QZV2P	QZV3P	
A	0.132 918	0.015 071	0.006 427	0.003 290	0.002 423	0.001 939	
B	0.141 244	0.012 862	0.006 484	0.003 183	0.002 284	0.001 741	
C	0.109 298	0.013 537	0.006 253	0.003 480	0.002 402	0.001 826	
		aug-DZVP	aug-TZVP	aug-TZV2P	aug-QZV2P	aug-QZV3P	
A		0.012 042	0.005 783	0.002 845	0.002 266	0.001 821	
B		0.010 455	0.005 907	0.002 792	0.002 193	0.001 656	
C		0.009 270	0.004 836	0.002 429	0.001 999	0.001 417	
	m-SZV	m-DZVP	m-TZVP	m-TZV2P			m-TZV2PX
A	0.067 889	0.003 399	0.002 471	0.001 892			0.000 563
B	0.081 205	0.003 989	0.002 852	0.002 143			0.000 908
C	0.093 991	0.004 973	0.002 580	0.002 132			0.001 322

(m-SZV) basis. The term containing the condition number has been added to avoid the generation of basis sets that would be badly conditioned. It is found that by using a relatively small but arbitrary value for γ ($\gamma=0.001$ hartree), sufficient flexibility in the basis set is retained, while at the same time condition numbers stay small. The larger and more balanced the molecular reference set M is, the smaller the bias imposed on the basis, and the better the transferability of the basis will be. On the other hand, by restricting the set of molecules, higher accuracy for a limited class of compounds could be obtained, but thereby the transferability of the basis would be reduced. The selected reference set of 31 molecules explores a rather wide range of bonding patterns: NF_3, HNO_2, HF, HCl, H_2SO_4, H_2O, H_2O_2, H_2CO, F_3PO, $ClFO_3$, $CHCl_3$, CH_4, CH_3SiH_3, CH_3SH, CH_3PH_2, CH_3CN, CH_2PH, CH_2F_2, CCl_2O, C_6H_5OH, C_5H_5N, $C_4H_4N_2$, C_2H_4, C_2H_2, SiO_2, SiH_3F, Si_2H_6, SF_6, SF_4, P_2H_4, F_2. The five targeted basis sets require the optimization of 49 (C, N, O, F) / 42 (H, Si, P, S, Cl) parameters per atom or 406 parameters for the full library. The objective function has been optimized using unconstrained optimization by quadratic optimization, which requires only function evaluations.[33] Initial values for the exponents and coefficients have been derived from atomic calculations. No global optimization has been attempted. The direct rigorous minimization of the target function Eq. (1) is computationally not tractable, and therefore an approximate multistep procedure has been adopted. Instead of optimizing all parameters at the same time, optimizations for each element are performed separately. Parameters that define the m-SZV and m-DZVP basis sets are determined first, and the additional parameters for the m-TZVP, m-TZV2P, and m-TZV2PX basis sets are optimized afterwards. It is important to optimize the Gaussian exponents using a target function that includes at least the m-DZVP basis set. Only in this way a significant weight for the more diffuse functions is achieved. In order to balance the basis set for all elements, several complete cycles of optimizations over all elements have been performed. Finally, whereas all 31 molecules mentioned above have been employed in the optimization process, only four molecules have been used simultaneously. The optimization procedure outlined above required several million molecular wave function optimizations.

III. BASIS SET TESTING

In order to assess the quality of the basis sets generated, we summarize here a number of properties as computed with these basis sets. For comparison, we also provide results obtained using more traditional basis sets, i.e., split valence with uncontracted polarization functions, as obtained previously.[2] These basis sets have been constructed for GTH potentials and are available from the CP2K website. We also refer to plane wave calculations performed with the CPMD package[34] and to all-electron calculations performed using the Gaussian code[35] or found in literature.

A. Total energies

In Table I, the average error in the total energy per atom is shown for a number of molecules. Results have been grouped to distinguish between molecules present in the training set and molecules that are not. The latter category has been split in equilibrium geometries and transition states. A first observation is that the best results are obtained with the molecularly optimized basis sets; they outperform basis sets of the same or slightly larger sizes. The effect is most pronounced for the m-DZVP basis, which reduces the error by a factor of 3–4 with respect to a traditional basis of the same size (DZVP). The m-TZV2PX basis yields the best total energies, with errors below 1 millihartree per heavy atom, the effect of adding *f* functions is particularly significant for SF_6. There is some difference for molecules that have been and molecules that have not been part of the fitting set. In general, the effect is minor, and even the transition states appear well described, which suggests good transferability for these basis sets. The poorest result is obtained for CO, with an atypical and short C–O bond, for which the molecularly optimized and split valence basis give similar

TABLE II. Bond lengths (angstrom) for (A) H_2O, (B) HCl, (C) CH_3CN (C–N bond), (D) CO as obtained with the different basis sets. All-electron calculations of the corresponding quantities using the aug-cc-pVQZ basis yield 0.969, 1.290, 1.162, and 1.135 Å.

	SZV	DZVP	TZVP	TZV2P	QZV2P	QZV3P	
A	1.114	0.978	0.972	0.970	0.971	0.970	
B	1.473	1.299	1.296	1.288	1.288	1.287	
C	1.312	1.180	1.167	1.162	1.162	1.162	
D	1.342	1.150	1.143	1.138	1.138	1.137	
		aug-DZVP	aug-TZVP	aug-TZV2P	aug-QZV2P	aug-QZV3P	
A		0.978	0.973	0.971	0.971	0.970	
B		1.299	1.296	1.288	1.288	1.287	
C		1.176	1.167	1.162	1.162	1.162	
D		1.148	1.143	1.138	1.138	1.137	
	m-SZV	m-DZVP	m-TZVP	m-TZV2P			m-TZV2PX
A	1.004	0.972	0.971	0.970			0.970
B	1.325	1.290	1.289	1.288			1.286
C	1.223	1.167	1.163	1.162			1.162
D	1.291	1.136	1.140	1.137			1.136

results. At this point, we would like to mention the link with response basis sets,[32] since the concept employed in the generation of these basis sets explains to some extent the performance and transferability observed here. In the case of response basis sets, the minimal basis is the atomic basis, and its size is increased by considering perturbations of the atomic charge density. The response of the spherical atom (to first and higher orders) to the addition and removal of electrons allows for increasing the size of valence space, while the response to nonhomogenous fields allows for adding basis functions with increased angular momentum. The molecularly optimized basis sets are similar, except that we have chosen to avoid isolated atoms as reference systems and instead used atoms in their molecular environment. Furthermore, the perturbations experienced by these atoms are not infinitesimal but, due to the molecular environment, of an appropriate finite magnitude. Finally, it is interesting to compare the results in Table I to all-electron calculations, in particular, to the polarization consistent basis sets of Jensen.[16,17] This carefully constructed series of basis sets (pc-*N*, with *N* from 0 to 4) has been shown to yield excellent total energies to within microhartrees of the exact results. The m-DZVP and m-TZV2PX bases have the same composition (neglecting core states) as the contracted versions of pc-1 and pc-2, respectively. The errors in the total energy obtained with the (uncontracted) pc-1 and pc-2 basis sets are approximately 35 and 3 millihartree/at. and thus higher than the error we observe. Nevertheless, it should be pointed out that the errors in the atomization energies of pc-1 and pc-2 are similar to the errors in total energy we observe with m-DZVP and m-TZV2PX. The atomization energy might be a more meaningful quantity for the comparison of all-electron to pseudopotential calculations. Finally, plane wave calculations for a single water molecule, as described by GTH pseudopotentials, at wave function cutoffs of 100, 200, and 250 Ry have errors in the total energies of 0.0902, 0.0037, 0.000 85 hartree, respectively. This is similar to the errors obtained with the m-SZV, m-DZVP, and m-TZV2PX basis sets. Particularly in the case of plane wave basis sets, experience has shown that the performance for total energies does imply little for other properties such as, for example, intermolecular interactions.

B. Bond lengths and molecular dipoles

Bond lengths and molecular dipoles are reported for a number of small molecules in Tables II and III, respectively. The results illustrate that the excellent performance of the m-DZVP basis holds not only for total energies, but also for geometries and for molecular dipoles. Bond lengths are within 0.005 A (the C–N triple bond in CH_3CN) of the converged all-electron results, while dipoles are within 0.1 D (CO). The latter property is sensitive to a good description of the density far from the center of the molecule. We observe that m-DZVP performance is in most cases similar to the aug-TZV2P performance and that m-TZV2PX does not improve these results significantly, with the exception of CO which was the worst case for the total energies. We note that the good agreement between pseudopotential and all-electron calculations is a further indication of the quality of the GTH pseudopotentials, but refer to Ref. 26 for a more detailed discussion and further comparison with all-electron calculations.

C. A comparison between molecular dipoles in gas and condensed phases

At this point, we would like to hint at the fact that the requirements for a basis in the gas and condensed phases might differ, and especially that for bulk molecular liquid diffuse functions might not be needed for a correct description of the electronic structure. This is in agreement with our earlier observations of liquid water[5] and liquid acetonitrile[36] at constant volume, where we find surprisingly little effect of the basis on structural and dynamical properties of these liquids. This, however, will not hold if these liquids are allowed to change density, for example, if they are in contact with a

TABLE III. Dipole (debye) for (A) H_2O, (B) HCl, (C) CH_3CN, and (D) CO as obtained with the different basis sets at the consistently optimized geometry (see Table II). All-electron calculations of the corresponding quantities using the aug-cc-pVQZ basis yield 1.803, 1.094, 4.000, and 0.200.

	SZV	DZVP	TZVP	TZV2P	QZV2P	QZV3P	
A	2.205	2.064	2.145	1.956	1.959	1.901	
B	1.498	1.375	1.372	1.180	1.171	1.121	
C	3.853	4.046	4.030	4.028	4.026	4.024	
D	0.346	0.191	0.192	0.209	0.195	0.194	
		aug-DZVP	aug-TZVP	aug-TZV2P	aug-QZV2P	aug-QZV3P	
A		1.925	1.911	1.843	1.840	1.823	
B		1.150	1.137	1.082	1.077	1.069	
C		4.035	4.030	4.018	4.015	4.014	
D		0.172	0.169	0.188	0.194	0.193	
	m-SZV	m-DZVP	m-TZVP	m-TZV2P			m-TZV2PX
A	2.479	1.847	1.857	1.860			1.869
B	2.091	1.115	1.135	1.109			1.133
C	3.911	4.019	4.046	4.022			4.011
D	0.926	0.273	0.199	0.187			0.184

vapor state. Recent examples of such *ab initio* calculations are the water-vapor interface[37,38] or calculations of the water vapor–liquid equilibrium,[6,7] where a good description of the diffuse wave function tails remains necessary. In particular for the vapor-liquid equilibrium, a clear effect of the basis could be established.[39] Here, we wish to investigate an easy-to-quantify aspect of the performance of different basis sets in the condensed phase and compute the molecular dipoles of the individual water molecules in the liquid based on the Berry phase formalism. For the 64 water molecules of a single liquid configuration, the root mean square error (RMSE) and the maximum error with respect to the results as obtained for the aug-QZV3P basis have been computed. In Table IV these are compared to the errors for a gas phase water molecule. The difference between the RMSE and the error in the gas phase is striking, as the error is reduced by almost a factor of 10, yielding dipoles within 0.01 D of the reference result, even for the relatively small split valence basis sets without diffuse functions. The maximum error is also reduced, albeit slightly. This could be related to the fact that this error occurs for a water molecule with a rather distorted configuration (H–O–H angle of 111°), i.e., in this case the error in the dipole is not only due to a poor description of the wave function tail.

D. Condition number and basis set superposition error

We report a number of tests that illustrate the central claim of the paper, i.e., that the molecularly optimized basis set is accurate for molecular interactions in the gas phase, yet maintains a good condition number of the overlap matrix, with the associated stability for the self-consistent field (SCF) and molecular dynamics procedures, in the condensed phase. The condition number in the condensed phase is reported for a number of systems, all molecular liquids simulated using periodic boundary conditions as described in more detail in previous work,[5,36,40] in Table V. The BSSE as estimated with the counterpoise correction[41] is reported in Table VI for a number of hydrogen bonded dimers. Hydro-

TABLE IV. Error in the molecular dipole (debye) relative to the aug-QZV3P results for water in gas and in a liquid sample containing 64 molecules. (A) the error for a gas phase molecule at equilibrium geometry. (B) the maximum error for the molecules of the liquid sample. (C) The RMSE for the molecules of the liquid sample.

	SZV	DZVP	TZVP	TZV2P	QZV2P	QZV3P	
A	0.383	0.242	0.322	0.134	0.136	0.078	
B	0.419	0.180	0.167	0.050	0.026	0.008	
C	0.046	0.017	0.012	0.003	0.001	0.000	
		aug-DZVP	aug-TZVP	aug-TZV2P	aug-QZV2P	aug-QZV3P	
A		0.102	0.088	0.020	0.017	0.000	
B		0.029	0.034	0.004	0.004	0.000	
C		0.001	0.003	0.000	0.000	0.000	
	m-SZV	m-DZVP	m-TZVP	m-TZV2P			m-TZV2PX
A	0.656	0.024	0.034	0.038			0.046
B	0.301	0.039	0.028	0.018			0.019
C	0.019	0.001	0.000	0.000			0.001

TABLE V. The condition number of the overlap matrix (ratio of largest to smallest eigenvalue) for different molecular liquids, represented using periodic boundary conditions. (A) 64 water molecules (Ref. 5), (B) 56 methanol molecules and a solvated benzoquinone (Ref. 40), (C) 45 acetonitrile molecules (Ref. 36). Details on how these configurations have been obtained can be found in the references provided.

	SZV	DZVP	TZVP	TZV2P	QZV2P	QZV3P	
A	1.0×10^{01}	9.3×10^{02}	1.2×10^{04}	2.9×10^{04}	2.9×10^{05}	4.4×10^{05}	
B	2.0×10^{01}	1.3×10^{05}	4.2×10^{06}	7.7×10^{06}	2.5×10^{08}	4.6×10^{08}	
C	2.2×10^{01}	1.4×10^{04}	2.8×10^{05}	4.9×10^{05}	1.8×10^{07}	2.9×10^{07}	
		aug-DZVP	aug-TZVP	aug-TZV2P	aug-QZV2P	aug-QZV3P	
A		1.3×10^{10}	2.1×10^{12}	3.5×10^{12}	3.3×10^{14}	1.3×10^{15}	
B		1.0×10^{11}	1.9×10^{13}	3.3×10^{13}	9.9×10^{13}	8.7×10^{13}	
C		7.7×10^{09}	1.4×10^{12}	2.3×10^{12}	3.8×10^{14}	1.7×10^{14}	
	m-SZV	m-DZVP	m-TZVP	m-TZV2P			m-TZV2PX
A	6.8×10^{00}	1.6×10^{03}	4.0×10^{03}	1.5×10^{04}			1.9×10^{04}
B	1.1×10^{01}	2.2×10^{03}	9.7×10^{03}	2.9×10^{04}			4.6×10^{04}
C	1.3×10^{01}	1.7×10^{03}	5.9×10^{03}	1.5×10^{04}			2.3×10^{04}

gen bonding is a relatively weak interaction that can be described rather well with DFT. However, the property is known to be sensitive to the BSSE. We note that the BSSE is not the only contribution to the error in the interaction energy between molecules, as the latter also depends on, e.g., the quality of the geometry. The errors in the interaction energy are typically even larger for basis sets that yield poor geometries in Table II, but we wish to separate to some extent these issues by considering the BSSE only. The condition numbers of the molecularly optimized basis sets clearly indicate that we have succeeded in our attempt to generate a basis set that is free of near degeneracies. Even the large m-TZV2PX basis sets have condition numbers in the condensed phase that are similar to the smaller split valence basis sets and that can be up to four orders of magnitude smaller than basis sets of similar size. The difference is significantly more pronounced if we compare with the augmented basis sets. Both basis sets contain primitives with similar diffuse exponents, but the molecularly optimized basis sets have condition numbers that are up to ten orders of magnitude better. At the same time, the molecularly optimized basis sets also provide a route toward removing the BSSE, the m-TZV2PX basis has errors below 0.16 kcal/mol for the systems studied, and can thus compete with the QZV2P and aug-TZV2P basis sets. The small m-DZVP basis performs better than the larger TZV2P basis set and has a BSSE of only 0.2 kcal/mol for the water dimer, a system relevant to most *ab initio* molecular dynamics simulations. The computational cost of the optimization did not allow for varying the weight γ of the condition number of the overlap matrix in the optimization [see Eq. (1)], but we speculate that reducing this weight would improve the performance of the m-TZV2PX for the BSSE, at the cost of increasing the condition number for this basis. Finally, we note that there appears to be a significant difference between typical all-electron and pseudopotential calculations with respect to the importance of the BSSE and the influence of diffuse functions. Indeed, the water dimer, for example, has BSSEs of 1.97, 3.55, and 1.65 kcal/mol for 6-31G^{**}, cc-pVDZ, and cc-pVTZ, respectively, which is significantly larger than the

TABLE VI. Estimated BSSEs (kcal/mol) for (A) H_2O-H_2O, (B) NH_3-NH_3, (C) HF–HF, and (D) NH_3-H_2O dimers.

	SZV	DZVP	TZVP	TZV2P	QZV2P	QZV3P	
A	0.54	0.64	0.60	0.32	0.27	0.16	
B	0.23	0.63	0.67	0.35	0.17	0.09	
C	0.77	0.76	0.24	0.20	0.10	0.06	
D	0.45	0.90	1.00	0.49	0.26	0.16	
		aug-DZVP	aug-TZVP	aug-TZV2P	aug-QZV2P	aug-QZV3P	
A		0.64	0.29	0.11	0.09	0.04	
B		0.60	0.25	0.07	0.07	0.03	
C		0.29	0.11	0.07	0.05	0.03	
D		0.67	0.37	0.12	0.09	0.04	
	m-SZV	m-DZVP	m-TZVP	m-TZV2P			m-TZV2PX
A	0.31	0.23	0.18	0.11			0.11
B	0.67	0.11	0.17	0.11			0.13
C	0.21	0.41	0.34	0.19			0.15
D	1.30	0.20	0.23	0.16			0.16

TABLE VII. Shown is, for a fully solvated protein (Ref. 3). (A) the fraction of nonzero matrix elements of the overlap matrix, (B) the fraction of nonzero matrix elements of S^{-1}, and (C) the number basis functions. The threshold to consider a matrix element nonzero is 10^{-3}, NA has been used to indicate that the overlap matrix, represented using double precision numbers, is not positive definite. (See text for details.)

	SZV	DZVP	TZVP	TZV2P	QZV2P	QZV3P	
A	0.013	0.008	0.010	0.008	0.009	0.007	
B	0.018	0.144	0.429	0.399	0.644	0.646	
C	6107	22 822	28 910	39 537	45 625	56 252	
		aug-DZVP	aug-TZVP	aug-TZV2P	aug-QZV2P	aug-QZV3P	
A		0.025	0.028	0.021	0.023	0.018	
B		0.868	0.914	0.891	NA	NA	
C		34 082	40 170	50 797	56 885	67 512	
	m-SZV	m-DZVP	m-TZVP	m-TZV2P			m-TZV2PX
A	0.014	0.037	0.044	0.077			0.083
B	0.013	0.150	0.348	0.406			0.350
C	6107	22 822	28 910	39 537			55 794

results we have obtained, 0.64 kcal/mol in the worst case (DZVP) for a basis without diffuse primitives. Augmenting these all-electron basis sets has a pronounced effect, reducing errors to 0.86, 0.23, and 0.05 kcal/mol for $6\text{-}31+G^{**}$, aug-cc-pVDZ, and aug-cc-pVTZ, respectively.

E. Sparsity of the inverse overlap matrix in the condensed phase

The purpose of Table VII is to illustrate the effect of the basis set on the sparsity of the overlap matrix (S) and more importantly its inverse (S^{-1}). The system used for testing is the iron-sulfur protein rubredoxin, fully solvated and described using periodic boundary condition. The system contains approximately 2800 atoms contained within a unit cell of $31 \times 28 \times 31$ Å^3. Its redox properties have previously been computed using CP2K based on a DFT description of the full system.[3] All elements of the overlap matrix have been computed to machine accuracy, while the inverse of the overlap matrix has been computed explicitly based on the Cholesky decomposition of S. Failure to compute the Cholesky decomposition indicates that the inverse condition number is roughly equal to machine accuracy. Due to computational constraints, a full optimization of the wave function has not been attempted. The sparsity of the matrices has been computed as the fraction of matrix elements larger than a given threshold. The threshold has been kept very high (10^{-3}) in order to be able to observe at least some sparsity in S^{-1} as obtained for the larger basis sets. For a larger system, a lower threshold could have been used to extract similar information. A first observation is that the sparsity of S is significantly reduced for the augmented and the molecularly optimized basis sets. This is no surprise, as both basis sets add diffuse primitives, which introduces nonzero matrix elements for atoms that are far apart. A difference between these two basis sets is that molecularly optimized basis set adds diffuse primitives to all basis functions, whereas in the augmented case, only a part of the basis is actually diffuse. This results in a larger number of nonzero matrix elements for the molecularly optimized set, even though the number of nonzero atomic blocks should be similar. The importance of the condition number (see Table V) is reflected in the sparsity of S^{-1}. Indeed, for extremely small and well conditioned basis sets (SZV), we observe a similar sparsity for S and S^{-1}. For all other basis sets, the filling of S^{-1} is increased by roughly a factor of 10 or more. This underlines the importance of testing linear scaling methods with basis sets that are nonminimal, as the computational cost of certain terms might be significantly different between SZV and DZVP or larger basis sets. For the augmented basis set, there is basically no sparsity in S^{-1}, while for the molecularly optimized basis sets the sparsity is similar, or even better than the sparsity obtained with the split valence basis sets. The results are fully in line with the condition numbers shown in Table V.

F. Computational cost

In order to discuss the computational cost of the basis sets presented here, it is important to realize that in the Gaussian and plane wave approaches as implemented in CP2K/QUICKSTEP,[2] the computational cost is not a simple function of the basis set size or composition. In particular, there are three main regimes or system types that are relevant, as each of them exposes a different computational bottleneck: (a) small gas phase or mixed quantum/classical (QM/MM) systems that contain a few dozen atoms: dominated by operations on the plane wave grids and relatively independent of the Gaussian basis set used. (b) Small to medium sized (100 s of atoms) condensed phase systems: dominated by the collocation of the density/integration of the matrix elements on the plane wave grids, sensitive to the number of matrix elements, and the composition in primitives of the corresponding basis function. (c) Large systems containing several thousands of atoms: dominated by the linear algebra, i.e., sensitive to the number of basis functions, but not their composition. Furthermore, since the computational bottlenecks have a different parallel efficiency, the boundaries between these classes are not rigid, but might depend on the number of CPUs or other hardware details. In Table VIII, the ruthenium tris(bipyridine) compound represents class a, the liquid water (64 water molecules, periodic boundary conditions) represents class b, and class c is repre-

TABLE VIII. Timings in seconds per SCF iteration for (A) ruthenium tris(bipyridine) in a QM/MM setup (classical acetonitrile solvent, quantum solute). (B) A liquid water sample (64 water molecules with periodic boundary conditions). (C) A solvated configuration of the iron-sulfur protein rubredoxin (high spin state). Timings for A and B are on a single core of an Intel Core2 at 2.4 GHz, C on 256 CPUs of a CRAY XT3. NA has been used to indicate calculations that failed due to a singular overlap matrix.

	SZV	DZVP	TZVP	TZV2P	QZV2P	QZV3P	
A	26	28	30	31	34	38	
B	8	14	20	36	48	77	
C	26	54	71	100	124	167	
		aug-DZVP	aug-TZVP	aug-TZV2P	aug-QZV2P	aug-QZV3P	
A		35	38	39	43	49	
B		58	92	124	151	200	
C		NA	NA	NA	NA	NA	
	m-SZV	m-DZVP	m-TZVP	m-TZV2P			m-TZV2PX
A	32	39	40	40			53
B	22	111	116	140			386
C	33	105	128	192			369

sented by the solvated iron sulfur protein rubredoxin. Comparing the time needed per SCF step for a small (SZV) and a large basis (m-TZV2P), the clear difference between regimes a and b can be seen, as the time for ruthenium tris(bipyridine) increases by less than a factor of 2, while for liquid water an almost 20-fold increase of computer time is observed. The most important reason for this is the diffuse nature of the basis, which for condensed phase system greatly increases the number of atoms that interact with other atoms for a given tolerance in the screening. Indeed, the time needed for the augmented and molecularly optimized basis sets is similar, even though the latter requires slightly more time, as all basis functions contain diffuse primitives and are highly contracted. Two more detailed results that can be extracted from the liquid water results in Table VIII are the small difference in timings going from the m-DZVP to the m-TZV2P basis and the relatively large increase going from m-TZV2P to m-TZV2PX. The former illustrates that the code is able to exploit the family basis aspect of the basis, indeed, no new primitives are added going from m-DZVP to the m-TZV2P, and so no significant increase in computational cost is observed for systems that are dominated by the integral evaluation. The increase of time going from m-TZV2P to m-TZV2PX is related to the increased angular momentum of all basis functions. However, this increase in computational cost is relatively modest, as the inner loop of the grid based integration and collocation scales linearly in the l quantum number.[2] The protein benchmark (≈2800 atoms) tries to explore the large systems regime, which is dominated by the linear algebra part. It can indeed be observed that the timings for the traditional and molecularly optimized basis sets become similar, for example, 100 and 192 s per iteration for the TZV2P and m-TZV2P basis sets, respectively. The difference in the timings can be attributed to the fact that the cubically scaling terms, which are independent of the composition of the basis, are not yet fully dominant in the m-TZV2P case, while linear, quadratic, and cubic parts account, respectively, for approximately 20%, 20%, and 60% of the time in the TZV2P case. The surprisingly small cost of the cubic part can, to a large part, be attributed to the efficiency of the orbital transformation scheme.[1] Indeed, all but one cubic term scale, at fixed system size, linearly in the size of the basis, as is illustrated by the near linear increase of cost going from a DZVP to a QZV2P basis. In actual calculations, only the preconditioning of wave function minimization is performed using a full matrix representation of the inverse of the overlap matrix (or related preconditioners), and hence scales quadratically in the size of the basis. For the TZV2P basis, this matrix multiply step accounts for approximately 50% of the time spent in linear algebra. Finally, with 369 s per iteration (117 s on 1024 CPUs), the results for the m-TZV2PX basis show that it is possible to compute the electronic structure of systems containing thousands of atoms with a basis set that typically has an error in the total energy below a millihartree/at., excellent electrostatic properties, and a small basis set superposition error.

IV. CONCLUSIONS

We have presented an approach to derive Gaussian basis sets with favorable properties for density functional calculations on molecular systems. The same basis sets are suitable for calculations in the gas and in the condensed phase. By optimizing all Gaussian exponents and contraction coefficients based on molecular calculations, we obtain optimally adapted radial functions for all angular momenta. This procedure results in basis sets with excellent performance for total energies and geometries, but especially the good performance of the m-DZVP basis is noteworthy. At the same time, we find that the basis remains transferable. Allowing for a sufficient number of primitives and optimizing the exponents not only within a SZV basis, diffuse primitives are added to the basis in a natural way. This leads to a basis set with good performance for properties that rely on a proper description of the wave function tails, such as molecular dipoles. Furthermore, we have shown that, for a number of hydrogen bonded dimers, the BSSE can be reduced to approximately the level of a traditional augmented basis. The main difference with respect to these basis sets, and a prime objective of this work, is the well conditioned nature of the molecularly optimized basis sets. Indeed, near degeneracies, as manifested by small eigenvalues of the overlap matrix, are absent even in the condensed phase. This makes them particularly well suited for geometry optimization, molecular dynamics simulations, and large or condensed phase systems. The computational cost of the basis is highly dependent on the type of system studied and most strongly influenced by the diffuse nature of the basis. With CP2K/QUICKSTEP, the cost is moderate for small gas phase (or QM/MM) systems and systems containing thousands of atoms, but significant for medium sized condensed phase systems. Finally, we demonstrated the feasibility of computing the electronic structure of a fully solvated iron-sulfur protein with an estimated error in the total energy below a millihartree/at., and nearly free of BSSE. Such a calculation runs, with an estimated parallel efficiency of approximately 80%, in less than 2 min minutes per SCF step on 1024 CPUs.

[1] J. VandeVondele and J. Hutter, J. Chem. Phys. **118**, 4365 (2003).

[2] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter, Comput. Phys. Commun. **167**, 103 (2005).

[3] M. Sulpizi, S. Raugei, J. VandeVondele, P. Carloni, and M. Sprik, J. Phys. Chem. B **111**, 3969 (2007).

[4] I.-F. W. Kuo, C. J. Mundy, M. J. McGrath *et al.*, J. Phys. Chem. B **108**, 12990 (2004).

[5] J. VandeVondele, F. Mohamed, M. Krack, J. Hutter, M. Sprik, and M. Parrinello, J. Chem. Phys. **122**, 014515 (2005).

[6] M. McGrath, J. Siepmann, I.-F. W. Kuo, C. Mundy, J. VandeVondele, J. Hutter, F. Mohamed, and M. Krack, ChemPhysChem **6**, 1894 (2005).

[7] M. J. McGrath, J. Siepmann, I.-F. W. Kuo, C. Mundy, J. VandeVondele, J. Hutter, F. Mohamed, and M. Krack, J. Phys. Chem. A **110**, 640 (2006).

[8] S. F. Boys, Proc. R. Soc. London, Ser. A **200**, 542 (1950).

[9] J. C. Slater, Phys. Rev. **36**, 57 (1930).

[10] J. Ihm, A. Zunger, and M. L. Cohen, J. Phys. C **12**, 4409 (1979).

[11] D. Marx and J. Hutter, in *Modern Methods and Algorithms of Quantum Chemistry*, NIC Series Vol. 1, edited by J. Grotendorst (FZ Jülich, Germany, 2000), pp. 329–477; see also http://www.fz-juelich.de/nic-series/Volume1/

[12] K. Cho, T. A. Arias, J. D. Joannopoulos, and P. K. Lam, Phys. Rev. Lett. **71**, 1808 (1993).

[13] B. Delley, J. Chem. Phys. **92**, 508 (1990).

[14] E. R. Davidson and D. Feller, Chem. Rev. (Washington, D.C.) **86**, 681 (1986).

[15] S. Huzinaga, J. Andzelm, M. Klobukowsi, E. Radzio-Andzelm, Y. Sakai, and H. Tatewaki, *Gaussian Basis Sets for Molecular Calculations* (Elsevier, Amsterdam, 1984).

[16] F. Jensen, J. Chem. Phys. **115**, 9113 (2001).

[17] F. Jensen, J. Chem. Phys. **116**, 7372 (2002).

[18] S. Goedecker and G. E. Scuseria, Comput. Sci. Eng. **5**, 14 (2003).

[19] G. Lippert, J. Hutter, and M. Parrinello, Mol. Phys. **92**, 477 (1997).

[20] G. Lippert, J. Hutter, and M. Parrinello, Theor. Chem. Acc. **103**, 124 (1999).

[21] S. Goedecker, Rev. Mod. Phys. **71**, 1085 (1999).

[22] R. C. Raffenetti, J. Chem. Phys. **58**, 4452 (1973).

[23] M. S. Lee and M. Head-Gordon, J. Chem. Phys. **107**, 9085 (1997).

[24] M. S. Lee and M. Head-Gordon, Computers & Chemistry **24**, 295 (2000).

[25] G. Berghold, M. Parrinello, and J. Hutter, J. Chem. Phys. **116**, 1800 (2002).

[26] S. Goedecker, M. Teter, and J. Hutter, Phys. Rev. B **54**, 1703 (1996).

[27] C. Hartwigsen, S. Goedecker, and J. Hutter, Phys. Rev. B **58**, 3641 (1998).

[28] M. Krack, Theor. Chem. Acc. **114**, 145 (2005).

[29] The CP2K developers group, http://cp2k.berlios.de/ (2007).

[30] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[31] See EPAPS Document No. E-JCPSA6-127-308733 for the exponents and coefficients of the molecularly optimized basis sets. This document can be reached through a direct link in the online article's HTML reference section or via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html).

[32] G. Lippert, J. Hutter, P. Ballone, and M. Parrinello, J. Chem. Phys. **100**, 6231 (1996).

[33] M. J. D. Powell, Math. Program. **92**, 555 (2002).

[34] J. Hutter *et al.*, CPMD (Car-Parrinello Molecular Dynamics), an *Ab Initio* Electronic Structure and Molecular Dynamics Program, IBM Zurich Research Laboratory (1990-2007) and Max-Planck-Institut für Festkörperforschung Stuttgart (1997–2001), http://www.cpmd.org/

[35] M. J. Frisch, G. W. Trucks, H. B. Schlegel *et al.*, GAUSSIAN 03, Revision C.02, Gaussian, Inc., Wallingford CT, 2004.

[36] J. VandeVondele, R. Lynden-Bell, E. J. Meijer, and M. Sprik, J. Phys. Chem. B **110**, 3614 (2006).

[37] I.-F. W. Kuo and C. J. Mundy, Science **303**, 658 (2004).

[38] C. J. Mundy and I.-F. W. Kuo, Chem. Rev. (Washington, D.C.) **106**, 1282 (2006).

[39] M. J. McGrath, J. I. Siepmann, I.-F. W. Kuo, and C. J. Mundy, Mol. Phys. **104**, 3619 (2006).

[40] J. VandeVondele, M. Sulpizi, and M. Sprik, Angew. Chem., Int. Ed. **45**, 1936 (2006).

[41] S. F. Boys and F. Bernardi, Mol. Phys. **19**, 553 (1970); [Mol. Phys. **100**, 65 (2002) (reprinted)].