JCTC Journal of Chemical Theory and Computation

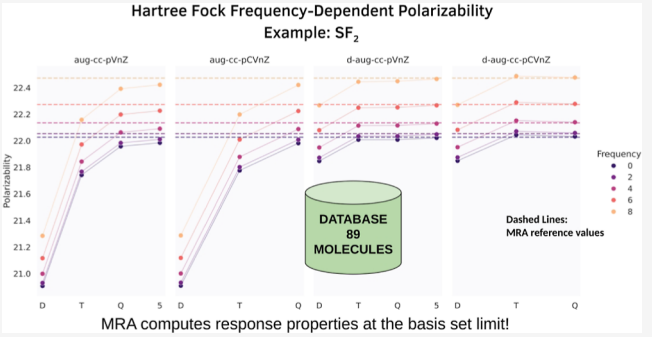
pubs.acs.org/JCTC Article

Benchmarking Correlation-Consistent Basis Sets for Frequency-Dependent Polarizabilities with Multiresolution Analysis

Adrian Hurtado,* Hideo Sekino,* and Robert J. Harrison*

Cite This: J. Chem. Theory Comput. 2024, 20, 5145−5156 Read Online

ABSTRACT: This paper presents the first converged frequency-dependent HF polarizability results for general molecules, on a set of 89 closed-shell atoms and molecules. The solver employs multiresolution analysis (MRA) in a multiwavelet basis to compute both ground and response states to a guaranteed precision, which are validated against independent numerical grid calculations on atoms and linear molecules. The MRA ground-state energies and response properties are used to evaluate results in correlation-consistent basis sets up to 5Z augmented with either single or double diffuse functions and core-polarization functions. Systematic trends are revealed through consideration of chemical composition as well as the use of machine learning to cluster convergence trends, the latter suggesting the possibility of learning and correcting basis-set error.



1. INTRODUCTION

We present benchmark Hartree−Fock (HF) frequency-dependent polarizabilities for a set of 89 molecules using a new multiresolution analysis (MRA) solver.[1−5]

Many quantum chemistry packages, such as Dalton[6] and NWChem[7] used herein, have solvers employing the standard linear combination of atomic orbitals numerical technique for computing response properties. Although Gaussian bases have several advantages,[8,9] convergence of results can be challenging since it is difficult to systematically improve the underlying representation for general molecules without introducing significant computational expense and, potentially, near numerical linear dependence. The provided reference-quality results make it possible to fully quantify the accuracy expected from a given basis, elemental composition, and bonding pattern of the molecule.

The frequency-dependent polarizability is a second-order response property with the quality of results dependent on the calculation of both ground and response states. MRA guarantees accuracy by adaptively refining the basis to meet specified numerical truncation thresholds in the solution,[2,10,11] with this being performed independently in both calculations. In contrast, despite their possibly different physical characteristics, the same Gaussian basis set is typically employed to calculate both the ground and response states, leading to known difficulties with basis-set imbalance[12] in which one or the other is more accurately described. Similarly, incompleteness in one region of the molecule (e.g., atomic core or inner valence) can "borrow" basis functions intended to describe another region (e.g., the tail of the wave function). These behaviors are discussed further in the Results section below.

Historically, fully converged numerical results were only available for atoms and linear molecules for which the use of symmetry reduces the numerical problem to 1 or 2 dimensions. To independently establish the accuracy of our MRA results, we compare with highly accurate results from numerical finite-difference grid calculations on atoms and linear molecules.[13] MRA within a multiwavelet basis[14,15] was the first approach capable of routinely computing converged results for general polyatomic molecules,[2,3,11] and there are now several independent codes[11,16−18] available for general use. MRA allows for a numerical solution of ground state and response properties[19−21] with guaranteed accuracy and speed. Recent results[22,23] demonstrate that for large molecules the MRA approach can be faster than Gaussian bases beyond low precision. Several previous studies have employed MRA to provide reference-quality quantum chemistry data[24−26] including evaluation of the aug-pc4 basis set for computation of DFT static polarizabilities on a set of 124 open and closed-shell systems.[24−26]

This paper presents the first converged frequency-dependent HF polarizability results for general molecules, employing a set of 89 closed-shell atoms and molecules. We focus upon basis-set errors in the computation of frequency-dependent HF polar-

Received: March 27, 2024
Revised: May 20, 2024
Accepted: May 21, 2024
Published: June 6, 2024

Table 1. Molecules and Atoms in the Test Set Grouped by Subsets[a]

	1	2	3	4	5	6	7	8	9
1	HOOH	C_6H_6	CH_3OH	NH_3O	O_3	C_2H_2	CO	He	NH_3
2	CH_2BH	CH_2NH	HBO	CH_3BH_2	BeH_2	Be	LiCN	$HCONH_2$	Ne
3	BH_3	C_2H_4	Li_2	H_2	CH_4	N_2H_2	NH_2OH	$LiBH_4$	HNC
4	LiH	N_2H_4	CO_2	HCN	H_2O	N_2	HCOOH	CH_3NH_2	HCHO
5	HNO	F_2	HOF	BH_2F	HCCF	FCN	HF	BHF_2	BF
6	NH_2F	OF_2	FNO	CH_3F	ClF	NOCl	SiH_4	HCP	SiH_3Cl
7	CS	NaCN	NH_2Cl	NaH	NaLi	HNS	SF_2	Na_2	SiO
8	Mg	BH_2Cl	SH_2	P_2	HOCl	NaCl	HCCCl	SiH_3F	HCl
9	P_2H_4	Mg_2	CSO	PH_3O	S_2H_2	OCl_2	HBS	SO_2	Cl_2
10	Ar	PH_3	SCl_2	CH_3Cl	LiCl	HCHS	ClCN	CH_3SH	–

[a]First-row elements in blue except for those with fluorine that are colored in orange; those containing second-row elements in green.

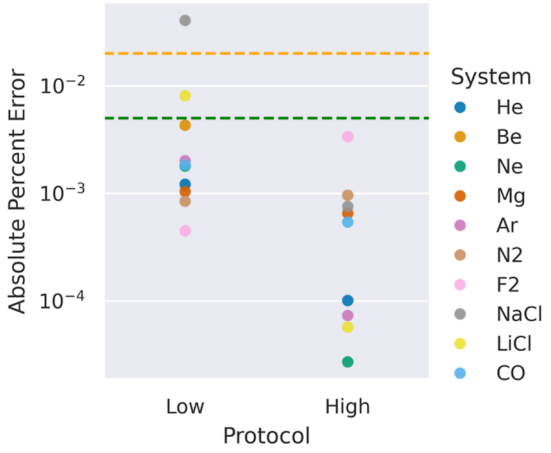


Figure 1. Comparison of MRA α_{zz} static polarizability to finite-difference reference data.[13] The calculations are performed at both low and high accuracy protocols with a convergence criterion for the residual of each response component equal to 5×10^{-5} and 5×10^{-7} relative error. The orange and green lines signify the MRA error bounds for low and high accuracy which are used to benchmark basis set calculations.

izability since the HF model is semiquantitatively accurate, is physically correct for large molecules since it does not suffer from the self-interaction error present in most DFT functionals, and eliminates numerical issues associated with higher-derivatives of density functionals and the presence of the kinetic energy operator in meta-GGA.[27]

The MRA ground-state energies and response properties are used to evaluate results in several correlation-consistent basis-set variants.[28−31] We explore basis-set incompleteness in response calculations by comparing Gaussian basis-set polarizabilities to MRA results. Specifically, we employ the singly and doubly augmented (aug/d-aug)-cc-pVnZ bases at the D, T, and Q valence levels (with limited results at the 5Z level) to explore the effect of diffuse function augmentation, and their core-polarized versions (aug/d-aug)-pCVnZ to investigate the impact of core polarization. Systematic trends are revealed through consideration of chemical composition as well as use of machine learning to cluster convergence trends, the latter suggesting the possibility of learning and correcting basis-set error.

2. METHODS

Total energies and frequency-dependent dipole polarizabilities were computed in the MRA basis using MADNESS[11] (git-**8c625b7**), and in Gaussian basis sets using the default parameter settings for Dalton.[6] Some results were reproduced with NWChem[7] for verification. We define the signed basis-set incompleteness errors (BSIE) in quantities Q, notably the total energy and polarizability, as

$$\Delta Q(\text{basis}) = Q(\text{basis}) - Q(\text{MRA}) \tag{1}$$

Unless otherwise stated, the term polarizability $[\alpha(\omega)]$ is used to refer to the spherically averaged (or isotropic) quantity computed at frequency ω

$$\alpha(\omega) = \frac{\alpha_{xx}(\omega) + \alpha_{yy}(\omega) + \alpha_{zz}(\omega)}{3} \tag{2}$$

We also examine the anisotropic polarizability $[\gamma(\omega)]$ following the definition from[32]

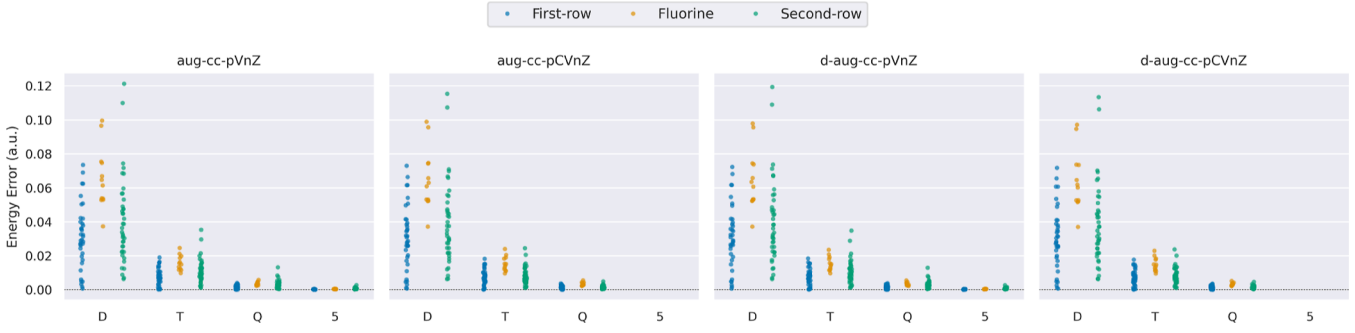


Figure 2. Basis set errors in total Hartree−Fock (HF) energy across different atomic systems. The scatter plots display the error in the total HF energy for the first-row (blue), fluorine (orange), and second-row (green) systems, across various basis sets. Each panel contains the results for a specific basis set variant. From left to right, these are aug-cc-pVnZ, aug-cc-pCVnZ, d-aug-cc-pVnZ, and d-aug-cc-pCVnZ. The valence basis set level increases from double-ζ (DZ) to quintuple-zeta (5Z).

https://doi.org/10.1021/acs.jctc.4c00394
J. Chem. Theory Comput. 2024, 20, 5145−5156

Table 2. Statistical Summary of Total Energy Errors in Basis Sets[a]

	count	mean	std	min	25%	50%	75%	max
aug-cc-pVDZ	89	3.99×10^{-2}	2.44×10^{-2}	6.43×10^{-4}	2.43×10^{-2}	3.62×10^{-2}	5.36×10^{-2}	1.21×10^{-1}
aug-cc-pCVDZ	89	3.89×10^{-2}	2.38×10^{-2}	6.43×10^{-4}	2.33×10^{-2}	3.51×10^{-2}	5.31×10^{-2}	1.15×10^{-1}
d-aug-cc-pVDZ	89	3.94×10^{-2}	2.40×10^{-2}	6.36×10^{-4}	2.40×10^{-2}	3.58×10^{-2}	5.32×10^{-2}	1.19×10^{-1}
d-aug-cc-pCVDZ	89	3.85×10^{-2}	2.35×10^{-2}	6.36×10^{-4}	2.31×10^{-2}	3.48×10^{-2}	5.26×10^{-2}	1.13×10^{-1}
aug-cc-pVTZ	89	1.02×10^{-2}	6.50×10^{-3}	1.47×10^{-4}	5.94×10^{-3}	9.37×10^{-3}	1.37×10^{-2}	3.53×10^{-2}
aug-cc-pCVTZ	89	8.73×10^{-3}	5.44×10^{-3}	1.06×10^{-4}	4.98×10^{-3}	8.15×10^{-3}	1.26×10^{-2}	2.44×10^{-2}
d-aug-cc-pVTZ	89	9.96×10^{-3}	6.33×10^{-3}	1.46×10^{-4}	5.86×10^{-3}	9.21×10^{-3}	1.35×10^{-2}	3.48×10^{-2}
d-aug-cc-pCVTZ	89	8.56×10^{-3}	5.28×10^{-3}	1.04×10^{-4}	4.91×10^{-3}	7.98×10^{-3}	1.22×10^{-2}	2.37×10^{-2}
aug-cc-pVQZ	89	2.56×10^{-3}	1.95×10^{-3}	5.28×10^{-5}	1.31×10^{-3}	2.29×10^{-3}	3.25×10^{-3}	1.32×10^{-2}
aug-cc-pCVQZ	89	1.67×10^{-3}	1.14×10^{-3}	5.28×10^{-5}	8.84×10^{-4}	1.39×10^{-3}	2.36×10^{-3}	5.42×10^{-3}
d-aug-cc-pVQZ	89	2.50×10^{-3}	1.90×10^{-3}	5.24×10^{-5}	1.29×10^{-3}	2.23×10^{-3}	3.14×10^{-3}	1.29×10^{-2}
d-aug-cc-pCVQZ	89	1.62×10^{-3}	1.08×10^{-3}	5.24×10^{-5}	8.77×10^{-4}	1.37×10^{-3}	2.29×10^{-3}	5.11×10^{-3}
aug-cc-pV5Z	75	4.63×10^{-4}	3.94×10^{-4}	1.75×10^{-5}	2.22×10^{-4}	3.83×10^{-4}	5.59×10^{-4}	2.68×10^{-3}
d-aug-cc-pV5Z	75	4.53×10^{-4}	3.89×10^{-4}	1.70×10^{-5}	2.16×10^{-4}	3.70×10^{-4}	5.42×10^{-4}	2.66×10^{-3}

[a]The data includes the count of data points and the mean, standard deviation, minimum, quartiles, and maximum error values.

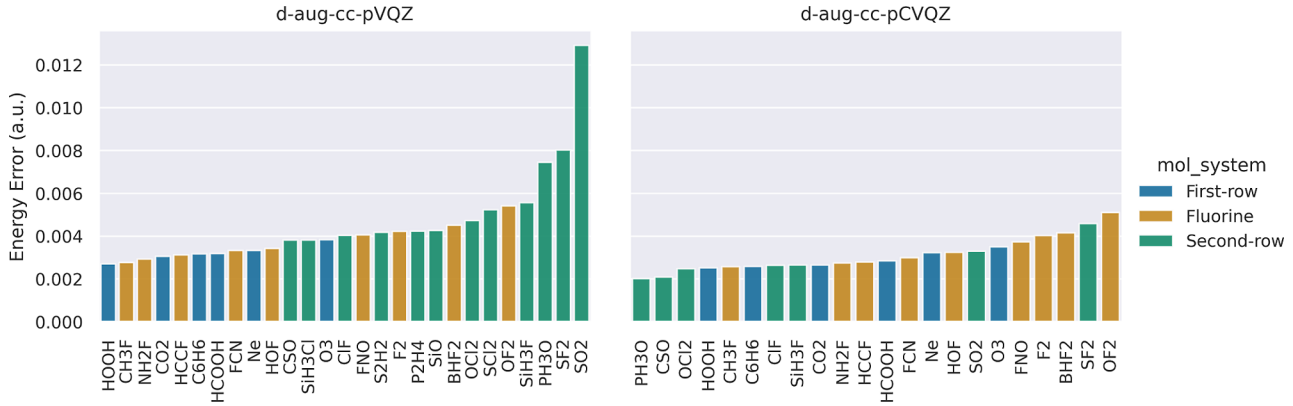


Figure 3. Energy outliers in aug-cc-pVQZ and d-aug-cc-pVQZ basis sets.

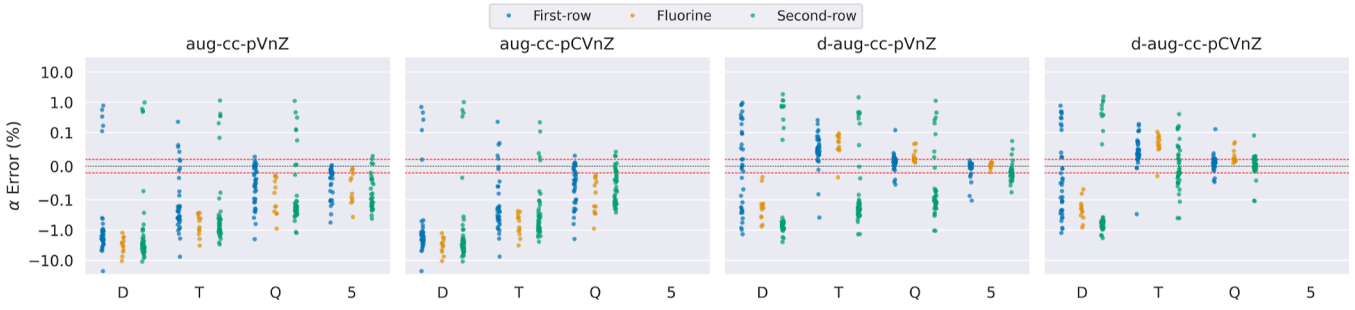


Figure 4. Basis set errors in the Hartree−Fock static isotropic polarizability. The scatter plots show the signed percentage error in the isotropic polarizability for the first-row (blue), fluorine (orange), and second-row (green) systems across various basis sets. In each plot, we use a symmetric-log scale with a linear threshold set to 0.1% to examine details at order of magnitude. We add red dashed lines at ±0.02% to each subplot, indicating MRA error bars.

$$\gamma = \frac{1}{\sqrt{2}}[(\alpha_{xx} - \alpha_{yy})^2 + (\alpha_{yy} - \alpha_{zz})^1 + (\alpha_{zz} - \alpha_{xx})^2 + 6(\alpha_{xy}^2 + \alpha_{yz}^2 + \alpha_{zx}^2)]^{1/2} \quad (3)$$

For each rotationally invariant measure of the polarizability, we examine the percentage error with respect to a reference value

$$\text{P. E.} = \frac{f - f_{\text{REF}}}{f_{\text{REF}}} \times 100 \quad (4)$$

where f refers to either the isotropic or anisotropic polarizability and the reference value depends on whether we are comparing basis set results or MRA values.

To examine the frequency dependence, for each molecule we estimated the first excitation energy (ω^*) using TDHF in an aug-cc-pVTZ basis. Subsequently, we computed the frequency-dependent polarizability at 9 frequencies spaced equally from 0 (i.e., static) to halfway to the first excitation energy. Thus, $\omega_i = i\omega^*/16$, with $i = 0, ..., 8$. In the next section, we establish the accuracy of the MRA results using reference value finite-difference results. Subsequently, we evaluate the performance of basis set results against our MRA reference results.

https://doi.org/10.1021/acs.jctc.4c00394
J. Chem. Theory Comput. 2024, 20, 5145−5156

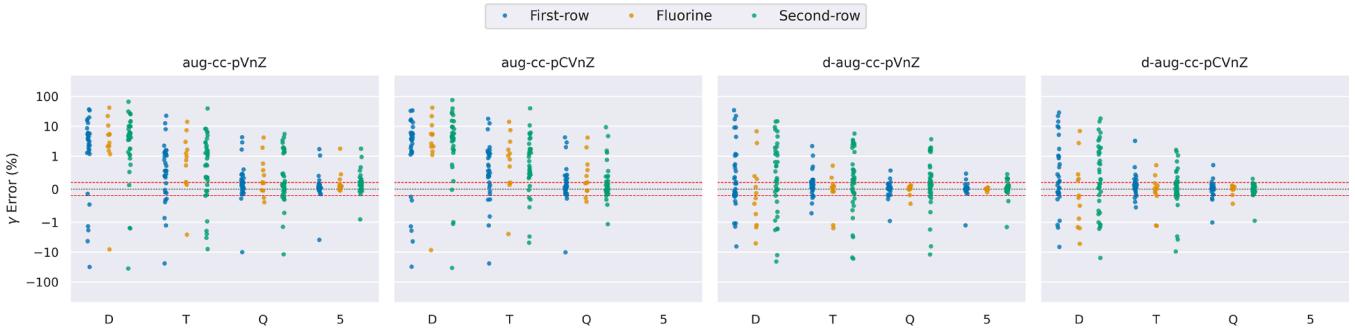


Figure 5. Basis set errors in the Hartree−Fock static anisotropic polarizability. As described in Figure 4 that examines the isotropic polarizability. In each plot, we use a symmetric-log scale with a linear threshold set to 1% to examine details at order of magnitude. We add red dashed lines at ±0.2% to each subplot, indicating MRA error bars.

Table 3. Statistical Summary of Isotropic Polarizability Errors in Basis Sets[a]

	count	mean	std	min	25%	50%	75%	max
aug-cc-pVDZ	89	−2.89	3.16	-2.29×10^{1}	−3.80	−2.29	−1.11	9.80×10^{-1}
aug-cc-pCVDZ	89	−2.87	3.11	-2.28×10^{1}	−3.75	−2.27	−1.25	1.00
d-aug-cc-pVDZ	89	-2.45×10^{-1}	6.65×10^{-1}	−2.50	-6.28×10^{-1}	-1.86×10^{-1}	6.40×10^{-2}	1.83
d-aug-cc-pCVDZ	89	-2.52×10^{-1}	5.62×10^{-1}	−1.85	-5.75×10^{-1}	-2.32×10^{-1}	-6.80×10^{-3}	1.54
aug-cc-pVTZ	89	-6.84×10^{-1}	1.03	−7.69	-9.43×10^{-1}	-4.34×10^{-1}	-1.97×10^{-1}	1.14
aug-cc-pCVTZ	89	-6.24×10^{-1}	9.54×10^{-1}	−7.59	-8.27×10^{-1}	-3.81×10^{-1}	-1.62×10^{-1}	2.25×10^{-1}
d-aug-cc-pVTZ	89	-3.46×10^{-2}	3.18×10^{-1}	−1.39	-1.69×10^{-1}	3.77×10^{-2}	7.27×10^{-2}	1.45
d-aug-cc-pCVTZ	89	2.48×10^{-2}	1.01×10^{-1}	-4.07×10^{-1}	-1.82×10^{-2}	3.45×10^{-2}	6.42×10^{-2}	4.03×10^{-1}
aug-cc-pVQZ	89	-1.58×10^{-1}	3.36×10^{-1}	−2.01	-2.24×10^{-1}	-1.09×10^{-1}	-2.55×10^{-2}	1.10
aug-cc-pCVQZ	89	-1.14×10^{-1}	2.42×10^{-1}	−1.99	-1.25×10^{-1}	-5.23×10^{-2}	-1.79×10^{-2}	4.26×10^{-2}
d-aug-cc-pVQZ	89	-3.25×10^{-2}	2.27×10^{-1}	−1.07	-9.34×10^{-2}	9.75×10^{-3}	2.18×10^{-2}	1.11
d-aug-cc-pCVQZ	89	8.90×10^{-3}	3.01×10^{-2}	-1.11×10^{-1}	-1.15×10^{-3}	1.13×10^{-2}	2.03×10^{-2}	1.29×10^{-1}
aug-cc-pV5Z	75	-9.15×10^{-2}	1.06×10^{-1}	-5.69×10^{-1}	-1.24×10^{-1}	-5.64×10^{-2}	-2.07×10^{-2}	3.00×10^{-2}
d-aug-cc-pV5Z	75	-1.07×10^{-2}	2.70×10^{-2}	-1.07×10^{-1}	-2.46×10^{-2}	-5.54×10^{-3}	4.83×10^{-3}	7.45×10^{-2}

[a]Data includes the count of data points and the mean, standard deviation, minimum, quartiles, and maximum error values.

Table 4. Statistical Summary of Anisotropic Polarizability Errors in Basis Sets[a]

	count	mean	std	min	25%	50%	75%	max
aug-cc-pVDZ	82	6.62	1.30×10^{1}	-3.56×10^{1}	1.43	3.98	7.51	6.58×10^{1}
aug-cc-pVTZ	82	1.78	6.33	-2.41×10^{1}	1.41×10^{-1}	7.62×10^{-1}	1.57	3.90×10^{1}
aug-cc-pVQZ	82	2.94×10^{-1}	2.17	-1.19×10^{1}	-4.88×10^{-2}	1.57×10^{-1}	4.69×10^{-1}	5.47
aug-cc-pV5Z	70	1.33×10^{-1}	6.39×10^{-1}	−3.89	6.60×10^{-3}	8.80×10^{-2}	2.13×10^{-1}	1.76
aug-cc-pCVDZ	82	6.79	1.36×10^{1}	-3.42×10^{1}	1.43	4.08	7.55	7.55×10^{1}
aug-cc-pCVTZ	82	1.77	6.12	-2.42×10^{1}	2.28×10^{-1}	6.69×10^{-1}	1.50	3.97×10^{1}
aug-cc-pCVQZ	82	3.27×10^{-1}	1.81	-1.02×10^{1}	-4.76×10^{-2}	6.85×10^{-2}	3.55×10^{-1}	9.24
d-aug-cc-pVDZ	82	1.50	6.43	-2.07×10^{1}	-2.64×10^{-1}	1.90×10^{-1}	1.15	3.43×10^{1}
d-aug-cc-pVTZ	82	-7.93×10^{-2}	2.79	-1.67×10^{1}	-7.72×10^{-2}	1.18×10^{-1}	3.49×10^{-1}	5.64
d-aug-cc-pVQZ	82	-3.11×10^{-2}	1.67	-1.20×10^{1}	-1.40×10^{-2}	4.22×10^{-2}	1.59×10^{-1}	3.67
d-aug-cc-pV5Z	70	-5.69×10^{-3}	2.68×10^{-1}	−1.46	-1.96×10^{-2}	1.03×10^{-2}	4.44×10^{-2}	4.71×10^{-1}
d-aug-cc-pCVDZ	82	1.61	5.66	-1.57×10^{1}	-2.55×10^{-1}	2.05×10^{-1}	1.25	2.86×10^{1}
d-aug-cc-pCVTZ	82	-7.65×10^{-2}	1.32	−9.61	-1.37×10^{-1}	6.53×10^{-2}	1.89×10^{-1}	3.24
d-aug-cc-pCVQZ	82	-9.75×10^{-3}	2.11×10^{-1}	−1.03	-3.17×10^{-2}	1.07×10^{-2}	6.41×10^{-2}	7.27×10^{-1}

[a]The data includes the count of data points and the mean, standard deviation, minimum, quartiles, and maximum error values.

2.1. Accuracy of MRA Polarizability Results. The accuracy of the MRA solver for the polarizability is primarily controlled by three parameters: the convergence criteria for the ground and response densities, and the numerical truncation threshold. To deliver the requested accuracy, MADNESS adaptively refines the numerical mesh and also adjusts the order of the polynomial basis to ensure computational efficiency. Since the ground-state orbitals are square-normalized to unity, the MADNESS solvers approximately maintain a relative error in the polarizability proportional to the truncation threshold and linear in the convergence criteria. To further simplify error control, in this work we define two MRA accuracy settings.

(1) Low—truncation threshold of 10^{-6} and a convergence criterion of 5×10^{-4} in the density. The convergence criterion for the residual of each response component is 5×10^{-5} relative error.

(2) High—truncation threshold of 10^{-8} and a convergence criterion of 5×10^{-6} in the density. The convergence

https://doi.org/10.1021/acs.jctc.4c00394
J. Chem. Theory Comput. 2024, 20, 5145−5156

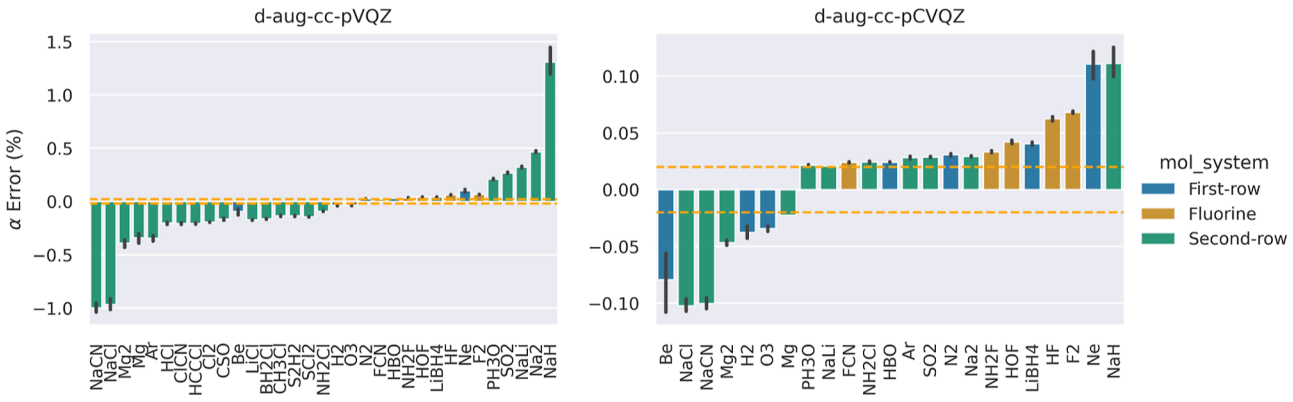


Figure 6. Isotropic polarizability outliers in aug-cc-pVQZ and d-aug-cc-pVQZ basis sets.

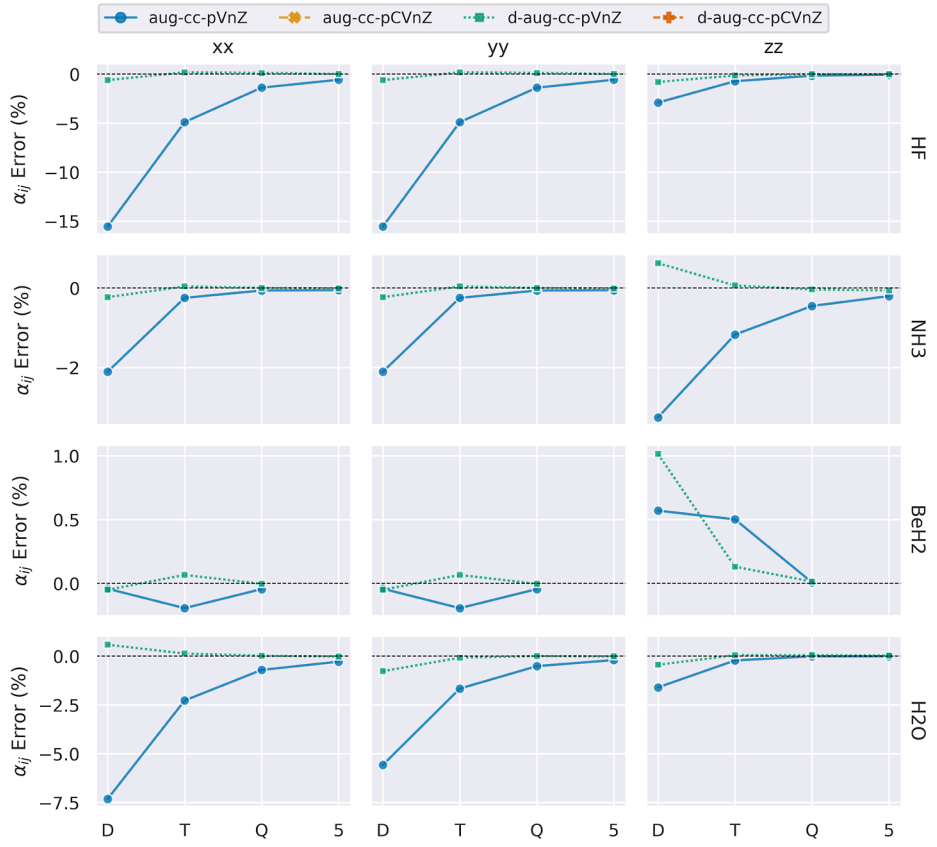


Figure 7. Percentage errors in the eigenvalues of the polarizability tensor for BeH_2, HF, NH_3, and H_2O molecules across various basis sets. The graph depicts errors for the *xx*, *yy*, and *zz* components of the tensor using singly augmented (aug-cc-pVnZ) and doubly augmented (d-aug-cc-pVnZ) basis sets, ranging from DZ to 5Z.

criterion for the residual of each response component is 5×10^{-7} relative error.

We compare against independent results for α_{zz} from very high-accuracy finite-difference calculations.[13] Figure 1, shows that we systematically obtain high accuracy by tightening convergence and truncation thresholds. For the routinely applicable low protocol we obtained for all systems a percent error approaching 10^{-2} that amounts to about 4 significant digits in the polarizability, except for NaCl, which is in error by up to 0.04%. For the high protocol 5 significant digits are obtained for most systems. For both the low and high protocols it appears that the convergence threshold on the ground and response densities is the limiting factor in accuracy as the long as the truncation threshold is sufficiently small.

The results in the main body of the paper below were obtained with the low protocol for which we adopt error bars of ±0.02% in the polarizability to guide interpretation. Energies should be accurate to at least 10^{-4} Hartree. A few fluorine systems were recomputed with the high protocol to confirm observed trends in the polarizability.

2.2. Molecular Test Set. In the Results section below, we evaluate the performance of the Gaussian basis sets on a set of 89 closed-shell molecules with geometries obtained from.[25,33] The list of molecules is given in Table 1. We computed basis set results for all molecules in the test set using basis sets provided

https://doi.org/10.1021/acs.jctc.4c00394
J. Chem. Theory Comput. 2024, 20, 5145−5156

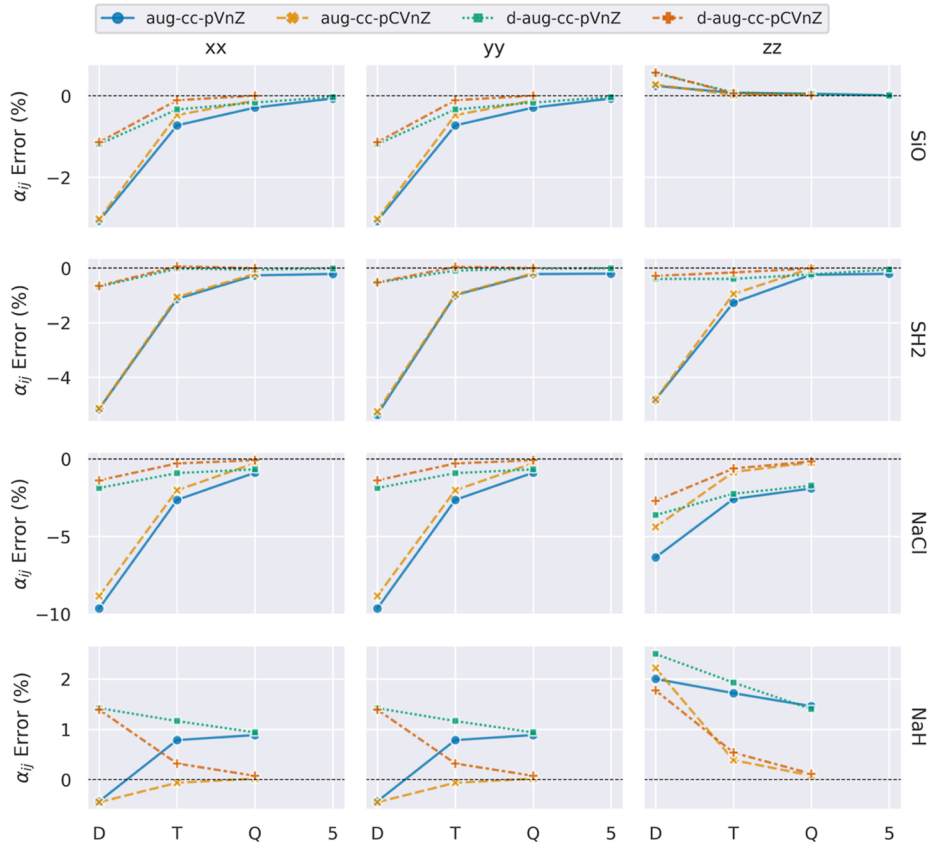


Figure 8. Percentage error in the eigenvalues of the polarizability tensor for select second row systems. The graph depicts errors for the xx, yy, and zz components of the tensor using correlation-consistent basis set variants, ranging from DZ to 5Z levels.

by default in Dalton.[6] The molecular systems containing Li, Be, Na, or Mg are not available in the 5Z basis.

3. RESULTS

3.1. Errors in the Total Energy. Figure 2 displays the distributions of signed errors in the Gaussian basis energies, while Table 2 reports the statistics of these distributions. As expected from the variational principle, all distributions are strictly positive and converge smoothly and monotonically as the basis set increases in the sequence n = D, T, Q, 5. From the figure and table, it is apparent that double augmentation does not significantly affect the total energy, as expected. However, the addition of core polarization functions does affect the energy, with the most significant effects being noticeable for outliers in the larger basis sets.

In Figure 3, we identify the molecules with the most significant errors in the d-aug-cc-pVQZ and d-aug-cc-pCVQZ basis sets. It is clear that the systems with the largest energy errors are both second-row systems and fluorine systems. When we compare the basis sets with and without the core polarization function, we discover that SO_2, SF_2, PH_3O, SiH_3F, and OF_2 have the largest errors in the d-aug-cc-pVQZ basis set (which lacks a core polarization function). However, when core-polarization functions are included, the errors in systems with second-row atoms are significantly reduced by nearly twofold. This is not the case for fluorine systems, where the largest errors are found in OF_2, SF_2, BHF_2, F_2, FNO.

3.2. Errors in the Static Polarizability. This and subsequent sections survey the polarizability results, examining the importance of valence level and augmentation with additional diffuse functions and core polarization functions. We begin here with the static results to highlight trends that remain largely consistent in the dynamic results.

Figures 4 and 5 show the distributions of the signed percentage errors in the Gaussian basis-set isotropic and anisotropic static polarizabilities, respectively, broken down by atomic composition. Statistics of those distributions are reported in Tables 3 and 4. Since the anisotropic polarizability is zero for groups T and higher (e.g., atoms and tetrahedral molecules), these were omitted from that figure.

The variational property of the polarizability applies to the individual components and to the isotropic average, but not to the anisotropic polarizability. Thus, assuming a converged ground state wave function, the errors in the isotropic polarizability should be negative, which is seen in the figures and perhaps most readily in mean error values, especially in the smaller basis sets, in Table 3. However, basis incompleteness can manifest itself differently in the ground state and response function leading to positive errors (i.e., polarizabilities that are too large) being routinely observed. Rappaport[12] demonstrated that using overly diffuse functions in a Gaussian basis set can result in non-variational polarizabilities above the complete basis set limit.[25] As explained,[12] when a Gaussian basis set lacking in the core or valence region is augmented with diffuse functions, the ground state can acquire excited-state character and hence polarizabilities above the CBS limit. Our data support this finding, with a more significant fraction of non-variational results in the doubly augmented data compared to the singly augmented data. We interpret varying convergence patterns in the isotropic results as indicating that singly augmented basis

https://doi.org/10.1021/acs.jctc.4c00394
J. Chem. Theory Comput. 2024, 20, 5145−5156

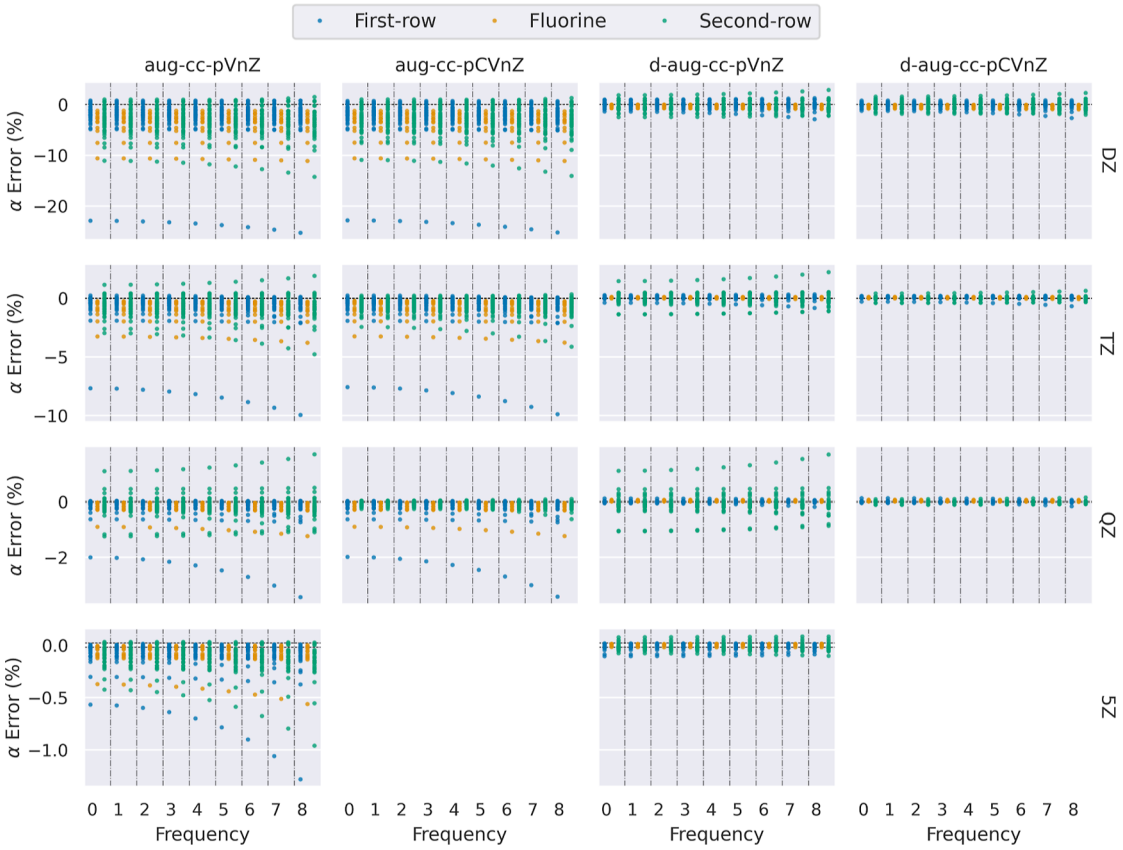


Figure 9. Errors in the Hartree−Fock frequency-dependent isotropic polarizability. The isotropic polarizability is displayed with increasing frequency from 0 (static) to 8 (maximum frequency, defined as halfway to the first excited state computed with the aug-cc-pVTZ basis set). Colors as described in 4 that examines the static polarizability.

sets have polarizability errors primarily in the response. In contrast, doubly augmented basis sets have more balanced errors between the ground and response densities.

The outliers in the isotropic polarizability without core polarization (left panel in Figure 6) are primarily second-row systems, just as they were for the total energy. With core polarization (right panel), an additional significant digit is recovered and the outliers are now a mix of first- and second-row systems.

In the largest QZ basis, the fluorine results display a small positive bias, whereas other first-row and the second-row atoms display larger, but more balanced errors. Since some of these results are within the nominal MRA error bars, we confirmed this conclusion by recomputing results at the higher precision that should yield at least 5 significant figures in the polarizability. The bias seems to be eliminated at the 5Z level, with the other systems also clustering more strongly around zero error.

Comparing the isotropic and anisotropic polarizabilities shown in Figure 4, we note that the error distributions for anisotropic polarizabilities have different shapes but similar patterns for the performance of each basis set variant. However, due to the cancellation of a significant digit when taking the differences to compute the anisotropic quantity, in any given basis, the relative errors in the anisotropy are ca. 10× those in the isotropic polarizability.

The different basis-set dependence of the anisotropic polarizability indicates a spatial directional bias in convergence. In molecules with different physics controlling polarization in different directions, components of the polarizability might converge at different rates. In Figures 7 and 8, we examine the convergence of eigenvalues of the polarizability tensor for first-row (BeH_2, HF, NH_3, and H_2O) and second-row (SiO, NaCl, LiCl, SH_2, and NaH) systems, respectively. We selected molecules that have an absolute anisotropic polarizability error greater than 4%. For example, for SiO, α_{xx} converges from below (i.e., variationally) requiring double-augmentation, core polarization and high valence level for accurate results, whereas α_{zz} converges rapidly. In contrast, in NaH, it is α_{zz} that converges most slowly. In SH_2, all components require large basis sets, however, the α_{xx} and α_{yy} converge from above whereas α_{zz} converges from below.

The results in Figures 7 and 8 demonstrate varying convergence rates for individual components of the polarizability, which complicates simple basis set extrapolations. The varying convergence rates of different components imply that the sum and difference of these components, such as isotropic and anisotropic polarizabilities, may not converge monotonically. Moreover, for many molecules, the rate of convergence of individual components varies between basis set variants, further exacerbating the challenge of obtaining convergence consistently.

3.3. Errors in the Frequency-Dependent Polarizability.

The errors in the frequency-dependent isotropic polarizability are displayed in Figure 9. It is apparent that most molecules show no or only weak frequency dependence in the error, though there are a few significant outliers. These outliers can show errors that both increase and decrease in absolute or signed error with frequency. Considering the sum-over-states ex-

https://doi.org/10.1021/acs.jctc.4c00394
J. Chem. Theory Comput. 2024, 20, 5145−5156

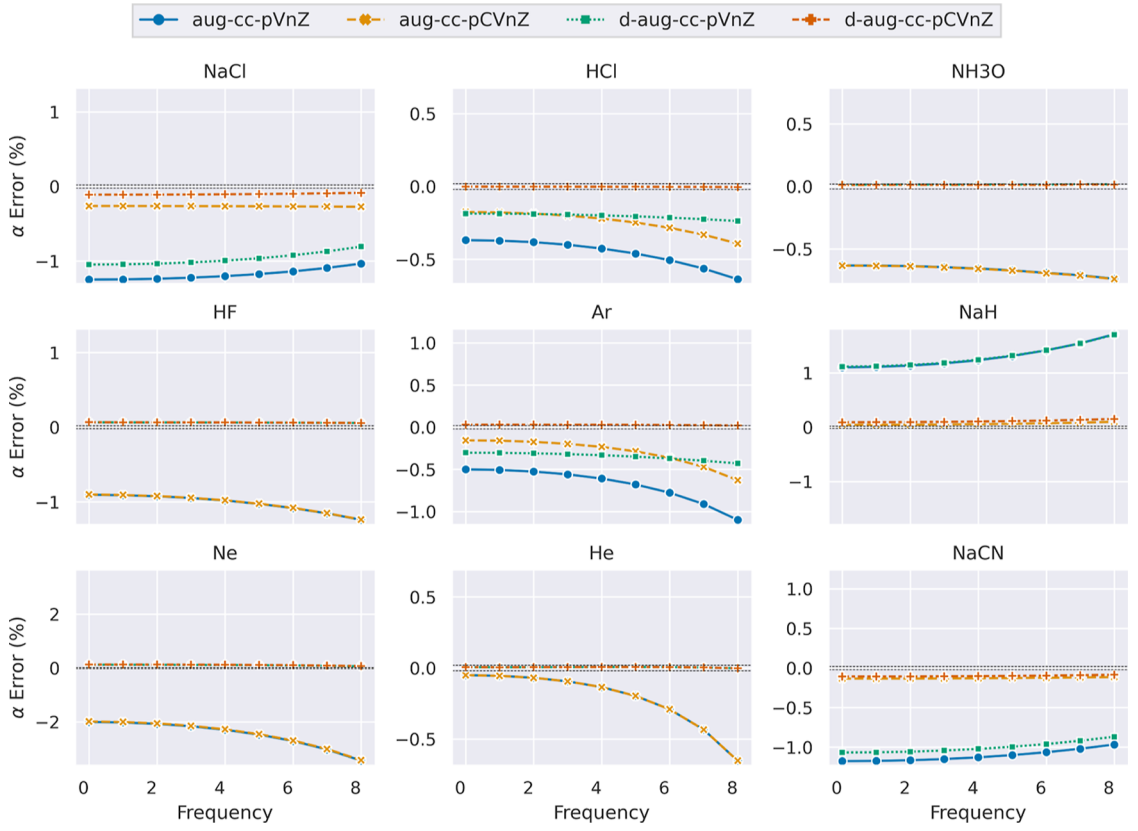


Figure 10. Errors in Hartree−Fock frequency-dependent isotropic polarizability for selected molecules in QZ basis sets. The figures show the effect of basis set type on the frequency-dependent isotropic polarizability at a given valence level. For each molecule, the frequency-dependent error varies between the basis sets types.

pression for the polarizability, the frequency dependence will arise primarily from errors in the relative position of the ground and first relevant excited state.

In Figure 10, we examine the impact of both double augmentation and core polarization on selected first- and second-row systems. For first-row systems such as Ne, HF, He, and NH_3O, the error becomes more negative with increasing frequency in the singly augmented basis. This indicates that the singly augmented excited states are too high in energy, leading to a negative bias in the polarizability, presumably due to the excited state being more diffuse than the ground state.

For second-row systems, we observe varying patterns. In the case of NaCN, NaCl, and NaH, we find that the basis sets without core-polarization functions have the largest frequency-dependent errors, which tend to increase positively. Conversely, Ar and HCl, for example, display frequency-dependent errors that increase more negatively in all but the largest d-aug-cc-pCQZ basis. For all such molecules, core-polarization functions are necessary to reduce the bias in polarizability and maintain accuracy as frequencies increase. While these molecules only represent a small subset of the data set, they exemplify the possible error trends present in frequency-dependent calculations.

In Figure 11, we examine the frequency-dependent error in the eigenvalues of the polarizability. In each component of the polarizability we find similar patterns as shown in the previous Figure 11 of the isotropic polarizability. However, the figure illustrates how this error varies among the individual components of a molecule molecules. For instance, in HF, we observe a larger frequency dependence in the transverse α_{xx} and α_{yy} directions compared to the longitudinal α_{zz} direction in singly augmented basis sets. This suggests that for HF it is the direction perpendicular to the molecular axis that is most sensitive to diffuse function augmentation. We notice this pattern in several other molecules.

3.4. Machine Learning Clustering Analysis of Basis Set Convergence. We used a Gaussian mixture model[34] (GMM) implemented using the Scikit-learn machine learning library[35] to cluster the convergence trends of the isotropic polarizability with respect to the basis set errors in each molecule. The GMM is a probabilistic model that assumes that the data is generated from a mixture of several Gaussian distributions with unknown parameters. We employed a spherical covariance type for each Gaussian distribution. To select the model parameters, i.e., the number of clusters and the covariance type, we used the Bayesian Information Criterion (BIC) as described in the sklearn documentation on model selection.[34]

From the BIC analysis, we found that the optimal number of clusters was sensitive to how the data is preprocessed. In the reported results, we preprocessed the data by first zeroing basis set errors less than a threshold of ±0.05% (motivated by the error bars on the MRA calculations) and then using a symmetric log transformation with a linear threshold of 0.05% followed by normalization so that each degree of freedom has the same range. With the preprocessed data, we determined the optimal model to contain 4 clusters. To evaluate how well our data was clustered, we used the silhouette score, which gave us a value of 0.36. Reducing the basis set truncation threshold and corresponding symmetric-log linear thresholding increases the number of clusters and the silhouette score.

https://doi.org/10.1021/acs.jctc.4c00394
J. Chem. Theory Comput. 2024, 20, 5145−5156

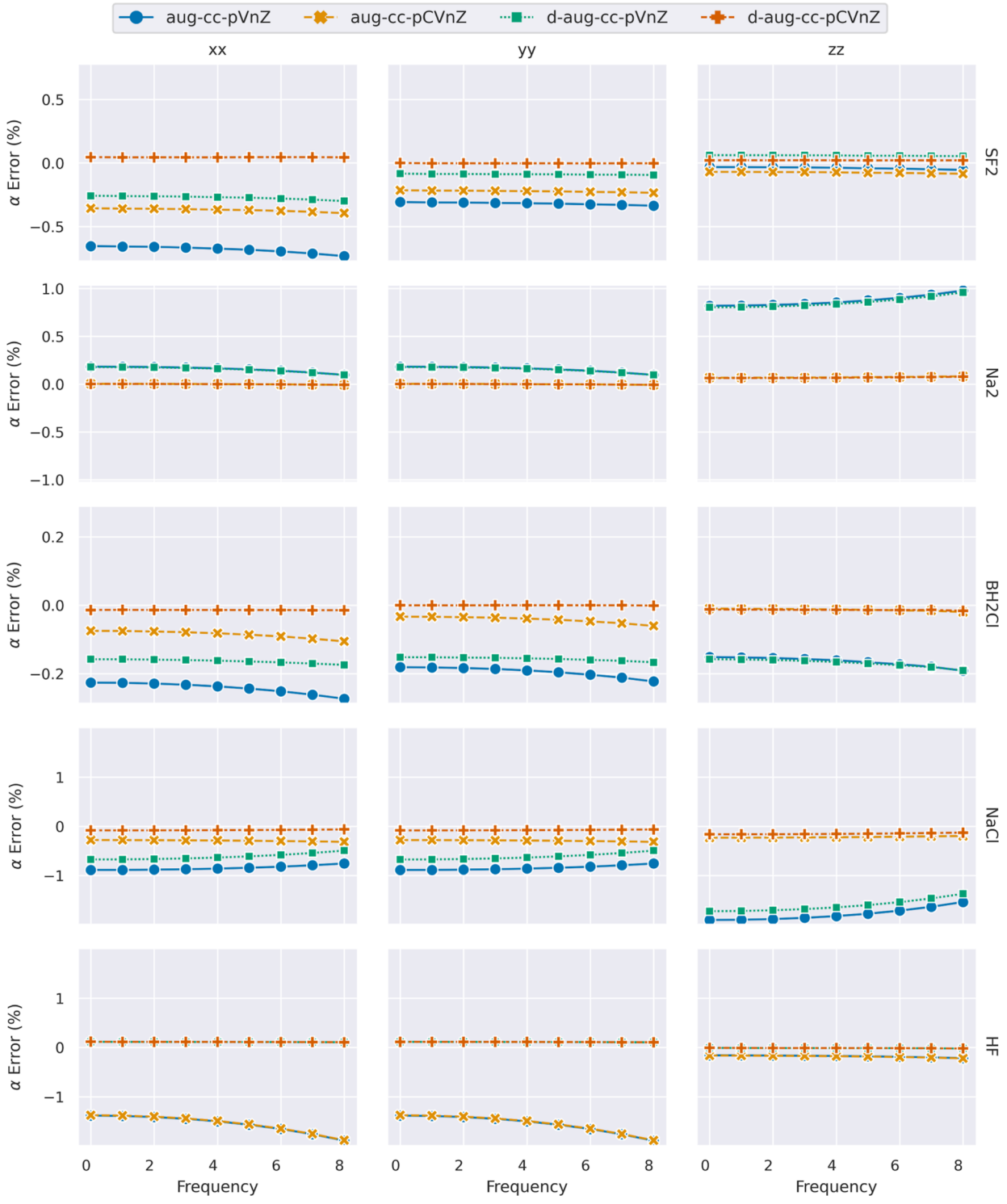


Figure 11. Errors in the Hartree−Fock frequency-dependent polarizability eigenvalues for selected molecules in QZ basis sets. The figure shows the effect of basis set type on the frequency-dependent anisotropic polarizability at a given valence level. For each molecule, the frequency-dependent error in each component of the polarizability tensor varies between the basis sets types.

The results of our clustering analysis are displayed in Figure 12, and the resulting cluster compositions are shown in Table 5. We have ordered the 4 clusters by their average basis set error in the isotropic polarizability across all basis sets, making it clear that the clusters are grouped by magnitude and pattern of their basis set convergence.

Cluster 1 composes first- and second-row that converge primarily from above in both singly and doubly augmented basis sets even with the largest QZ basis sets. The core polarization functions reduce the polarizability toward the CBS limit for these systems.

Cluster 2 contains first-row and second-row systems with basis set errors within about 2% in all basis sets, primarily converging from below in the singly augmented basis sets and from above in the doubly augmented basis sets.

https://doi.org/10.1021/acs.jctc.4c00394
J. Chem. Theory Comput. 2024, 20, 5145−5156

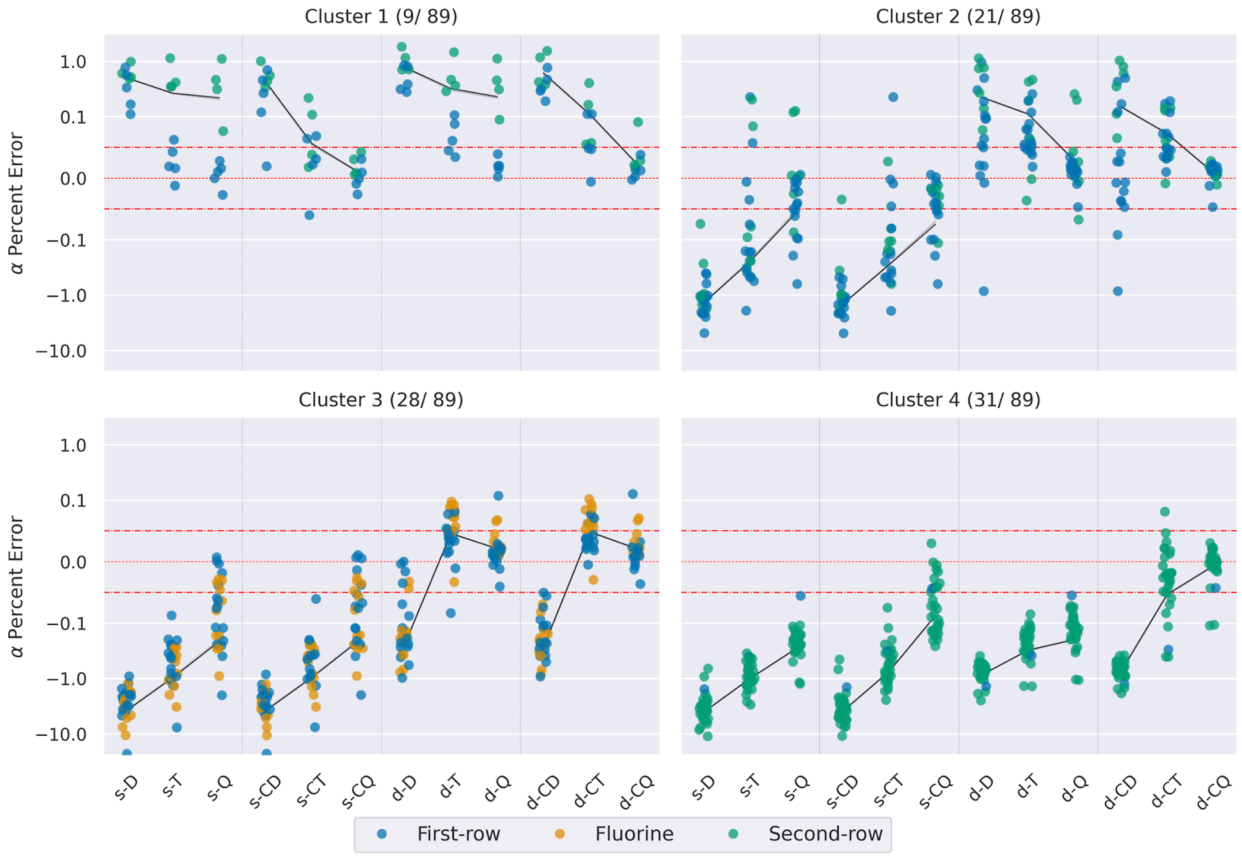


Figure 12. Gaussian mixture model (GMM) clustering of basis set convergence trends. 89 molecules have been clustered 4 groups their basis set error in the isotropic polarizability. In each plot the average basis set error is shown for each cluster in all basis sets separated by first-row, second-row and fluorine systems.

Cluster 3 is the largest cluster, containing most first-row and all fluorine systems. The systems in this cluster have a similar convergence pattern to cluster 2 but with more significant basis set errors in the singly augmented basis set and a non-monotonic convergence pattern in the doubly augmented basis sets going from negative to positive errors from DZ to TZ. Consistent with its chemical composition the errors are largely insensitive to core polarization.

Cluster 4 has the convergence patterns most sensitive to core polarization. This cluster only contains mostly second-row systems and converges from below in both the singly and doubly augmented basis sets. As we see in the figure, the core-polarization functions are essential to converge systems in cluster 4, raising the polarizability toward the CBS limit.

In summary, no chemical information was included in this clustering and yet the resulting clusters were grouped in a chemically meaningful way. This is a promising avenue for future work to improve the accuracy of basis-set extrapolations in the polarizability and other molecular properties.

4. CONCLUSIONS

As the Gaussian basis set quality improves, both the isotropic and anisotropic polarizabilities converge toward the MRA reference limit, as expected. Since the error distributions are strongly non-normal they cannot be characterized by just the mean and standard deviation. The tails of the distributions, i.e., outliers, dominate the errors.

For the static isotropic polarizability.

- For most first-row systems including fluorine, to attain 1% error in the total polarizability, a doubly augmented, DZ basis set usually suffices, with TZ required to guarantee the error is less than 1%. Single augmentation can have errors exceeding 10%.
- For second-row systems, double augmentation with core polarization at the TZ level are all essential to achieve 1% error.
- For both first- and second-row systems, doubly augmented QZ with core-polarization for second-row systems, ensure most errors are less than 0.1%.
- Neither QZ nor 5Z can guarantee errors are less than 0.1%, though it is clear from the figures and the statistics that the largest basis sets do indeed cause errors to cluster more strongly near the origin.
- Fluorine systems in the largest QZ basis have a residual positive bias in the polarizability that was verified with high accuracy calculations. This is resolved at the 5Z level.

For the static anisotropic polarizability, cancellation of significant digits exacerbates basis set convergence issues for both the Gaussian and MRA basis sets.

- For both first- and second-row systems, doubly augmented TZ, with core-polarization for second row, is required to ensure most errors are less than 10%, i.e., no smaller basis can reliably deliver even one significant decimal digit.
- The corresponding QZ basis sets deliver ca. 1% errors, with no significant improvement at the 5Z level, except for fluorine systems.

https://doi.org/10.1021/acs.jctc.4c00394
J. Chem. Theory Comput. 2024, 20, 5145−5156

Table 5. Composition of Clusters from GMM Analysis[a]

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
30	–	–	–	CH_3SH
29	–	–	–	ClCN
28	–	–	–	HCHS
27	–	–	CH_3F	LiCl
26	–	–	FNO	CH_3Cl
25	–	–	OF_2	SCl_2
24	–	–	NH_2F	Ar
23	–	–	BF	Cl_2
22	–	–	BHF_2	HBS
21	–	–	HF	OCl_2
20	–	PH_3	FCN	S_2H_2
19	–	SO_2	HCCF	CSO
18	–	PH_3O	BH_2F	Mg_2
17	–	P_2H_4	HOF	HCl
16	–	SiH_3F	F_2	HCCCl
15	–	SiH_3Cl	HNO	NaCl
14	–	CH_3NH_2	HCHO	HOCl
13	–	HCOOH	N_2	P_2
12	–	N_2H_4	H_2O	SH_2
11	–	LiH	HCN	BH_2Cl
10	–	HNC	CO_2	Mg
9	–	CH_4	NH_2OH	SiO
8	Na_2	H_2	N_2H_2	SF_2
7	NaLi	C_2H_4	Ne	HNS
6	NaH	$HCONH_2$	LiCN	NH_2Cl
5	SiH_4	CH_2BH	HBO	NaCN
4	$LiBH_4$	NH_3	CH_2NH	CS
3	Li_2	He	CO	HCP
2	BH_3	NH_3O	C_2H_2	NOCl
1	BeH_2	CH_3OH	O_3	ClF
0	CH_3BH_2	C_6H_6	HOOH	Be

[a]The table shows the composition of each cluster color-coded by system type.

- For systems such as near linear molecules in which a variety of physics contributes to the polarizability (e.g., valence charge transfer along a bond versus deformation of the tail of the wave function orthogonal to a bond), components of the polarizability may converge at different rates or be sensitive to different elements in the basis. This results in BSIE not canceling in the computation of the anisotropic polarizability, and, hence, larger overall errors.

Errors in the frequency-dependent polarizability largely follow those of the static quantity. Frequency-dependent errors are interpreted as primarily arising from the value of the relevant first excitation energy, which can originate from errors in the position of either the ground or excited states.

A follow-on study will explore convergence of these properties in larger molecules, for which less basis-set sensitivity is anticipated.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c00394.

pertinent computed data in atomic units (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Adrian Hurtado − *Department of Materials Science and Chemical Engineering, Stony Brook University, Stony Brook, New York 11794-2275, United States*; orcid.org/0000-0002-2460-8059; Email: adrian.hurtado@stonybrook.edu

Hideo Sekino − *Institute for Advanced Computational Science, Stony Brook University, Stony Brook, New York 11794-5250, United States*; Email: hideo.sekino@stonybrook.edu

Robert J. Harrison − *Institute for Advanced Computational Science, Stony Brook University, Stony Brook, New York 11794-5250, United States*; Email: robert.harrison@stonybrook.edu

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c00394

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank Stony Brook Research Computing and Cyberinfrastructure and the Institute for Advanced Computational Science at Stony Brook University for access to the high-performance SeaWulf computing system, which was made possible by $1.85M in grants from the National Science Foundation (awards 1531492 and 2215987) and matching funds from the Empire State Development's Division of Science, Technology and Innovation (NYSTAR) program (contract C210148). This research was supported in part by the Department of Energy via award DE-SC0022327 under subcontract from Virginia Tech.

■ REFERENCES

(1) Hurtado, A. Multiresolution Analysis for Accurate and Efficient Computation of Linear and Non-linear Response Properties in Quantum Chemistry. Doctoral dissertation, Stony Brook University; **2023**.

(2) Harrison, R. J.; Fann, G. I.; Yanai, T.; Gan, Z.; Beylkin, G. Multiresolution Quantum Chemistry: Basic Theory and Initial Applications. *J. Chem. Phys.* **2004**, *121*, 11587−11598.

(3) Yanai, T.; Fann, G. I.; Gan, Z.; Harrison, R. J.; Beylkin, G. Multiresolution Quantum Chemistry in Multiwavelet Bases: Analytic Derivatives for Hartree-Fock and Density Functional Theory. *J. Chem. Phys.* **2004**, *121*, 2866−2876.

(4) Sekino, H.; Maeda, Y.; Yanai, T.; Harrison, R. J. Basis Set Limit Hartree-Fock and Density Functional Theory Response Property Evaluation by Multiresolution Multiwavelet Basis. *J. Chem. Phys.* **2008**, *129*, 034111.

(5) Furche, F. On the Density Matrix Based Approach to Time-Dependent Density Functional Response Theory. *J. Chem. Phys.* **2001**, *114*, 5982−5992.

(6) Aidas, K.; Angeli, C.; Bak, K. L.; Bakken, V.; Bast, R.; Boman, L.; Christiansen, O.; Cimiraglia, R.; Coriani, S.; Dahle, P.; et al. The Dalton Quantum Chemistry Program System. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 269−284.

(7) Aprà, E.; Bylaska, E. J.; de Jong, W. A.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; van Dam, H. J. J.; Alexeev, Y.; Anchell, J.; Anisimov, V.; Aquino, F. W.; Atta-Fynn, R.; Autschbach, J.; Bauman, N. P.; Becca, J. C.; Bernholdt, D. E.; Bhaskaran-Nair, K.; Bogatko, S.; Borowski, P.; Boschen, J.; Brabec, J.; Bruner, A.; Cauët, E.; Chen, Y.; Chuev, G. N.; Cramer, C. J.; Daily, J.; Deegan, M. J. O.; Dunning, T. H.; Dupuis, M.; et al. NWChem: Past, Present, and Future. *J. Chem. Phys.* **2020**, *152*, 184102.

(8) Hill, J. G. Gaussian Basis Sets for Molecular Applications. *Int. J. Quantum Chem.* **2013**, *113*, 21−34.

(9) Jensen, F. Atomic Orbital Basis Sets. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 273−295.

(10) Sundahl, B.; Bunner, C.; Gentles, E.; Anderson, J.; Sekino, H.; Harrison, R. J. *The Basis Set Truncation Error Revisited: Comparison of Multiwavelet and Gaussian Basis Sets for Computing Properties from First Principles*, 2021, https://zenodo.org/records/4514999.

(11) Harrison, R. J.; Beylkin, G.; Bischoff, F. A.; Calvin, J. A.; Fann, G. I.; Fosso-Tande, J.; Galindo, D.; Hammond, J. R.; Hartman-Baker, R.; Hill, J. C.; Jia, J.; Kottmann, J. S.; Yvonne Ou, M.-J.; Pei, J.; Ratcliff, L. E.; Reuter, M. G.; Richie-Halford, A. C.; Romero, N. A.; Sekino, H.; Shelton, W. A.; Sundahl, B. E.; Thornton, W. S.; Valeev, E. F.; Vázquez-Mayagoitia, Á.; Vence, N.; Yanai, T.; Yokoi, Y. MADNESS A Multiresolution, Adaptive Numerical Environment for Scientific Simulation. *SIAM J. Sci. Comput.* **2016**, *38*, S123−S142.
(12) Rappoport, D. Basis-Set Quality and Basis-Set Bias in Molecular Property Calculations. *ChemPhysChem* **2011**, *12*, 3404−3413.
(13) Kobus, J. Hartree-Fock Limit Values of Multipole Moments, Polarizabilities, and Hyperpolarizabilities for Atoms and Diatomic Molecules. *Phys. Rev. A: At., Mol., Opt. Phys.* **2015**, *91*, 022501.
(14) Alpert, B. K. A Class of Bases in $L2$ for the Sparse Representation of Integral Operators. *SIAM J. Math. Anal.* **1993**, *24*, 246−262.
(15) Alpert, B.; Beylkin, G.; Gines, D.; Vozovoi, L. Adaptive Solution of Partial Differential Equations in Multiwavelet Bases. *J. Comput. Phys.* **2002**, *182*, 149−190.
(16) Bast, R.; Bjorgve, M.; Di Remigio, R.; Durdek, A.; Frediani, L.; Gerez, G.; Jensen, S. R.; Juselius, J.; Monstad, R.; Wind, P. *MRChem.: MultiResolution Chemistry*, 2024.https://zenodo.org/records/4146976.
(17) Jensen, S. R.; Bjorgve, M.; Wind, P.; Remigio, R. D.; Frediani, L.; Gerez, G. *MRChemSoft/Mrcpp: MRCPP-v1.2.0: Minor Release*, 2020, https://zenodo.org/records/3749748.
(18) Frediani, L.; Sundholm, D. Real-Space Numerical Grid Methods in Quantum Chemistry. *Phys. Chem. Chem. Phys.* **2015**, *17*, 31357−31359.
(19) Yanai, T.; Fann, G. I.; Beylkin, G.; Harrison, R. J. Multiresolution Quantum Chemistry in Multiwavelet Bases: Excited States from Time-Dependent Hartree-Fock and Density Functional Theory via Linear Response. *Phys. Chem. Chem. Phys.* **2015**, *17*, 31405−31416.
(20) Yanai, T.; Harrison, R. J.; Handy, N. C. Multiresolution Quantum Chemistry in Multiwavelet Bases: Time-dependent Density Functional Theory with Asymptotically Corrected Potentials in Local Density and Generalized Gradient Approximations. *Mol. Phys.* **2005**, *103*, 413−424.
(21) Kottmann, J. S.; Höfener, S.; Bischoff, F. A. Numerically Accurate Linear Response-Properties in the Configuration-Interaction Singles (CIS) Approximation. *Phys. Chem. Chem. Phys.* **2015**, *17*, 31453−31462.
(22) Wind, P.; Bjørgve, M.; Brakestad, A.; Gerez S, G. A.; Jensen, S. R.; Eikås, R. D. R.; Frediani, L. MRChem Multiresolution Analysis Code for Molecular Electronic Structure Calculations: Performance and Scaling Properties. *J. Chem. Theory Comput.* **2023**, *19*, 137−146.
(23) Pitteloud, Q.; Wind, P.; Jensen, S. R.; Frediani, L.; Jensen, F. Quantifying Intramolecular Basis Set Superposition Errors. *J. Chem. Theory Comput.* **2023**, *19*, 5863−5871.
(24) Jensen, S. R.; Saha, S.; Flores-Livas, J. A.; Huhn, W.; Blum, V.; Goedecker, S.; Frediani, L. The Elephant in the Room of Density Functional Theory Calculations. *J. Phys. Chem. Lett.* **2017**, *8*, 1449−1457.
(25) Brakestad, A.; Jensen, S. R.; Wind, P.; D'Alessandro, M.; Genovese, L.; Hopmann, K. H.; Frediani, L. Static Polarizabilities at the Basis Set Limit: A Benchmark of 124 Species. *J. Chem. Theory Comput.* **2020**, *16*, 4874−4882.
(26) Jensen, S. R.; Flå, T.; Jonsson, D.; Monstad, R. S.; Ruud, K.; Frediani, L. Magnetic Properties with Multiwavelets and DFT: The Complete Basis Set Limit Achieved. *Phys. Chem. Chem. Phys.* **2016**, *18*, 21145−21161.
(27) Lehtola, S.; Marques, M. A. L. Many Recent Density Functionals Are Numerically Ill-Behaved. *J. Chem. Phys.* **2022**, *157*, 174114.
(28) Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(29) Peterson, K. A.; Dunning, T. H. Accurate Correlation Consistent Basis Sets for Molecular Core−Valence Correlation Effects: The Second Row Atoms Al−Ar, and the First Row Atoms B−Ne Revisited. *J. Chem. Phys.* **2002**, *117*, 10548−10560.
(30) Woon, D. E.; Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. IV. Calculation of Static Electrical Response Properties. *J. Chem. Phys.* **1994**, *100*, 2975−2988.
(31) Woon, D. E.; Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. V. Core-valence Basis Sets for Boron through Neon. *J. Chem. Phys.* **1995**, *103*, 4572−4585.
(32) Hickey, A. L.; Rowley, C. N. Benchmarking Quantum Chemical Methods for the Calculation of Molecular Dipole Moments and Polarizabilities. *J. Phys. Chem. A* **2014**, *118*, 3678−3687.
(33) Hait, D.; Head-Gordon, M. How Accurate Are Static Polarizability Predictions from Density Functional Theory? An Assessment over 132 Species at Equilibrium Geometry. *Phys. Chem. Chem. Phys.* **2018**, *20*, 19800−19810.
(34) Mixture Models and EM. In *Pattern Recognition and Machine Learning. Information Science and Statistics*; Bishop, C. M., Ed.; Springer: New York, NY, 2006, DOI: 10.1007/978-0-387-45528-0_9.
(35) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-Learn: Machine Learning in Python *Machine Learning in Python*, 2011.