PCCP

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Polarization Consistent Basis Sets with Projector Augmented Wave Method: A Renovation Brought by PAW into Gaussian Basis Sets†

Quan Manh Phung,*[a,b] Masaya Hagai,[a] Xiao-Gen Xiong,[c] and Takeshi Yanai*[a,b]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

A recently introduced framework incorporating the Projector Augmented Wave method and Gauss-type function (GTF-PAW) [*J. Chem. Theory Comput.* 2017, 13, 3236–3249] opens alternative possibilities to perform low-cost molecular computational chemistry calculations. In this work, we present our first attempt to expand the applicability of this method by developing a family of compact general contracted polarization consistent basis sets (PAW-L*n*) as optimized GTF basis in combination with PAW. The results show that PAW-L*n*, despite having small numbers of primitives, can provide not only better performance than effective core potential (ECP) but also good accuracy and desirable systematic convergence compared to larger all-electron basis sets. This demonstrates that GTF-PAW using the PAW-L*n* basis sets could be a better alternative to both conventional all-electron- and ECP-based approaches for routine DFT calculations.

1 Introduction

Kohn-Sham density functional theory (KS-DFT) is undoubtedly the most popular theoretical method to study a wide range of molecular systems[1,2]. Over the years, it is not only employed by computational chemists/physicists but also routinely used in many experimental groups as a supplemental tool to give insights into the understanding of their experimental results. Nowadays, KS-DFT calculations of systems containing hundreds of atoms can be done with reasonable computational resources (computer time, memory, and human time). Large-scale DFT simulations are achievable with linear-scaling technique on supercomputer[3,4].

The enormous popularity of KS-DFT is perhaps attributable to its low cost, augmented with reduced scaling methods[5–16] and the Resolution of the Identity (RI) approximation for the electron integrals[17–24], as well as the development of various accurate density functional approximations (DFAs)[2]. Obviously, a straightforward way to achieve lower computational cost is to decrease the number of Kohn-Sham orbitals or the basis set size, preferably without loss in accuracy. The pseudopotential based method[25–43] makes use of this idea by replacing core electrons of an atom and its nucleus with an effective potential. Thus, only valence electrons are explicitly described and a large set of basis functions required to characterize the core electrons is eliminated. The elimination can be further promoted by the combination of the pseudopotential with augmented basis (or plane-wave) method[44,45]. Among different pseudopotential and augmented-basis schemes, Projector Augmented Wave (PAW), introduced by Blöchl[46], is considered one of the best methods in terms of accuracy and computational efficiency. Furthermore, the PAW method allows the all-electron wavefunctions to be reconstructed from the pseudo-wavefunctions, thus properties depending on the core wavefunction, for example NMR chemical shift[47], can be calculated. PAW has attracted much interest within the solid-state community and has been implemented in many solid-state physics software employing plane-wave basis sets, e.g. VASP[48], ABINIT[49], CASTEP[50], Quantum ESPRESSO[51], and using other type basis, such as tabulated numerical basis in GPAW[52], wavelets in BIGDFT[53], Lagrange-sinc basis in ACE[54], etc.

Unfortunately, within the quantum chemistry community the PAW method seems to be overlooked. One reason for this is that the effective core potential (ECP) method has been well-established long before and widely-recognized as a highly-successful pseudopotential (PP) method that can give satisfactory accuracy[55]. The ECP has been implemented in most quantum chemistry (QC) packages and mainly serves as a relativistic method to treat heavy elements. A central difference in formal-

[a]*Department of Chemistry, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8602, Japan; E-mail: quan.phung@chem.nagoya-u.ac.jp; yanait@chem.nagoya-u.ac.jp*
[b]*Institute of Transformative Bio-Molecules (WPI-ITbM), Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8602, Japan*
[c]*Sino-French Institute of Nuclear Engineering and Technology, Sun Yat-Sen University, Zhuhai, 519082, China*

† Electronic Supplementary Information (ESI) available: PDF file containing (i) parameters used to construct the PAW atomic dataset; (ii) list of molecules used in the test set; (iii) mean and maximum absolute errors of atomization energies per atom, vertical ionization energies, and noncovalent binding energies; (iv) mean and maximum root-mean-square deviation (RMSD) of atomic positions. Spreadsheet containing (i) atomic and diatomic total energies of H, Be, Mg, N, and P, using different basis set compositions; (ii) total energies of molecules of the G2-97′ and S22 datasets; (iii) RMSD values of the molecules of the G2-97′ dataset. Text files containing PAW-L*n* basis sets, the PAW dataset in modified ABINIT format, and ATOMPAW input to generate the dataset. See DOI: 10.1039/cXCP00000x/

ism between ECP and PAW methods is that ECP is classified to the norm-conserving type[56–60], whereas PAW is of the ultrasoft type[46,61]. In ECP, the PP associated with the frozen atomic core states is written as $\hat{V}_{\rm PP}(r) = U_L(r) + \sum_{l=0}^{L-1}\sum_{m=-l}^{l}|Y_{lm}\rangle U_l(r)\langle Y_{lm}|$, given that we here neglect the spin-orbit effect for simplicity. The indices L (as well as l) and m are angular and magnetic quantum numbers, respectively, $|Y_{lm}\rangle$ are the spherical harmonics as projectors, and $U_l(r)$ are potential functions. In the QC implementations, a linear combination of Gaussians multiplied by powers of the electron-core distance is used for the representation of $U_l(r)$ for ease of integration. The potential function $\hat{V}_{\rm PP}(r)$ is kept fixed at the predetermined atomic form in molecular calculations. The resultant valence orbitals $\psi_v(r)$ have nodeless shape in the inner-core region, but usually agree well with all-electron orbitals in the valence region. The density is defined using the normalized ψ_v ($\langle\psi_v|\psi_v\rangle = 1$) as the method is typed as norm-conserved.

A distinct difference of PAW from ECP arises from the starting formalism that is to introduce the exact transformation of the atomic state $|\psi_v^a\rangle = |\tilde{\psi}_v^a\rangle + (|\psi_v^a\rangle - |\tilde{\psi}_v^a\rangle) = \hat{T}|\tilde{\psi}_v^a\rangle$, where $|\psi_v^a\rangle$ and $|\tilde{\psi}_v^a\rangle$ are true (or all-electron) and pseudo orbitals, respectively, and the superscript a indicates the atomic state. The ultrasoft formalism built into PAW means that $|\tilde{\psi}_v^a\rangle$ can be numerically insignificant and indeed its norm is usually much less than 1, whereas in the ECP method, orbital states are normalized to a norm of 1, as mentioned above. The transformation operator $\hat{T}$ is written as $\hat{T} = 1 + \sum_{v'}(|\psi_{v'}^a\rangle - |\tilde{\psi}_{v'}^a\rangle)\langle p_{v'}^a|$ using PAW's projectors $|p_v^a\rangle$, which satisfy the biorthogonality $\langle p_v^a|\psi_{v'}^a\rangle = \delta_{vv'}$. The reconstruction of the all-electron shape from the pseudo orbitals via the predetermined $\hat{T}$ is at the heart of the augmented wave scheme of PAW and unavailable in ECP. This scheme is exploited as a key to derive the pseudized one-electron Schrödinger equation in a rigorous manner as $\hat{T}^\dagger\hat{F}\hat{T}\,|\tilde{\psi}_v^a\rangle = \varepsilon_v\hat{T}^\dagger\hat{T}\,|\tilde{\psi}_v^a\rangle$ where $\hat{F}$ is the all-electron Fock operator and ε_v is the orbital energy. The pseudized Fock operator is written as $\hat{T}^\dagger\hat{F}\hat{T} = -\frac{1}{2}\nabla^2 + U_{\rm loc}^a(r) + \sum_{vv'}|p_v^a\rangle U_{vv'}^a\langle p_{v'}^a|$, in which the potential term at first glance takes a form similar to the ECP but critically differs from it in the sense that PAW's potential is adaptively modulated in response to the all-electron molecular $\hat{F}$ in a self-consistent manner. The coefficients of the nonlocal term $U_{vv'}^a$ are computed using all-electron atomic orbital information from a unique mapping of all-electron $\hat{F}$ to the pseudization representation, rigorously derived as $U_{vv'}^a = \langle\psi_v^a|\hat{F}|\psi_{v'}^a\rangle - \langle\tilde{\psi}_v^a|\hat{F}|\tilde{\psi}_{v'}^a\rangle$ for atomic state. This mapping, built upon accurate projection, can totally avoid artificial fitting parameterization with few exceptions. In this sense, PAW is not simply a PP method but thus characterized as a generalization of the PP and augmented-wave methods, as emphasized in the original work of Blöchl[46]. In contrast, there is no posterior adjustment taking place in the given PPs in the ECP calculations. In addition, the construction of ECP's PP entails a certain arbitrariness in its fitting method, which is based on either atomic shapes or energies[55]. It should also be noted that $U_l(r)$ of ECP's PP represented with the Gaussians is uniformly sphere or isotropic, whereas the PAW formally retains anisotropic structures of the local and nonlocal potentials. We can construct the PAW pseudization without the frozen core (FC) approximation unlike ECP, although the FC approximation is employed in most cases.

In a recent work of Xiong and Yanai[62], a framework to incorporate the PAW method into the Gauss-type function- (GTF) based molecular DFT code, referred to as GTF-PAW, was introduced. The generality of the PAW method allows the GTF-PAW method to be implemented in the existing quantum chemistry program in a rather straightforward manner. Our prototype implementation has been achieved by linking our in-house Gaussian basis code with a portable PAW library LIBPAW, developed by Rangel *et al.*[53] for the plane-wave code ABINIT[49,63,64]. Most of PAW procedures can be performed with the existing subroutines offered by LIBPAW. In the previous study, the complementary PAW treatments to handle GTF basis was newly and sub-optimally implemented[62]. Using the atomic code ATOMPAW, developed by Holzwarth *et al.*[65,66], the PAW atomic data required by the LIBPAW library can be easily generated in a nearly blackbox manner.

In the GTF-PAW method, we fully utilize all features of the PAW method and employ Gaussian basis functions for basis to represent PAW's pseudo waves, as was reported in our previous paper[62]. In this paper, we do not repeat the detailed explanation of the original PAW method including basis transformation, pseudization, frozen core treatment, and various technical aspects, for which the readers should consult the original paper[46] or other well-written literatures[28,52,65].

Similar to ECP, high-exponent primitive GTFs are no longer required in GTF-PAW. Moreover, it is possible to reproduce the conventional all-electron total energies (to within 1 m$E_{\rm h}$) with GTF-PAW, providing that adequate frozen core treatments and a near-complete basis (e.g. uncontracted cc-pVQZ) were employed. The work of Xiong and Yanai[62] demonstrated that GTF-PAW is a reduced-cost promising alternative to ECP and even conventional all-electron methods. The work serves as a starting point for our future research on efficient and user-friendly PAW-based methods (both DFT and ab initio). In our very recent work[67], the level of the functional treatment in GTF-PAW code was raised to the generalized gradient approximation (GGA). In addition, the uniform mesh grid for the DFT quadrature was introduced in the Gaussian code as a feasible alternative to the traditional Becke multicenter fuzzy cell grid[68].

In this work, we tackle one of the important issues that were raised in the previous work[62], that is the need of atomic basis sets well-prepared for GTF-PAW. Two strategies may be employed to overcome this issue. The first one is to simply use existing uncontracted all-electron basis sets and truncate them, i.e. remove all inappropriate tight basis functions[62]. However, for practical applications that can best exploit the efficiency of GTF-PAW, some questions remain unanswered: Which uncontracted all-electron basis set should be used? Which tight basis functions should be eliminated? What is the accuracy of such truncated basis set? Nevertheless, this kind of truncated basis set is potentially either unbalanced or inefficient, hence diminishing the appeal of the GTF-PAW method. Obviously, a superior strategy is to design new basis sets for GTF-PAW, explicitly optimize their compositions and exponents, and benchmark their performance against all-electron calculations. In this work, we pursue this strategy and propose a sequence of polarization consistent basis sets as atomic GTF basis for routine molecular calculations with GTF-PAW, hereafter

referred to as PAW-Ln. The index n denotes the level of polarization compared to the isolated atom, e.g. for carbon, PAW-L0 has sp, PAW-L1 has spd, PAW-L2 has $spdf$ basis functions, etc. These basis sets were built in a manner similar to the all-electron polarization consistent basis sets (pc-n) introduced in the works of Jensen *et al.*[69–73] The most important characteristic of this family is that functions contributing similar amounts of energy are included at the same stage. It was shown that using this hierarchy of polarization consistent basis sets, total energies, atomization energies, equilibrium distances, and dipole moments converge monotonically towards the complete basis set (CBS) limit. The major disadvantage of this family is that the pc-3 and pc-4 basis sets are quite large. For instance carbon pc-4 has 131 primitives ($18s11p6d3f2g1h$) contracted to 109 basis functions [$8s7p6d3f2g1h$]. It is much larger than the largest Karlsruhe basis set def2-QZVPPD ($16s8p4d2f1g$)→[$8s4p4d2f1g$] or the cc-pV5Z (108 primitives, 91 contracted basis functions). Such large basis sets are apparently applicable for small molecules and may be redundant for standard DFT applications. It should also be noted that due to the (typically large) intrinsic error of DFAs, results calculated with sizeable basis sets are not necessarily better than those with smaller basis sets. Considering this fact, we will only focus on developing small- to medium-sized basis sets *explicitly* used for GTF-PAW. The PAW-Ln series is limited to $n = 1$ and 2. Nevertheless, due to the PAW frozen core treatment, we expect that this series is able to yield comparable results to those obtained from more expensive all-electron calculations. The incorporation of the relativistic effect into the PAW for the plane-wave code was investigated by earlier studies; however, originally, the PAW was not introduced as a relativistic method but developed as a general numerical framework, which was formulated based on the non-relativistic Kohn-Sham equation[46]. In a similar manner, this study attempts to show the development of basis sets with GTF-PAW in the non-relativistic regime. This naturally serves as an essential proof-of-concept to verify the validity and usability of our direct adaptation of PAW to quantum chemical methods, which usually handle chemical systems mostly consisting of light elements. As a preliminary study, we will thus work with the first-, second-, and third-period elements, except noble-gas elements.

2 Generation of PAW Atomic Dataset

The PAW formalism requires the so-called "PAW atomic data" file for each element. Over the years, many PAW atomic datasets have been constructed[28,52,66,74,75], unfortunately, they are primarily designed for solid-state simulations. Thus, we decided to generate a new PAW atomic dataset for our molecular calculations. The dataset was produced following the recipe described in the work of Jollet *et al.*[66]. The calculations were done with the orz program package[76]. The SVWN (or LDA) exchange-correlation functional[77–79] and uncontracted pc-4 basis set were used (see Tables 1 and 2). For light elements, the contribution of relativistic effects is minimal, therefore, we employed the non-relativistic wave equation. In the partial-waves basis generation part, we chose the Vanderbilt polynomial pseudization scheme[27] for the wavefunctions and Troullier–Martins pseudization scheme for the local potential[29]. The values of the most critical parameter determining the accuracy of the PAW atomic data, the radial cutoffs r_{paw}, are listed in Table S1. To describe molecular systems with short bond lengths, we used smaller r_{paw} values than those in the existing PAW datasets[28,52,66,74,75]. r_{paw} for each element was chosen to be smaller than half the bond length of its diatomic molecule. Also the r_{paw} values were selected so that molecules with very short bond distances (according to the Computational Chemistry Comparison and Benchmark Database[80]) can be calculated. We note that one can eventually further reduce r_{paw}, thus improving the accuracy of the dataset. However, such too "hard" dataset will require large atomic basis sets, consequently, deteriorate the efficiency of the GTF-PAW method. For alkali and alkaline-earth elements (Li, Be, Na, and Mg), the "semi-core" s and p electrons were treated as valence. Hence, we avoid the so-called "ghost states"[81] and can properly describe ionic bonds with high electronegative elements (e.g. fluorine).

3 Uncontracted Basis Sets

With the generated PAW dataset, we first construct the uncontracted PAW-L1 and PAW-L2 basis sets by determining their optimal compositions and optimizing their exponents with respect to GTF-PAW total energies. All GTF-PAW calculations were performed using the SVWN (or LDA) exchange-correlation functional, D_{2h} symmetry, tight grid equivalent to Grid7 implemented in the ORCA program package[82] (770-point Lebedev angular grid and radial grid with IntAcc = 5.67). Open-shell species were calculated with the unrestricted formalism (UKS). No symmetry equivalencing was applied in the calculations. One might argue that LDA usually gives poor results and the generated basis sets would be of low quality. However, as has been shown in the work of Jensen[70], we expect that the composition of the basis sets, i.e. the number of s, p, and d functions, will remain the same. Furthermore, later Jensen[83] showed that the difference between using different XC functionals (including LDA) for basis set optimization is small and well below the basis error as compared to CBS. We thus expect that optimizing the basis sets at a higher level of theory such as B3LYP, although might improve the accuracy, should not significantly change the quality of our basis sets optimized with LDA. Because it is difficult to simultaneously optimize all exponents α, we instead generated them using the so-called even-tempered series[84]: $\alpha_{i,l} = \alpha_l \beta_l^{i-1}$, $\alpha_l, \beta_l > 0$, $i = 1$–N_l, where l is the angular momentum quantum number, α_l and β_l are parameters optimized with respect to the total (atomic or molecular) energies. The advantages of this approach are its simplicity (the dimension of the search space is drastically reduced), stability (variational collapse, i.e. two exponents collapse to the same value, is prevented), and extensibility (diffuse functions can be naturally included by extending the series). The optimization of the atomic or molecular energy E with respect to α_l and β_l was done with the BFGS algorithm. Energy gradients with respect to α_l and β_l were calculated by central finite differences. During the optimization procedure, multiple minima in the even-tempered parameter space could be located. These minima can have very similar energies with quite different α_l and β_l values. In this case, we selected the minimum with the lowest total energy. We finally note that whereas our optimization approach offers some advan-

tages, the optimized exponents are still far from optimal.

In Figure 1, we summarize the procedure to develop the PAW-L*n* basis sets. This procedure resembles the one employed in the work of Jensen [69]. Here, we describe in detail the procedure applied for hydrogen. First, we determined the compositions of the basis sets by calculating the contribution of each basis function to the total energy of the hydrogen atom (step 1). The results for the *s* functions are shown in Figure 2 (full blue line). Undoubtedly, the atomic DFT energy is independent of polarization (*p* and *d*) functions. Therefore, the impacts of these polarization functions could be only evaluated from molecular calculations (steps 2 and 3), in this case, the calculations of H_2 at d(H–H) = 0.76 Å. One can argue that this choice of bond distance is arbitrary, furthermore, the optimized exponents are better at describing the targeted molecule (H_2) instead of general molecules. Nevertheless, Jensen [85] demonstrated that basis sets obtained with this approach are capable of providing highly accurate results over a wide range of molecular systems. In both steps 2 and 3, we restricted the perturbation of the *s* functions by using up to 11*s* functions and fixing the exponents at their atomic values. The results are presented in Figure 2 (dotted red line and dashed green line). It can be seen that the convergence of the total energy with respect to the number of functions is roughly exponential as expected.

The composition of primitive GTFs ($4s1p$) is commonly employed in the pc-1, def2-SVP, cc-pVDZ, and 6-31G(d,p) basis sets and was shown above to provide an energy prediction accurate to 10^{-3} E_h with the GTF-PAW scheme. In case of hydrogen, we did not employ the frozen core approximation; however, PAW's pseudization can render the 1*s* valence wavefunction drastically smooth, so that the conventional size of the *s* primitives, i.e., ($4s$), is a bit conservative. To accommodate the the double-ζ level quality at a minimal cost, we decided to use ($3s$) for PAW-L1, anticipating that the quality of ($3s1p$) should be similar to ($4s1p$) in terms of the DZP-accuracy basis. Figure S1 shows the comparison of the σ molecular orbital (MO) of H_2 determined at $4s1p$ and $3s1p$, illustrating that the errors of these two basis representations appear to be rather small. In GTF-PAW, the smooth pseudo wave function represented with a small number of GTF functions is corrected by the PAW transformation, resulting in high-accuracy all-electron wave function. Figure 3(a) displays the all-electron wave function resulting from the pseudo wave function based on PAW-L1 ($3s1p$), which shows an even better description than those obtained using the all-electron ($4s1p$) basis from the conventional GTF implementation. PAW-L2 can be either ($6s3p1d$) or ($5s2p1d$). We chose the latter option, taking into consideration that PAW-L2 should not be larger than pc-2 and cc-pVTZ, which are ($6s2p1d$) and ($5s2p1d$), respectively. Figure 3(b) shows all-electron form of σ MO obtained with PAW-L2, which again produced a higher-accuracy description including nuclear cusp structures than that calculated using the conventional approach with the cc-pVTZ basis.

In Figures 4 and 5, we show similar plots for Be, Mg, N, and P. For Be, the compositions of PAW-L1 and PAW-L2 should be ($4s2p$) and ($6s3p1d$), respectively. These basis sets contain significantly smaller numbers of *s* functions than the all-electron basis sets, although the frozen core approximation was not employed. The reason for this reduction is that the $2s$ wavefunction (with one node) is transformed into a nodeless $2s$ smoothed pseudo-wavefuntion, therefore, tight *s* basis functions are no more required. As compared to Be, Mg requires more attention as we found that basis sets produced with the standard procedure give inferior performance. The reason is that the atomic *p* basis functions, optimized for the atomic energy, are too tight. This problem could be resolved by augmenting these tight atomic *p* basis functions with *p* polarization functions, denoted as p_{pol}. We found that one extra p_{pol} function is sufficient to improve the accuracy of the Mg basis sets. The compositions of PAW-L1 and PAW-L2 for Mg are therefore ($5s3p1p_{\mathrm{pol}}$) and ($6s4p1p_{\mathrm{pol}}1d$), respectively.

For nitrogen and phosphorus, [core] $ns^2\,np^3$, the radial pseudo-wavefunction *ns* and *np* are quite similar, thus we expect that *s* and *p* basis functions give comparable contributions to the atomic total energy. Indeed, one can observe the similar behavior of the *s* and *p* lines depicted in Figure 5. As compared to the results of Jensen [69], the lines are not very smooth, e.g. there are several prominent kinks in phosphorus. The reasons might be related to the fact that the total energy function $E(\alpha_l, \beta_l)$ has multiple minima and the BFGS algorithm is prone to converge to local minima. Nevertheless, we propose that for both N and P, PAW-L1 is ($3s3p1d$) and PAW-L2 is ($4s4p2d1f$).

Commonly, basis sets of elements within the same block and period have an identical composition (see also Tables 1 and 2). We also applied this principle in designing the PAW-L*n* basis sets, i.e. for Li–Be, PAW-L1 = ($4s2p$) and PAW-L2 = ($6s3p1d$); for Na–Mg, PAW-L1 = ($5s3p1p_{\mathrm{pol}}$) and PAW-L2 = ($6s4p1p_{\mathrm{pol}}1d$); for B–F and Al–Cl, PAW-L1 = ($3s3p1d$) and PAW-L2 = ($4s4p2d1f$). The basis set compositions are also summarized in Tables 1 and 2. As noted above, the optimized exponents of the polarization functions were obtained in calculations of symmetric homonuclear molecules X_2, d(X–X) = 2.70, 1.60, 1.26, 1.21, 1.42, 2.98, 2.45, 2.28, 1.90, and 2.00 Å, with X = Li, B, C, O, F, Na, Al, Si, S, and Cl, respectively.

After defining the basis set compositions for all elements, we then reoptimized the exponents of the polarization functions using the reduced *s* or *sp* basis to achieve maximum accuracy. The procedure is illustrated in Figure 1. For example for the PAW-L2 basis set of N ($4s4p2d1f$), we fixed the exponents of the $4s$ and $4p$ basis functions at their atomic values, while simultaneously optimizing the $2d$ and $1f$ exponents.

4 Contracted Basis Sets

Even though the PAW-L*n* basis sets are relatively small compared to uncontracted all-electron basis sets (see Tables 1 and 2), it is still beneficial in terms of computational efficiency to contract them. Similar to the work of Jensen [69], we employed the general contraction scheme introduced by Raffenetti [86], using orbital coefficients obtained from atomic DFT calculations as contraction coefficients. Here, for open-shell atoms, we used restricted open-shell DFT in combination with D_{2h} symmetry and averaged coefficients of *p*-type orbitals. As compared to the segmented contraction scheme, it is simpler, furthermore, atomic SCF energies are preserved upon contraction.

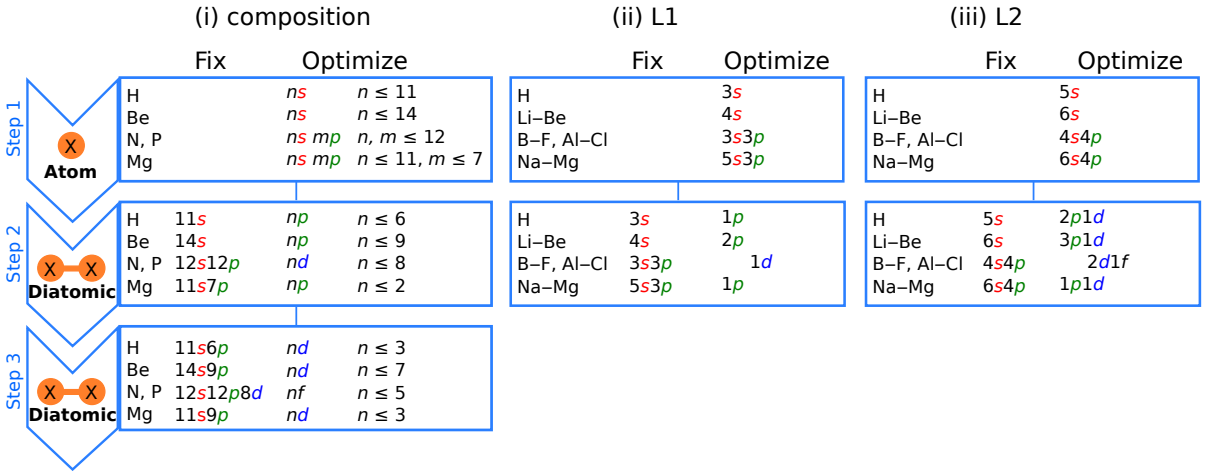

Fig. 1 Procedure to construct the PAW-Ln basis sets in this work. (i) Determine the composition of the basis sets by calculating the energy contribution of each basis function in atom X and diatomic molecule X_2, X = H, Be, N, P, and Mg. (ii) Optimize the exponents of the PAW-L1 basis set: H ($3s1p$); Li–Be ($4s2p$); Na–Mg ($5s4p$); B–F and Al–Cl ($3s3p1d$). (iii) Optimize the exponents of the PAW-L2 basis set: H ($5s2p1d$); Li–Be ($6s3p1d$); Na–Mg ($6s5p1d$); B–F and Al–Cl ($4s4p2d1f$). Polarization basis functions were optimized in diatomic calculations, whereas other basis functions were obtained from atomic calculations.

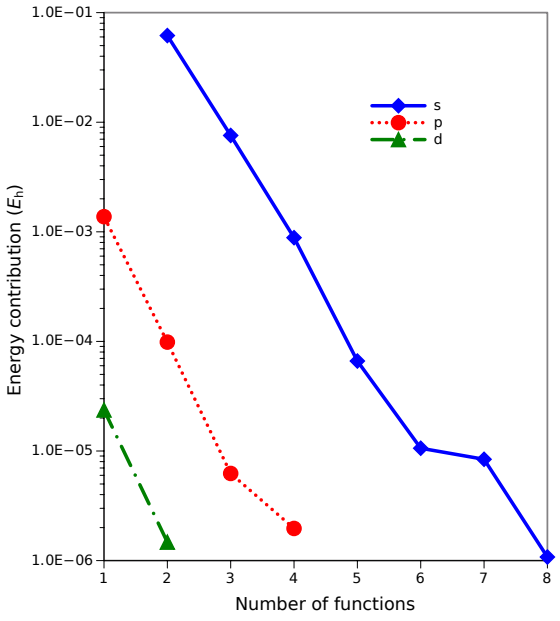

Fig. 2 Energy contribution (per atom, in E_h) of s, p, and d functions in H. The latter two were calculated in H_2 with d(H–H) = 0.76 Å, fixing the exponents of 11s functions at their atomic values. The calculations were done with the SVWN functional.

We selected the PAW-Ln contracted basis sets based on two criteria, they should be (i) smaller than (or equal to) and (ii) as accurate as the corresponding pc-n basis sets. For all elements, we were able to find contractions satisfying these two criteria. In Tables 1 and 2, we summarize the compositions of the contracted and uncontracted PAW-Ln basis sets, in comparison with other commonly used all-electron basis set families, i.e. polarization-consistent pc-n[69–73], Karlsruhe def2-[87], correlation-consistent cc-[88–90], and Pople basis sets[91–97]. pc-n and cc- are general contracted whereas the other two are segmented contracted. We note that corresponding segmented versions of pc-n having slightly fewer contracted functions and giving similar accuracy are available (pcseg-n)[98]. All basis sets, except for 6-311G(2df,2pd) of the third-period elements, were taken from the Basis Set Exchange (BSE) library[99]. The 6-311G(2df,2pd) basis set of the third-period elements was collected from the Gaussian 16 software package[100–102]. We also note that we constructed a "minimal" basis set (PAW-L0.5) which will be discussed in Section 7.

5 Performance of PAW-Ln as Compared to All-Electron Basis Sets

To assess the performance of the PAW-Ln basis sets, we performed a benchmark study on three properties: atomization energy, vertical ionization energy, and noncovalent binding energy. The systems used in the test set are listed in Table S2. This test set comprises two datasets: (i) the G2-97 dataset[103,104] augmented with additional molecules containing second- and third-period elements (255 molecules in total, referred to as G2-97′); (ii) S22 dataset consisting of small to medium-sized complexes (up to 30 atoms) of common molecules containing C, N, O, and H[105]. The results calculated with PAW-Ln (n = 1 and 2) were compared with those obtained from all-electron basis sets (shown in Tables 1 and 2). The results with the largest basis set (pc-4) were used as references. All all-electron calculations were performed with the

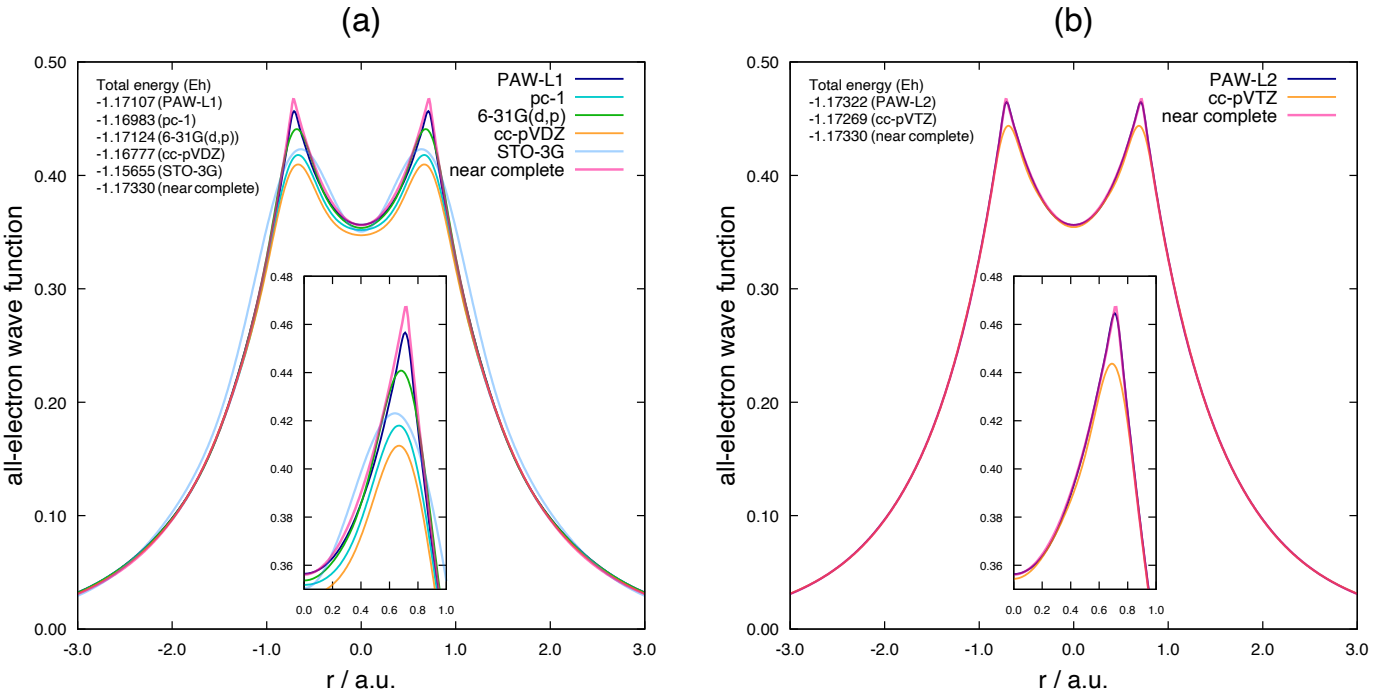

Fig. 3 All-electron wave functions obtained with GTF-PAW method using (a) uncontracted PAW-L1 and (b) uncontracted PAW-L2 basis for H_2 with d(H–H) = 0.76 Å. The plots of the corresponding pseudo waves for PAW-L1 and PAW-L2 basis are shown in Figure S1. The all-electron wave functions obtained with conventional approach using the all-electron basis sets are included for comparison. The calculations were done with the SVWN functional.

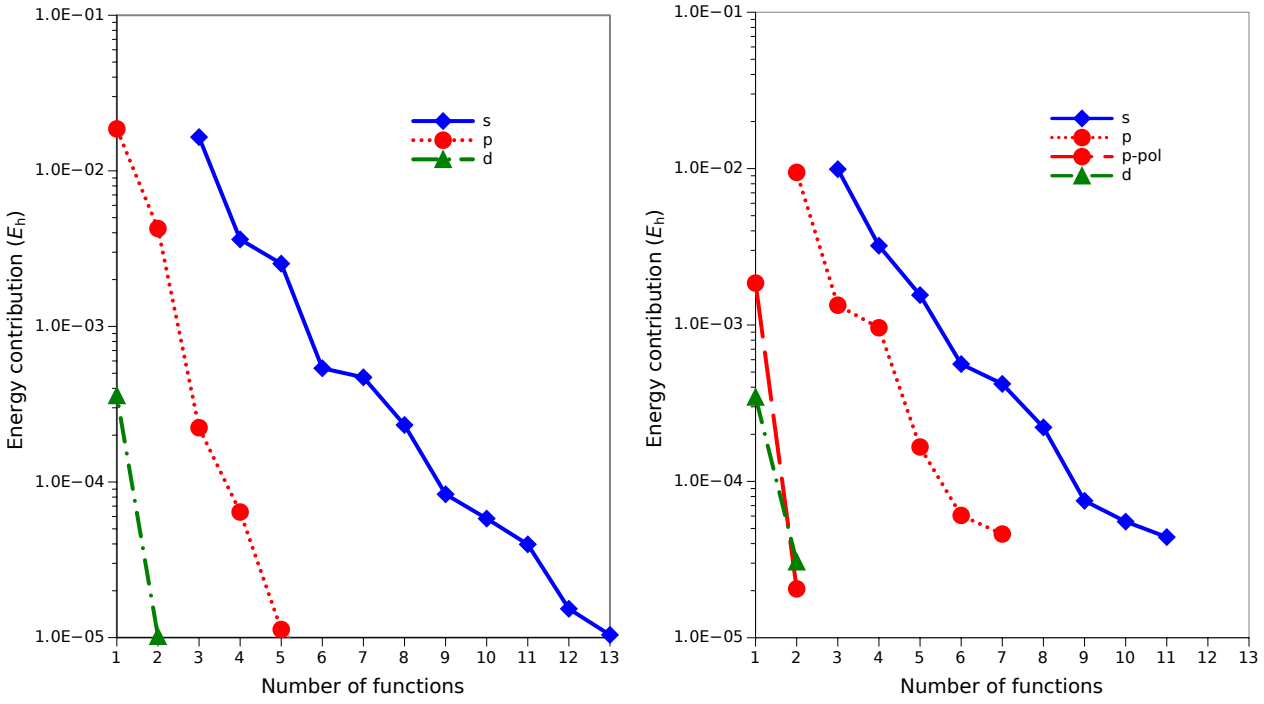

Fig. 4 Left plot: Energy contribution (per atom, in E_h) of s, p, and d functions in Be. The latter two were calculated in Be_2 with d(Be–Be) = 2.40 Å. Right plot: Energy contribution (per atom, in E_h) of s, p, p-pol, and d functions in Mg. The latter two were calculated in Mg_2 with d(Mg–Mg) = 3.38 Å. The calculations were done with the SVWN functional.

Table 1 Compositions of the first- and second-period elements basis sets used in this work

	H		Li–Be		B–F	
PAW-L0.5	$(2s1p)$		$(3s2p)$		$(2s2p1d)$	
PAW-L1	$(3s1p)$	$[2s1p]$	$(4s2p)$	$[3s2p]$	$(3s3p1d)$	$[2s2p1d]$
PAW-L2	$(5s2p1d)$	$[3s2p1d]$	$(6s3p1d)$	$[4s2p1d]$	$(4s4p2d1f)$	$[3s3p2d1f]$
pc-1	$(4s1p)$	$[2s1p]$	$(7s3p)$	$[3s2p]$	$(7s4p1d)$	$[3s2p1d]$
pc-2	$(6s2p1d)$	$[3s2p1d]$	$(10s4p1d)$	$[4s2p1d]$	$(10s6p2d1f)$	$[4s3p2d1f]$
pc-3	$(9s4p2d1f)$	$[5s4p2d1f]$	$(14s6p2d1f)$	$[6s3p2d1f]$	$(14s9p4d2f1g)$	$[6s5p4d2f1g]$
pc-4	$(11s6p3d2f1g)$	$[7s6p3d2f1g]$	$(19s8p3d2f1g)$	$[8s4p3d2f1g]$	$(18s11p6d3f2g1h)$	$[8s7p6d3f2g1h]$
def2-SVP	$(4s1p)$	$[2s1p]$	$(7s3p)^c$	$[3s2p]$	$(7s4p1d)$	$[3s2p1d]$
def2-TZVP	$(5s1p)$	$[3s1p]$	$(11s3p)^d$	$[5s3p]^d$	$(11s6p2d1f)$	$[5s3p2d1f]$
def2-QZVP	$(7s3p2d1f)$	$[4s3p2d1f]$	$(15s6p2d1f)^e$	$[6s4p2d1f]^e$	$(15s8p3d2f1g)$	$[7s4p3d2f1g]$
cc-pVDZ	$(4s1p)$	$[2s1p]$	$(9s4p1d)$	$[3s2p1d]$	$(9s4p1d)$	$[3s2p1d]$
cc-pVTZ	$(5s2p1d)$	$[3s2p1d]$	$(11s5p2d1f)$	$[4s3p2d1f]$	$(10s5p2d1f)$	$[4s3p2d1f]$
cc-pVQZ	$(6s3p2d1f)$	$[4s3p2d1f]$	$(12s6p3d2f1g)$	$[5s4p3d2f1g]$	$(12s6p3d2f1g)$	$[5s4p3d2f1g]$
6-31G(d,p)	$(4s1p)$	$[2s1p]$	$(10s4p1d)$	$[3s2p1d]$	$(10s4p1d)$	$[3s2p1d]$
6-311G(2df,2pd)	$(5s2p1d)$	$[3s2p1d]$	$(11s5p2d1f)$	$[4s3p2d1f]$	$(11s5p2d1f)$	$[4s3p2d1f]$
STRLC	$(4s1p)^a$	$[2s1p]^a$	$(4s4p)$	$[2s2p]$	$(4s4p)^b$	$[2s2p]^b$
SBK	$(4s4p)$	$[3s2p]$	$(4s5p4d)$	$[2s3p2d]$	$(4s5p4d)$	$[2s3p2d]$

[a]Since there is no basis set for H, we used def2-SVP instead. [b]$(4s5p)\rightarrow[2s3p]$ for O and F. [c]$(7s4p)$ for Be. [d]$(11s4p1d)\rightarrow[5s3p1d]$ for Be. [e]$(15s7p2d1f)\rightarrow[7s4p2d1f]$ for Be.

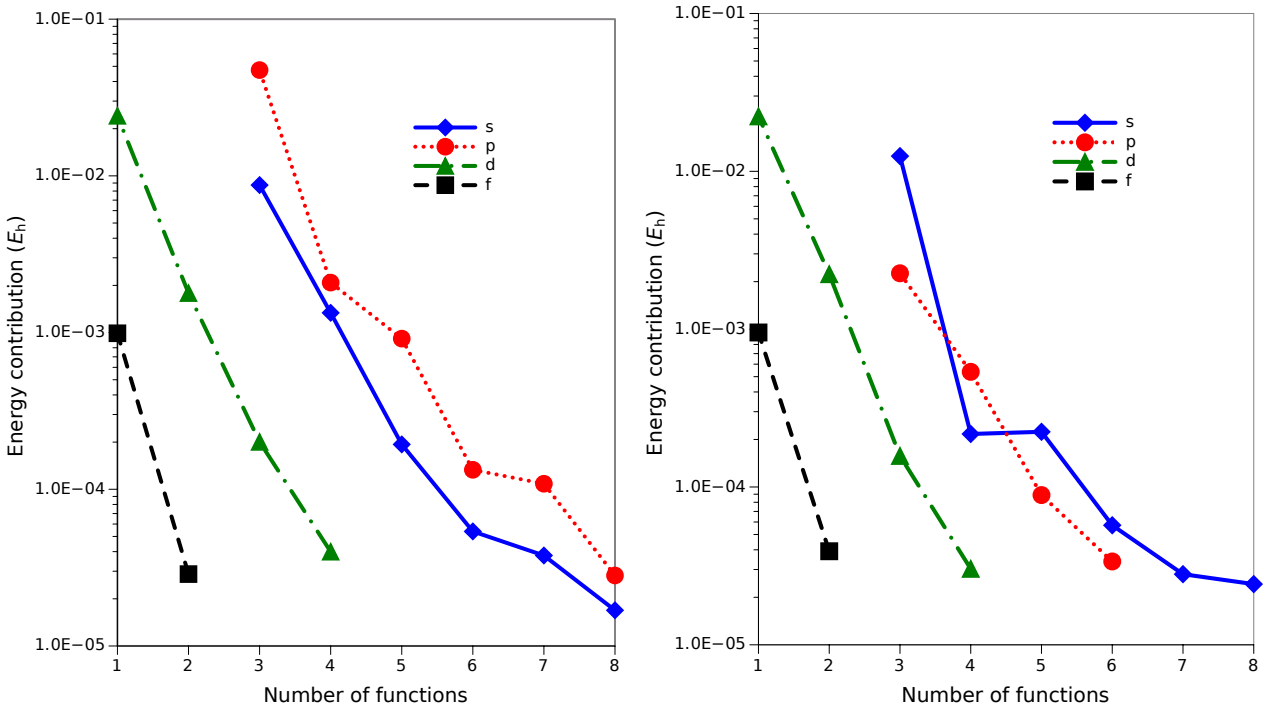

Fig. 5 Left plot: Energy contribution (per atom, in E_h) of s, p, d, and f functions in N. The latter two were calculated in N_2 with d(N–N) = 1.10 Å, Right plot: Energy contribution (per atom, in E_h) of s, p, d, and f functions in P. The latter two were calculated in P_2 with d(P–P) = 1.90 Å. The calculations were done with the SVWN functional.

Table 2 Compositions of third-period elements basis sets used in this work

	Na–Mg		Al–Cl	
PAW-L0.5	$(4s2p)$		$(2s2p1d)$	
PAW-L1	$(5s4p)$	$[4s2p]$	$(3s3p1d)$	$[2s2p1d]$
PAW-L2	$(6s5p1d)$	$[5s3p1d]$	$(4s4p2d1f)$	$[3s3p2d1f]$
pc-1	$(11s7p)$	$[4s2p]$	$(11s8p1d)$	$[4s3p1d]$
pc-2	$(13s9p1d)$	$[5s3p1d]$	$(13s10p2d1f)$	$[5s4p2d1f]$
pc-3	$(17s12p2d1f)$	$[6s4p2d1f]$	$(17s13p4d2f1g)$	$[6s5p4d2f1g]$
pc-4	$(21s15p3d2f1g)$	$[7s5p3d2f1g]$	$(21s16p6d3f2g1h)$	$[7s6p6d3f2g1h]$
def2-SVP	$(10s6p1d)^a$	$[4s2p1d]^a$	$(10s7p1d)$	$[4s3p1d]$
def2-TZVP	$(14s8p3d)$	$[5s4p3d]$	$(14s9p3d1f)$	$[5s5p2d1f]$
def2-QZVP	$(20s12p3d1f)^b$	$[9s5p3d1f]^b$	$(20s14p4d2f1g)$	$[9s6p4d2f1g]$
cc-pVDZ	$(12s8p1d)$	$[4s3p1d]$	$(12s8p1d)$	$[4s3p1d]$
cc-pVTZ	$(15s10p2d1f)^c$	$[5s4p2d1f]^c$	$(15s9p2d1f)$	$[5s4p2d1f]$
cc-pVQZ	$(16s12p3d2f1g)^d$	$[6s5p3d2f1g]^d$	$(16s11p3d2f1g)$	$[6s5p3d2f1g]$
6-31G(d,p)	$(16s10p1d)$	$[4s3p1d]$	$(16s10p1d)$	$[4s3p1d]$
6-311G(2df,2pd)	$(13s9p2d1f)$	$[6s5p2d1f]$	$(13s9p2d1f)^e$	$[6s5p2d1f]^e$
STRLC	$(4s4p)$	$[2s2p]$	$(4s4p)^f$	$[2s2p]^f$
SBK	$(4s5p4d)$	$[2s3p2d]$	$(4s5p4d)$	$[2s3p2d]$

[a] $(10s7p1d)\rightarrow[4s3p1d]$ for Mg. [b] $(20s12p4d1f)\rightarrow[9s5p4d1f]$ for Mg. [c] $(16s10p2d1f)\rightarrow[5s4p2d1f]$ for Na.
[d] $(19s12p3d2f1g)\rightarrow[6s5p3d2f1g]$ for Na. [e] $(13s10p2d1f)\rightarrow[6s5p2d1f]$ for Cl. [f] $(4s5p)\rightarrow[2s3p]$ for S and Cl.

SVWN functional and quadrature grids of size 7, implemented in the Turbomole v.7.4 program package [106].

We note that for the G2-97′ dataset containing small molecules (up to 14 atoms), we were able to perform geometry optimization calculations. This allows us to assess the performance of the basis sets in describing equilibrium structures. Because PAW-LDA energy gradients with respect to nuclear coordinates are not (yet) implemented, we optimized the structures using numerical gradients in combination with the Berny python package [107]. For the S22 dataset containing larger molecules (up to 30 atoms), we only performed single-point calculations.

Before discussing in detail the results, two important questions should be addressed: Which statistical metrics should be used to assess the performance of the basis sets? and more importantly, how to interpret these metrics? We can either use mean errors (MEs) and their standard deviation σ, as employed by Weigend and Ahlrichs [87] when designing the def2- basis set family; or mean absolute errors (MAEs) and maximum absolute errors (MaxAEs), used by Jensen [70] when benchmarking the pc-*n* basis sets. In both works, the "errors" were estimated using references calculated with either a large basis set [87] or the basis set limit [70]. Both approaches should lead to similar conclusions, even though the latter appears to give more penalty to basis sets with outliers, i.e. large MaxAEs. In this work, we employed the latter approach, additionally, we used error ranges, defined as the differences between the largest positive and negative error magnitudes. We seek for the best basis set with the smallest MAE, MaxAE, and error range values. Unfortunately, this is not a trivial task since often a basis set can have small MAE but very large MaxAE and error range and vice versa. In such cases, we prefer basis sets giving more consistent results (modest MAE, small MaxAE and error range) over those with slightly smaller MAE but large MaxAE and error range values. The reason for this favor is that, due to typically large intrinsic errors caused by DFAs, an "accurate" basis set close to the CBS limit does not necessarily give good results compared to experimental data.

5.1 Atomization Energies

In Figure 6, we plot the MAEs, MaxAEs, and error ranges of the atomization energies per atom calculated with different (contracted and uncontracted) basis sets. The atomization energies obtained with pc-4, which were shown to be close to the CBS limits (to within 0.02 kJ/mol) [70], are used as references. The results are also summarized in Table S3.

Since the atomic energies are independent of the polarization functions, the atomization energies (AE) are "unbalanced", i.e. the results are usually underestimated. Thus, it has been suggested that the "atomization energies", denoted as AE2, should be calculated relative to diatomic systems instead [70]. We found that AE2 errors are indeed smaller than the corresponding AE values. More importantly, both AE and AE2 results behave similarly (see Figure 6 and S2) and support our conclusion (*vide infra*). The AE2 results can be found in Table S3.

We first discuss the performance of the uncontracted double-ζ type basis sets (PAW-L1, pc-1, def2-SVP, cc-pVDZ, and 6-31G(d,p)). Overall, we found all basis sets provide practically similar performance. def2-SVP and 6-31G(d) give the smallest MAEs ($\sim$1.2 kcal/mol) and moderate MaxAEs (11.1 and 9.7 kcal/mol, respectively). PAW-L1, pc-1, and cc-pVDZ produce larger MAE values, 1.6, 1.9, and 1.7 kcal/mol, respectively. Among all the basis sets, PAW-L1 has the smallest MaxAE (8.4 kcal/mol) and error range (8.5 kcal/mol). The fact that the MaxAE and error range of PAW-L1 almost coincide is expected, as can also be seen with the pc-1 basis set. The reason is that the protocol of Jensen [69] tends to describe atomic energies better than molecular energies. Consequently, PAW-L1 and pc-1 *systematically* underestimate the atomization energies (244/255 and 255/255 molecules, respectively). The systematic behaviors of PAW-L1 and pc-1 are certainly beneficial, as the results will be always improved with larger basis sets (PAW-L2 and pc-2, *vide infra*). PAW-L1 should be favored over pc-1 because it has a smaller number of primitives and potentially higher computational efficiency.

Going from the uncontracted to contracted basis sets, the results are expected to deteriorate to some extent (Figure 6). The deterioration is characterized by contraction error, defined as the difference between the MAE values of the contracted and uncontracted basis sets. We found that in *all* molecules, contracted PAW-L1 and pc-1 further underestimate the atomization energies compared to the corresponding uncontracted basis sets. This is easy to clarify as the general contraction scheme of Raffenetti [86] only deteriorates molecular energies, whereas atomic energies are basically unchanged. The contraction errors of PAW-L1 and pc-1 are 0.74 and 0.39 kcal/mol, respectively. In contrast to PAW-L1 and pc-1, we observed that in many molecules, contracted def2-SVP, cc-pVDZ, and 6-31G(d,p) perform *better* than the corresponding uncontracted. Nevertheless, they still have positive contraction errors, 0.11, 0.46, and 0.20 kcal/mol, respectively.

Based on the MAE values, one can see that def2-SVP and 6-31G(d,p) perform the best with the smallest MAEs of 1.26 and 1.44 kcal/mol, respectively. They also give the smallest MaxAE value of 9.7 kcal/mol. The good performance of 6-31G(d,p) was also observed in ref. 70. The other three basis sets (PAW-L1, pc-1, and cc-pVDZ) perform slightly worse than def2-SVP and 6-31G(d,p). There is a small difference in performance between these three basis sets, with MAE values of 2.1–2.4 kcal/mol and MaxAE values of 10.8–12.6 kcal/mol. PAW-L1 appears to be a bit better than cc-pVDZ and pc-1, as it has smaller MaxAE (10.8 kcal/mol). Again, we found that PAW-L1 has the smallest error range of 10.8 kcal/mol (12.5 kcal/mol for pc-1 up to 14.4 kcal/mol for cc-pVDZ). Taking into consideration not only the results shown in Figure 6 but also the number of basis functions (see Tables 1 and 2), we argue that PAW-L1 should provide better performance in terms of accuracy and computational cost. It performs as well as the other basis sets but has a significantly smaller number of primitives, e.g. in carbon, the number of primitives is reduced by 30–40%, whereas in silicon the reduction is up to 50–70%.

In Figure 6, we also present the results calculated with the triple-ζ type basis sets (PAW-L2, pc-2, def2-TZVP, cc-pVTZ, and 6-311G(2df,2pd)). Here we only focus on the performance of the

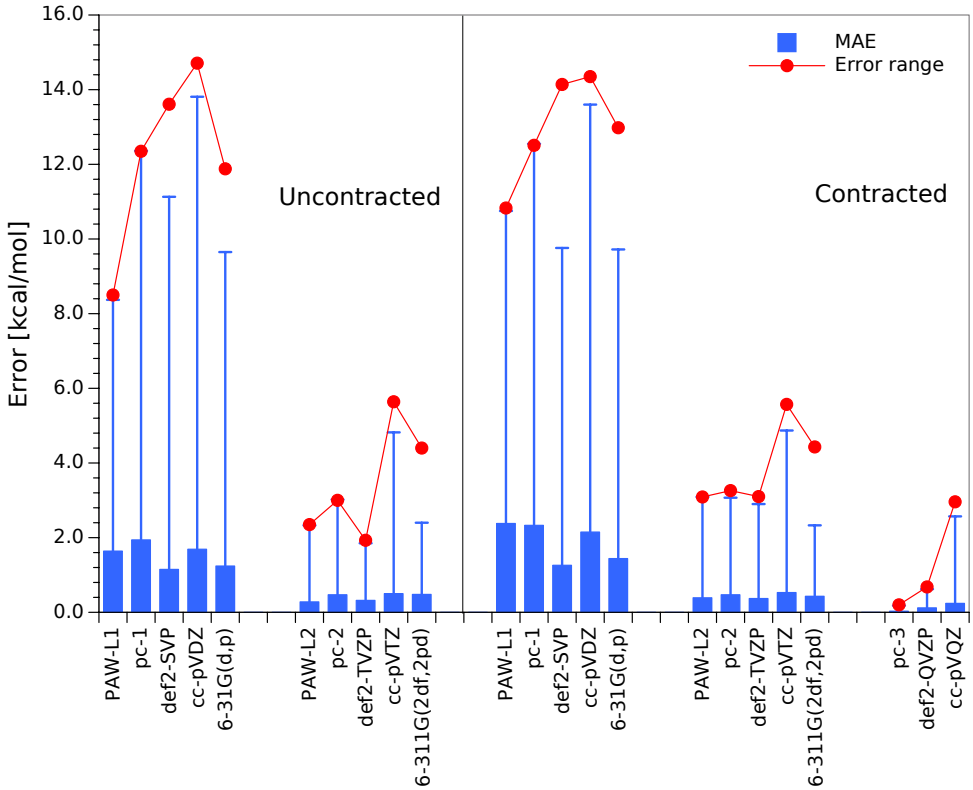

Fig. 6 Mean absolute errors (blue bars), maximum absolute errors (vertical blue lines), and error ranges (red circles) of the atomization energies per atom calculated with the PAW-Ln and different all-electron basis sets and the SVWN functional. The atomization energies calculated with pc-4 are used as references. Energies are in kcal/mol.

contracted basis sets. Based on the MAE values, we can conclude that all basis sets perform equally well, MAE is 0.39 kcal/mol (PAW-L2) up to 0.53 kcal/mol (cc-pVTZ). Based on the MaxAE and error range values, all basis sets, except for cc-pVTZ, give resembling performance. Among all five basis sets, PAW-L2 is apparently more attractive in terms of the accuracy–computational cost ratio. First, it inherits the appeal of PAW-L1, i.e. it *always* underestimates the atomization energies. Furthermore, it has smaller numbers of primitive and contracted basis functions. Using PAW-L2 instead of the equivalent triple-ζ basis sets can lead to 50% reduction in the number of primitives and 15% reduction in the number of contracted basis functions.

Finally, to examine whether PAW-L2 can be as accurate as larger basis sets containing higher angular momentum functions (up to g), we also performed calculations with pc-3, def2-QZVP, and cc-pVQZ (see Figure 6). The MAE values are small, $\sim$0.24 kcal/mol with cc-pVQZ and only 0.03 kcal/mol with pc-3. This shows that even though PAW-L2 provides distinct advantages over other equivalent all-electron basis sets, it cannot surpass the performance of basis sets containing higher polarization functions.

5.2 Vertical Ionization Energies

In Figure 7, we present the MAEs, MaxAEs, and error ranges of the vertical ionization energies calculated with the contracted PAW-Ln and all-electron basis sets. As compared to the all-electron double-ζ basis sets, PAW-L1 performs quite well. It has the smallest MAE value of 1.9 kcal/mol, a MaxAE value of only 8.6 kcal/mol, and a modest error range of 17 kcal/mol. The 6-31G(d,p) basis set, despite showing good performance in the atomization energies, performs the poorest in this case. It has the largest MAE value of 4.3 kcal/mol, a large MaxAE value of 15.2 kcal/mol, and an error range of up to 20 kcal/mol.

Among the triple-ζ basis sets, we found that the pc-2, def2-TZVP, cc-pVTZ, and PAW-L2 basis sets give comparable performance. The latter slightly outperforms the others, with the smallest MAE, MaxAE, and error range values of 0.4, 2.7, and 4.2 kcal/mol, respectively. To our surprise, the performance of PAW-L2 is even on a par with that of the cc-pVQZ basis set (MAE = 0.35 kcal/mol, MaxAE = 2.0 kcal/mol), however, still not as good as def2-QZVP (MAE = 0.19 kcal/mol) and especially pc-3 (MAE = 0.08 kcal/mol). Nonetheless, the results clearly demonstrate the appeal of the PAW-Ln basis sets, at least in the case of vertical ionization energies: (i) PAW-Ln, despite having smaller numbers of primitive and contracted basis functions, are as good as the other equivalent all-electron basis sets and (iii) PAW-L2 in some cases can be comparable to all-electron basis sets containing higher polarization functions.

5.3 Noncovalent Binding Energies

Finally, we assess the performance of PAW-Ln in the calculations of the S22 dataset noncovalent binding energies. This dataset comprises 22 complexes with different sizes and types of bonds

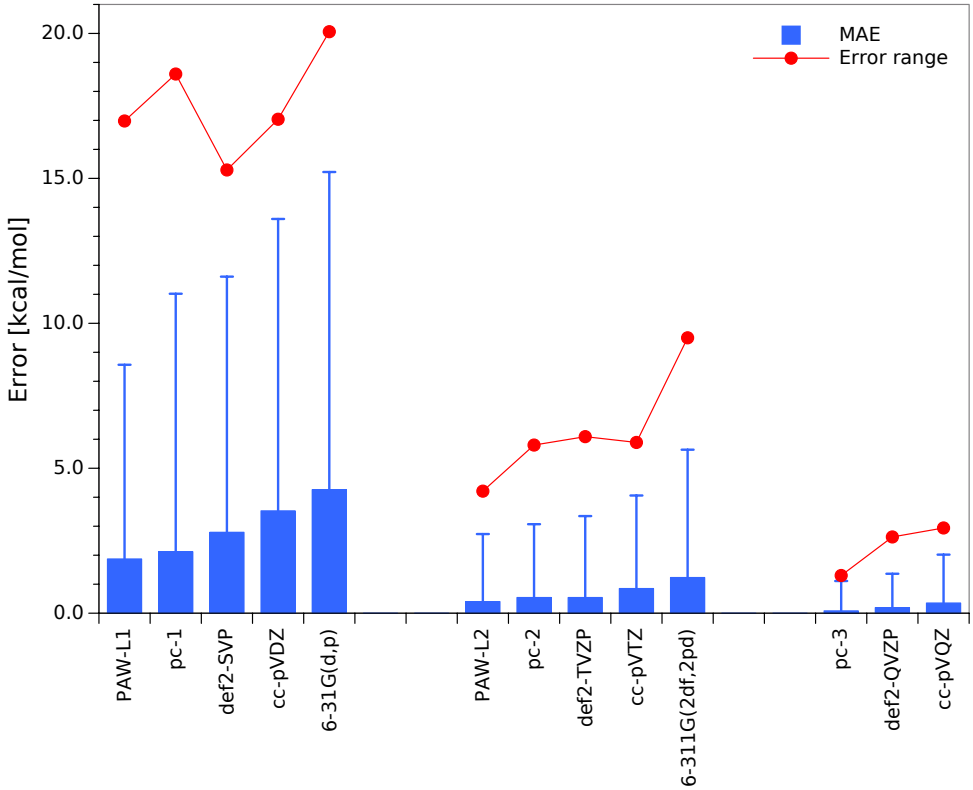

Fig. 7 Mean absolute errors (blue bars), maximum absolute errors (vertical blue lines), and error ranges (red circles) of the vertical ionization energies calculated with the contracted PAW-Ln and different all-electron basis sets and the SVWN functional. The ionization energies calculated with pc-4 are used as references. Energies are in kcal/mol. Results of boric acid, isobutane, trifluoroacetonitrile, and methylmagnesium chloride calculated with the PAW-Ln basis sets were not included because the calculations did not converge due to electronic degeneracy.

Table 3 MAEs, MaxAEs, and error ranges (in kcal/mol) of the binding energies of S22 complexes calculated with different contracted basis sets and SVWN functional. The binding energies calculated with pc-4 are used as references.

Basis set	MAE	MaxAE	Error range
PAW-L1	1.47	3.49	3.36
pc-1	2.00	4.34	4.30
def2-SVP	2.34	6.07	6.05
cc-pVDZ	2.03	5.82	5.79
6-31G(d,p)	1.97	4.89	4.80
aug-pc-1	1.81	4.61	4.35
def2-SVPD	1.86	4.69	4.36
aug-cc-pVDZ	0.74	1.87	1.71
6-31++G(d,p)	0.81	1.95	2.01
PAW-L2	0.17	0.44	0.48
pc-2	0.34	0.95	0.97
def2-TZVP	0.46	1.06	1.05
cc-pVTZ	0.77	1.60	1.59
6-311G(2df,2pd)	1.63	3.24	3.18
aug-pc-2	0.27	0.74	0.72
def2-TZVPD	0.23	0.73	0.91
aug-cc-pVTZ	0.17	0.44	0.43
6-311++G(2df,2pd)	0.50	1.46	1.46
pc-3	0.02	0.05	0.05
def2-QZVP	0.12	0.33	0.33
cc-pVQZ	0.36	0.78	0.78
aug-pc-3	0.02	0.05	0.06
def2-QZVPD	0.04	0.13	0.14
aug-cc-pVQZ	0.09	0.31	0.30
aug-pc-4[a]	0.00	0.01	0.01

[a] Only 18/22 complexes were calculated due to SCF convergence problem related to basis set linear dependencies.

(hydrogen, dispersion, and mixed bonds). It is widely accepted that diffuse functions play important roles in noncovalent interactions. Thus, in addition to the results calculated with the contracted basis sets listed in Tables 1 and 2, we also report the data obtained with basis sets augmented with diffuse functions. The results are shown in Table 3. For the aug-pc-4 basis set, only 18/22 complexes could be calculated as we encountered great difficulties in converging the SCF wavefunction (because of basis set linear dependencies). Thus, the values calculated with pc-4 are used as references, although aug-pc-4 basically gives the same results (to within 0.01 kcal/mol).

The results shown in Table 3 again highlight the good performance of the PAW-Ln basis sets. PAW-L1 is better than all double-ζ basis sets and even to some extent outperforms aug-pc-1 and def2-SVPD. Similarly, PAW-L2 is better to the other triple-ζ type basis sets. Interestingly, despite having a modest number of basis functions, PAW-L2 (MAE = 0.17 kcal/mol) can provide results close to def2-QZVP (MAE = 0.12 kcal/mol) and even outperforms cc-pVQZ (MAE = 0.36 kcal/mol).

6 Performance of PAW-Ln Compared to Effective Core Potential (ECP) Basis Sets

In light of the good performance (both accuracy and computational efficiency) of GTF-PAW in combination with PAW-Ln, it is also interesting to compare GTF-PAW to its alternative, the effective core potential (ECP) approach, which were generally constructed to reproduce as close as possible all-electron properties[108–112]. Unfortunately, ECP is traditionally employed for molecules containing heavy elements, e.g. transition metals, thus, there are few ECPs *and* Gaussian basis sets developed for light atoms. Furthermore, it has been shown that the accuracy of some popular ECPs can be disappointing, at least for geometries[113] and energetic properties[114] of some transition-metal complexes. In this part, we compare the performance of PAW-L1 with two popular ECP basis sets developed for light atoms: the Stuttgart relativistic large core ECP basis set, denoted as STRLC[108–110]; and the Stevens–Basch–Krauss ECP basis set, denoted as SBK[111,112]. The compositions of the ECP basis sets are shown in Tables 1 and 2. STRLC is smaller than PAW-L1, therefore, we did not expect that it can provide results better than PAW-L1. On the other hand, SBK has a significantly larger number of primitives than PAW-L1, allowing us to highlight the attractiveness of GTF-PAW. Of course, ECP results could be improved by using all-electron basis sets, e.g. pc-1[115,116] or employing new generations of ECP basis sets used for correlated many-body methods[38–43]. Unfortunately, the basis sets are still large, e.g. the double-ζ basis set of carbon in combination with the correlation consistent ECP of Bennett *et al.*[40] is (10s10p1d)→[2s2p1d], thus this defeats the purpose of reducing the computational cost. Since the ECPs were parameterized to implicitly describe relativistic effects, a direct comparison between ECP and PAW-L1 results might be inappropriate. However, we expect that in molecules containing light atoms the relativistic effects impact is minimal. Indeed, test calculations using the X2C scalar relativistic Hamiltonian[117] and the uncontracted pc-1 basis set confirmed that the relativistic effects negligibly change the atomization and ionization energies by ∼0.1 kcal/mol (see Table S3). We finally emphasize that all ECP basis sets employed in this work are "large-core", thus, large errors are expected.

The results of the atomization energy per atom, vertical ionization energy, and noncovalent binding energy, calculated with PAW-L1 as well as different ECP basis sets are summarized in Table 4. For comparison purposes, we also report the data calculated with 6-31G(d,p). The results calculated with the all-electron pc-4 basis set are used as references. The results clearly indicate that PAW-L1 is superior to the two ECP basis sets. For the atomization energies, all ECP calculations give very large MAE and MaxAE values. Two sources are attributed to the poor accuracy of these calculations: the intrinsic error of the ECPs and the error caused by limited valence basis sets. In the case of the SBK basis set, the former is expected to be large for the second-period elements whereas for the third-period elements, the latter is the dominant source of error[115]. While ECPs give very poor results for the atomization energies, they yield much better data for the ionization energies. The results are comparable to 6-31G(d,p), MAE = 5.7 kcal/mol. The numbers are still not as good as those

calculated with PAW-L1. For the noncovalent binding energies, the two ECP basis sets perform reasonably well (MAE = 2.18–3.10 kcal/mol) compared to PAW-L1 (MAE = 1.47 kcal/mol). This seems unexpected since these basis sets were not initially designed to describe noncovalent interactions. The remarkable performance of these basis sets is perhaps related to the fact that they contain some diffuse functions with exponents as small as 0.01.

In practice, ECPs are usually employed to quickly obtain equilibrium structures. Here, we evaluate the performance of PAW-L1 versus ECPs in describing not only equilibrium bond distances but also equilibrium angles and dihedrals. This can be done by calculating the root-mean-square deviation (RMSD) of the atomic positions (using pc-4 optimized structures as references) [118,119]. In Table 4 we report the mean and maximum of the RMSD values. PAW-L1 and 6-31G(d,p) give very similar structures, i.e. they yield almost the same mean RMSD values, 0.019 and 0.015 Å, respectively. ECPs perform slightly worse, nevertheless, the optimized structures are still in reasonable agreement with both PAW-L1 and 6-31G(d,p). The ECP results are in line with what were found in ref. 116.

7 Pushing the Performance to the Limit: The "Minimal" Basis Set

Considering the good performance of PAW-L1 in describing geometric and energetic properties, we explore whether an even smaller basis set can be constructed without drastically degrade the performance. As PAW-L1 is a double-ζ basis with polarization functions, we can recontract it to either a minimal basis, or a minimal basis with polarization functions, or a double-ζ basis. For instance with carbon, we can obtain $(3s3p)\rightarrow[1s1p]$, $(3s3p1d)\rightarrow[1s1p1d]$, or $(3s3p)\rightarrow[2s2p]$. Unfortunately, preliminary calculations for the atomization energy show that none of the reduced basis sets are able to provide results of sufficient quality. It is clear that double-ζ and polarization functions are prerequisites to accurately describe molecular bonding. Thus, a simple strategy is to design a double-ζ polarization basis set composed of only primitives: $(2s1p)$ for H, $(3s2p)$ for Li and Be, $(2s2p1d)$ for B–F and Al–Cl, $(4s2p)$ for Na and Mg. This basis set, denoted as PAW-L0.5, obviously has fewer primitives than PAW-L1, and drops the polarization consistent character. The MAE and MaxAE values of the atomization energies per atoms and ionization energies, as well as the RMSD of atomic positions can also be found in Table 4. First and foremost, PAW-L0.5 gives significantly better atomization energy results (MAE = 1.74 kcal/mol) than the ECP basis sets (MAE = 6–9 kcal/mol), and surprisingly, than PAW-L1 (MAE = 2.38 kcal/mol). This does not necessarily mean that PAW-L0.5 is more accurate than PAW-L1. In fact, because PAW-L0.5 drastically overestimates atomic energies, its good results are likely to be associated with error cancellations. For the ionization energies, the MAE values decrease in the order PAW-L1 > SBK > PAW-L0.5 > STRLC. We stress that the better performance of SBK compared to PAW-L0.5 is understandable given that SBK is 3–4 times larger than PAW-L0.5.

The results in Table 4 indicate that PAW-L0.5 is able to provide accurate equilibrium structures. The mean RMSD value is only 0.021 Å, similar to PAW-L1 (0.019 Å), and much smaller than the values calculated with ECPs (0.04–0.06 Å). This shows that PAW-L0.5 is certainly attractive, as one can obtain good equilibrium structures at a very cheap computational cost.

The compactness of PAW-L0.5 comes with some disadvantages. As compared to PAW-L1, PAW-L0.5 does not contain functions with a small exponent anymore, consequently, it is not able to describe molecular systems dominated by weak interactions, e.g. the S22 dataset. We found that for this dataset, PAW-L0.5 has a high MAE value (6.9 kcal/mol), much larger than PAW-L1 (1.5 kcal/mol).

8 Transferability of the PAW-L*n* Basis Sets: GGA Results

As the PAW-L*n* basis sets were optimized with the LDA functional, one can argue that a benchmark study at the LDA level of theory is biased towards these basis sets. In this section, we show that such bias is minimal by performing the same benchmark calculations but with the BLYP (GGA) functional [77,78,120]. The basis sets optimized at LDA level were reused whereas the PAW dataset was regenerated at the BLYP level. Our implementation of GTF-PAW has been recently extended to allow it to use the GGA functional [67]. The results are summarized in Figure 8 and Tables S4–S5.

We again observed the same trends found in the previous sections. PAW-L*n* are as good as the other equivalent all-electron basis sets in the descriptions of the atomization energy, ionization energy, and equilibrium structure. On the other hand, for noncovalent binding energies, PAW-L*n* give slightly better results than the other basis sets. PAW-L1 outperforms the ECP basis sets in all properties. The "minimal" basis set PAW-L0.5 again gives a small MAE value for the atomization energies (due to error cancellations) and is particularly suitable for the determination of equilibrium structures.

9 Conclusions

In this preliminary work, we introduced a new family of polarization consistent basis set (denoted as PAW-L*n*, $n = 1, 2$) that is specifically designed for GTF-PAW calculations. This family of basis set is our first attempt to expand the applicability of the GTF-PAW method. It resembles the one developed by Jensen [69], but provides some advantages such as a significantly smaller number of primitives, loose contraction, and comparable accuracy. Based on a benchmark study on the atomization energy, vertical ionization energy, and noncovalent binding energy, we firmly believe that qualitatively correct results are obtainable with the PAW-L1 basis set. The results are expected to be comparable to those calculated with double-ζ all-electron basis sets and certainly much better than ECP basis sets. For quantitative results that are not too far from the DFT basis set limit, PAW-L2 should be a good candidate. We also proposed a "minimal" basis set derived from PAW-L1, that can be useful to obtain equilibrium structures. Eventually, our results proved that GTF-PAW, in combination with small Gaussian basis sets is a promising approach for routine DFT applications.

Table 4 Mean and maximum absolute errors (in kcal/mol) of the atomization energies per atom, vertical ionization energies, and noncovalent binding energies; mean and maximum root-mean-square deviation (RMSD) of atomic positions (in Å), calculated with PAW-L1, PAW-L0.5, 6-31G(d,p), and different ECP basis sets and the SVWN functional. The values calculated with pc-4 are used as references.

	Atomization energies per atom		Ionization energies		Noncovalent binding energies		RMSD	
Basis set	MAE	MaxAE	MAE	MaxAE	MAE	MaxAE	Mean	Max
GTF-PAW								
PAW-L0.5	1.74	15.22	5.13	34.58	6.87	14.08	0.021	0.276
PAW-L1	2.38	10.75	1.87	8.57	1.47	3.49	0.019	0.374
ECP								
STRLC	9.37	50.97	7.02	34.50	2.18	6.29	0.064	0.485
SBK	6.19	35.78	4.03	17.19	3.10	7.08	0.041	0.484
All-electron								
6-31G(d,p)	1.44	9.72	5.7	15.2	1.97	4.89	0.015	0.215

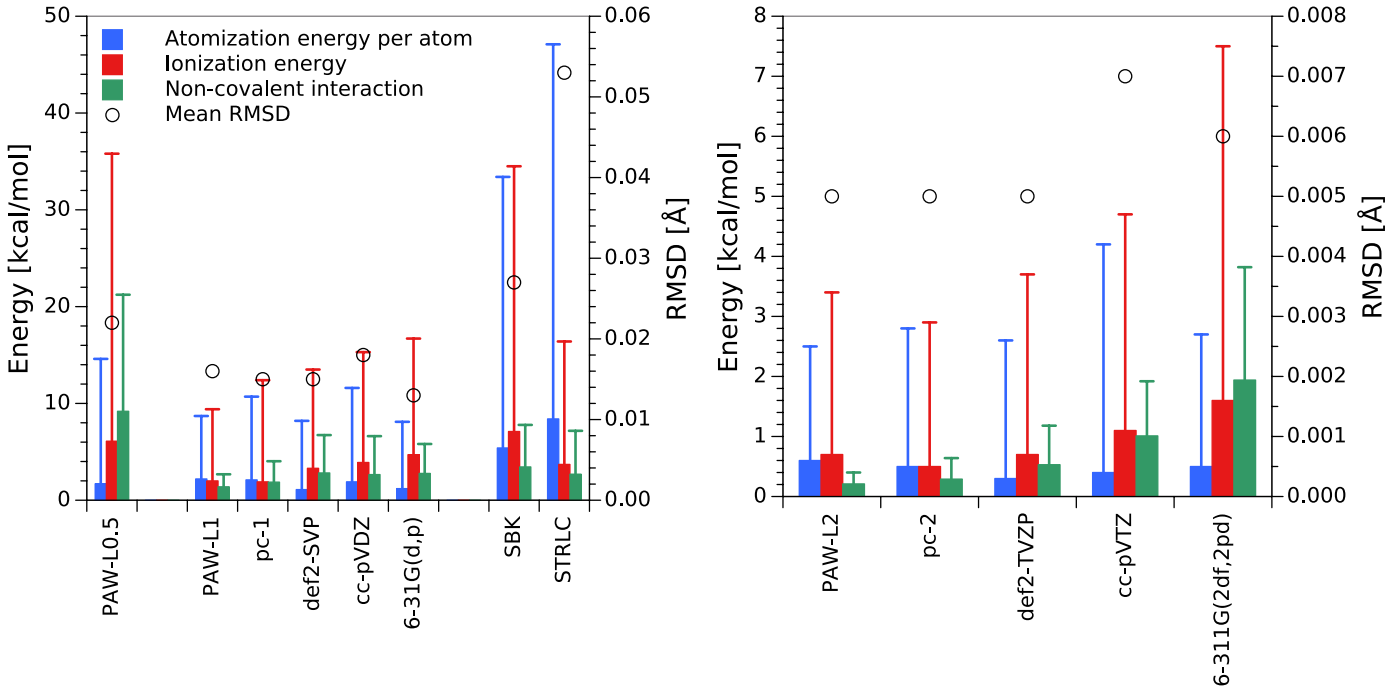

Fig. 8 Mean absolute errors (in kcal/mol) of the atomization energies per atom, vertical ionization energies, and noncovalent binding energies, calculated with PAW-L0.5, contracted PAW-Ln, all-electron, and ECP basis sets and the BLYP functional. The vertical lines represent the MaxAE values. The open circles represent the mean RMSD of the equilibrium structures. The values calculated with pc-4 are used as references. The ionization energies of boric acid calculated with the PAW-Ln basis sets were not included because the calculations did not converge due to electronic degeneracy. The potential energy curve of Mg_2 is too shallow, hence, the RMSD values of Mg_2 were not included.

The results presented in this work do *not* imply that our PAW-L*n* basis sets are always able to provide better results compared to other all-electron basis sets. Modifications of these all-electron basis sets will certainly improve the results. For instance, the cc-pV*n*Z basis sets can be augmented with additional tight *d* functions (cc-pV(n+d)Z) [121,122] that give better thermochemical properties of compounds containing third-period elements. These all-electron basis sets can also be reoptimized and recontracted at a better level of theory, e.g. BLYP and B3LYP for cc-pV*n*Z [123,124]. We are also aware that our basis sets in the current state have some limitations. Thus, this work can be continued in several directions. First and foremost, we will extend the PAW-L*n* basis sets to other elements of periods 4 and 5. The gain of computational efficiency should be more pronounced. The basis sets can be augmented with diffuse functions to correctly describe weakly bound electrons in anions and noncovalently bound systems. Also, we can reoptimize the basis sets at different levels of theory, such as HF or B3LYP. Such task is rather straightforward because the optimal compositions of the basis sets have been already determined and the optimization procedure is relatively fast and simple. The so-called P-orthogonalization method [98] can be employed in order to convert the general contracted basis sets to segmented contracted basis sets, which are expected to be more efficient. Finally, the smoothness of the PAW-L*n* basis with the GTF-PAW method should play a beneficial role in downsizing the auxiliary basis sets for the resolution of the identity (RI) approximation; thus, the development of the RI basis optimized for our orbital basis seems to be promising for accelerating the RI treatment.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

T.Y. is supported by JST, PRESTO Grant Number 17937609. X.-G. X. is supported by Fundamental Research Funds for the Central Universities Grant Number 19lgpy301 and Natural Science Foundation of China Grant Number 21501189. The computations were partly performed using Research Center for Computational Science, Okazaki, Japan.

Notes and references

1 A. D. Becke, *J. Chem. Phys.*, 2014, **140**, 18A301.
2 N. Mardirossian and M. Head-Gordon, *Mol. Phys.*, 2017, **115**, 2315–2372.
3 M. Arita, S. Arapan, D. R. Bowler and T. Miyazaki, *J. Adv. Simul. Sci. Eng.*, 2014, **1**, 87–97.
4 D. R. Bowler and T. Miyazaki, *J. Phys. Condens. Matter*, 2010, **22**, 074207.
5 E. Artacho, D. Sánchez-Portal, P. Ordejón, A. Garcia and J. M. Soler, *Phys. Status Solidi B*, 1999, **215**, 809–817.
6 C. F. Guerra, J. Snijders, G. t. te Velde and E. J. Baerends, *Theor. Chem. Acc.*, 1998, **99**, 391–403.
7 C.-K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne, *J. Chem. Phys.*, 2005, **122**, 084119.
8 C. Ochsenfeld, C. A. White and M. Head-Gordon, *J. Chem. Phys.*, 1998, **109**, 1663–1669.
9 J. Kussmann, M. Beer and C. Ochsenfeld, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2013, **3**, 614–636.
10 E. Schwegler and M. Challacombe, *J. Chem. Phys.*, 1996, **105**, 2726–2734.
11 F. Aquilante, T. B. Pedersen and R. Lindh, *J. Chem. Phys.*, 2007, **126**, 194106.
12 E. Rudberg, E. H. Rubensson and P. Sałek, *J. Chem. Phys.*, 2008, **128**, 184106.
13 J. C. Burant, G. E. Scuseria and M. J. Frisch, *J. Chem. Phys.*, 1996, **105**, 8969–8972.
14 G. E. Scuseria, *J. Phys. Chem. A*, 1999, **103**, 4782–4790.
15 X. He and K. M. Merz Jr, *J. Chem. Theory Comput.*, 2010, **6**, 405–411.
16 C. Ochsenfeld, *Chem. Phys. Lett.*, 2000, **327**, 216–223.
17 J. L. Whitten, *J. Chem. Phys.*, 1973, **58**, 4496–4501.
18 K. Eichkorn, F. Weigend, O. Treutler and R. Ahlrichs, *Theor. Chem. Acc.*, 1997, **97**, 119–124.
19 K. Eichkorn, O. Treutler, H. Öhm, M. Häser and R. Ahlrichs, *Chem. Phys. Lett.*, 1995, **240**, 283–290.
20 R. A. Kendall and H. A. Früchtl, *Theor. Chem. Acc.*, 1997, **97**, 158–163.
21 B. I. Dunlap, J. Connolly and J. Sabin, *J. Chem. Phys.*, 1979, **71**, 3396–3402.
22 F. Neese, *J. Comput. Chem.*, 2003, **24**, 1740–1747.
23 F. Neese, F. Wennmohs, A. Hansen and U. Becker, *Chem. Phys.*, 2009, **356**, 98–109.
24 C. Köppl and H.-J. Werner, *J. Chem. Theory Comput.*, 2016, **12**, 3122–3134.
25 D. Hamann, M. Schlüter and C. Chiang, *Phys. Rev. Lett.*, 1979, **43**, 1494.
26 L. Kleinman and D. Bylander, *Phys. Rev. Lett.*, 1982, **48**, 1425.
27 D. Vanderbilt, *Phys. Rev. B*, 1990, **41**, 7892.
28 G. Kresse and D. Joubert, *Phys. Rev. B*, 1999, **59**, 1758.
29 N. Troullier and J. L. Martins, *Phys. Rev. B*, 1991, **43**, 1993.
30 S. Goedecker, M. Teter and J. Hutter, *Phys. Rev. B*, 1996, **54**, 1703.
31 M. Dolg, H. Stoll, A. Savin and H. Preuss, *Theor. Chim. Acta*, 1989, **75**, 173–194.
32 D. Andrae, U. Haeussermann, M. Dolg, H. Stoll and H. Preuss, *Theor. Chim. Acta*, 1990, **77**, 123–141.
33 M. Dolg, U. Wedig, H. Stoll and H. Preuss, *J. Chem. Phys.*, 1987, **86**, 866–872.
34 W. R. Wadt and P. J. Hay, *J. Chem. Phys.*, 1985, **82**, 284–298.
35 P. J. Hay and W. R. Wadt, *J. Chem. Phys.*, 1985, **82**, 270–283.
36 P. J. Hay and W. R. Wadt, *J. Chem. Phys.*, 1985, **82**, 299–310.
37 J. Trail and R. Needs, *J. Chem. Phys.*, 2005, **122**, 174109.
38 J. Trail and R. Needs, *J. Chem. Phys.*, 2013, **139**, 014101.
39 J. R. Trail and R. J. Needs, *J. Chem. Phys.*, 2017, **146**, 204107.
40 M. C. Bennett, C. A. Melton, A. Annaberdiyev, G. Wang, L. Shulenburger and L. Mitas, *J. Chem. Phys.*, 2017, **147**,

224106.
41 M. C. Bennett, G. Wang, A. Annaberdiyev, C. A. Melton, L. Shulenburger and L. Mitas, *The Journal of chemical physics*, 2018, **149**, 104108.
42 A. Annaberdiyev, G. Wang, C. A. Melton, M. C. Bennett, L. Shulenburger and L. Mitas, *J. Chem. Phys.*, 2018, **149**, 134108.
43 G. Wang, A. Annaberdiyev, C. A. Melton, M. C. Bennett, L. Shulenburger and L. Mitas, *J. Chem. Phys.*, 2019, **151**, 144110.
44 J. C. Slater, *Phys. Rev.*, 1937, **51**, 846–851.
45 O. K. Andersen, *Phys. Rev. B*, 1975, **12**, 3060–3083.
46 P. E. Blöchl, *Phys. Rev. B*, 1994, **50**, 17953.
47 T. Charpentier, *Solid State Nucl. Magn. Reson.*, 2011, **40**, 1–20.
48 G. Kresse and J. Furthmüller, *Phys. Rev. B*, 1996, **54**, 11169.
49 X. Gonze, B. Amadon, G. Antonius, F. Arnardi, L. Baguet, J.-M. Beuken, J. Bieder, F. Bottin, J. Bouchet, E. Bousquet *et al.*, *Comput. Phys. Commun.*, 2020, **248**, 107042.
50 S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. Probert, K. Refson and M. C. Payne, *Z. Kristallogr. Cryst. Mater.*, 2005, **220**, 567–570.
51 P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo *et al.*, *J. Phys. Condens. Matter*, 2009, **21**, 395502.
52 J. e. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. Hansen *et al.*, *J. Phys. Condens. Matter*, 2010, **22**, 253202.
53 T. Rangel, D. Caliste, L. Genovese and M. Torrent, *Comput. Phys. Commun.*, 2016, **208**, 1–8.
54 S. Kang, S. Ryu, S. Choi, J. Kim, K. Hong and W. Y. Kim, *International Journal of Quantum Chemistry*, 2016, **116**, 644–650.
55 M. Dolg *et al.*, *Modern methods and algorithms of quantum chemistry*, 2000, **3**, 507–540.
56 J. C. Phillips and L. Kleinman, *Phys. Rev.*, 1959, **116**, 287–294.
57 M. L. Cohen and T. K. Bergstresser, *Phys. Rev.*, 1966, **141**, 789–796.
58 D. R. Hamann, M. Schlüter and C. Chiang, *Phys. Rev. Lett.*, 1979, **43**, 1494–1497.
59 D. R. Hamann, *Phys. Rev. B*, 2013, **88**, 085117.
60 S. Goedecker, M. Teter and J. Hutter, *Phys. Rev. B*, 1996, **54**, 1703–1710.
61 D. Vanderbilt, *Phys. Rev. B*, 1990, **41**, 7892–7895.
62 X.-G. Xiong and T. Yanai, *J. Chem. Theory Comput.*, 2017, **13**, 3236–3249.
63 M. Torrent, F. Jollet, F. Bottin, G. Zérah and X. Gonze, *Comp. Mat. Sci.*, 2008, **42**, 337–351.
64 M. Torrent, N. Holzwarth, F. Jollet, D. Harris, N. Lepley and X. Xu, *Comp. Phys. Comm.*, 2010, **181**, 1862–1867.
65 N. Holzwarth, A. Tackett and G. Matthews, *Comp. Phys. Comm.*, 2001, **135**, 329–347.
66 F. Jollet, M. Torrent and N. Holzwarth, *Comput. Phys. Commun.*, 2014, **185**, 1246–1254.
67 X.-G. Xiong, A. Sugiura and T. Yanai, *Journal of Chemical Theory and Computation*, 2020, **16**, 4883–4898.
68 A. D. Becke, *J. Chem. Phys.*, 1988, **88**, 2547–2553.
69 F. Jensen, *J. Chem. Phys.*, 2001, **115**, 9113–9125.
70 F. Jensen, *J. Chem. Phys.*, 2002, **116**, 7372–7379.
71 F. Jensen, *J. Chem. Phys.*, 2002, **117**, 9234–9240.
72 F. Jensen and T. Helgaker, *J. Chem. Phys.*, 2004, **121**, 3463–3470.
73 F. Jensen, *J. Phys. Chem. A*, 2007, **111**, 11198–11204.
74 K. F. Garrity, J. W. Bennett, K. M. Rabe and D. Vanderbilt, *Comput. Mater. Sci.*, 2014, **81**, 446–452.
75 A. Dal Corso, *Comput. Mater. Sci.*, 2014, **95**, 337–350.
76 T. Yanai, Y. Kurashige, W. Mizukami, J. Chalupský, T. N. Lan and M. Saitow, *Int. J. Quantum Chem.*, 2015, **115**, 283–299.
77 P. A. M. Dirac, *Proc. R. Soc. Lond.*, 1929, **123**, 714–733.
78 J. C. Slater, *Phys. Rev.*, 1951, **81**, 385.
79 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200–1211.
80 *NIST Computational Chemistry Comparison and Benchmark Database*, `http://cccbdb.nist.gov/`, (Release 20, August 2019).
81 X. Gonze, P. Käckell and M. Scheffler, *Phys. Rev. B*, 1990, **41**, 12264.
82 F. Neese, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2018, **8**, e1327.
83 F. Jensen, *Chem. Phys. Lett.*, 2005, **402**, 510–513.
84 R. D. Bardo and K. Ruedenberg, *J. Chem. Phys.*, 1974, **60**, 918–931.
85 F. Jensen, *J. Phys. Chem. A*, 2017, **121**, 6104–6107.
86 R. C. Raffenetti, *J. Chem. Phys.*, 1973, **58**, 4452–4458.
87 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
88 T. H. Dunning Jr, *J. Chem. Phys.*, 1989, **90**, 1007–1023.
89 B. P. Prascher, D. E. Woon, K. A. Peterson, T. H. Dunning and A. K. Wilson, *Theor. Chem. Acc.*, 2011, **128**, 69–82.
90 D. E. Woon and T. H. Dunning Jr, *J. Chem. Phys.*, 1993, **98**, 1358–1371.
91 J. S. Binkley and J. A. Pople, *J. Chem. Phys.*, 1977, **66**, 879–880.
92 J. D. Dill and J. A. Pople, *J. Chem. Phys.*, 1975, **62**, 2921–2923.
93 R. Ditchfield, W. J. Hehre and J. A. Pople, *J. Chem. Phys.*, 1971, **54**, 724–728.
94 M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. DeFrees and J. A. Pople, *J. Chem. Phys.*, 1982, **77**, 3654–3665.
95 M. S. Gordon, J. S. Binkley, J. A. Pople, W. J. Pietro and W. J. Hehre, *J. Am. Chem. Soc.*, 1982, **104**, 2797–2803.
96 P. C. Hariharan and J. A. Pople, *Theor. Chim. Acta*, 1973, **28**, 213–222.
97 W. J. Hehre, R. Ditchfield and J. A. Pople, *J. Chem. Phys.*,

1972, **56**, 2257–2261.

98 F. Jensen, *J. Chem. Theory Comput.*, 2014, **10**, 1074–1085.

99 B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson and T. L. Windus, *J. Chem. Inf. Model.*, 2019, **59**, 4814–4820.

100 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16 Revision C.01*, 2016, Gaussian Inc. Wallingford CT.

101 A. McLean and G. Chandler, *J. Chem. Phys.*, 1980, **72**, 5639–5648.

102 R. Krishnan, J. S. Binkley, R. Seeger and J. A. Pople, *J. Chem. Phys.*, 1980, **72**, 650–654.

103 L. A. Curtiss, K. Raghavachari, P. C. Redfern and J. A. Pople, *J. Chem. Phys.*, 1997, **106**, 1063–1079.

104 L. A. Curtiss, P. C. Redfern, K. Raghavachari and J. A. Pople, *J. Chem. Phys.*, 1998, **109**, 42–55.

105 P. Jurečka, J. Šponer, J. Černỳ and P. Hobza, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1985–1993.

106 R. Ahlrichs, M. Bär, M. Häser, H. Horn and C. Kölmel, *Chem. Phys. Lett.*, 1989, **162**, 165–169.

107 J. Hermann, *Berny - Molecular Optimizer*, `https://github.com/jhrmnn/pyberny`, 2020.

108 A. Bergner, M. Dolg, W. Küchle, H. Stoll and H. Preuß, *Mol. Phys.*, 1993, **80**, 1431–1441.

109 P. Fuentealba, H. Preuss, H. Stoll and L. Von Szentpály, *Chem. Phys. Lett.*, 1982, **89**, 418–422.

110 P. Fuentealba, L. Von Szentpaly, H. Preuss and H. Stoll, *J. Phys. B*, 1985, **18**, 1287.

111 N. P. Labello, A. M. Ferreira and H. A. Kurtz, *Int. J. Quantum Chem.*, 2006, **106**, 3140–3148.

112 W. J. Stevens, H. Basch and M. Krauss, *J. Chem. Phys.*, 1984, **81**, 6026–6033.

113 M. P. Waller, H. Braun, N. Hojdis and M. Bühl, *J. Chem. Theory Comput.*, 2007, **3**, 2234–2242.

114 X. Xu and D. G. Truhlar, *J. Chem. Theory Comput.*, 2012, **8**, 80–90.

115 C. Janfelt and F. Jensen, *Chem. Phys. Lett.*, 2005, **406**, 501–503.

116 F. Jensen and C. Janfelt, *Chem. Phys. Lett.*, 2005, **412**, 12–15.

117 D. Peng and M. Reiher, *Theor. Chem. Acc.*, 2012, **131**, 1081.

118 J. C. Kromann, *Calculate Root-Mean-Square Deviation (RMSD) of Two Molecules Using Rotation*, `http://github.com/charnley/rmsd`, 2020.

119 W. Kabsch, *Acta Crystallogr. A*, 1976, **32**, 922–923.

120 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.

121 T. H. Dunning Jr, K. A. Peterson and A. K. Wilson, *J. Chem. Phys.*, 2001, **114**, 9244–9253.

122 N. X. Wang and A. K. Wilson, *J. Phys. Chem. A*, 2005, **109**, 7187–7196.

123 B. P. Prascher and A. K. Wilson, *Mol. Phys.*, 2007, **105**, 2899–2917.

124 A. Mahler, J. J. Determan and A. K. Wilson, *J. Chem. Phys.*, 2019, **151**, 064110.