Supporting Information for Extending Density-Corrected Density Functional Theory to Large Molecular Systems

Youngsam Kim,[a,†] Mingyu Sim,[a,†] Minhyeok Lee,[a] Sehun Kim,[a] Suhwan Song,[a,‡]
Kieron Burke,[b] and Eunji Sim[a,*]

[a]Department of Chemistry, Yonsei University, 50 Yonsei-ro Seodaemun-gu, Seoul 03722, Korea
[b]Department of Chemistry, University of California, Irvine, CA 92697, USA

Contents

PySCF Code for Dual-Basis HF-DFT s2

Figure S1 L13 (L7+S6L) database structures s3

Figure S2 C_{60} relative total energy and corresponding wall-time for HF-r^{2}SCAN with different basis sets s4

Figure S3 Relative HF total energy (He atom) and the HF isomerization energies (C_{60}) s4

Figure S4 MABD for Bauza30 .... s5

Figure S5 MABD of D^{2}C-PBE and D^{2}C-B3LYP on GMTKN55 database s5

Figure S6 MAD of D^{2}C-PBE, -r^{2}SCAN, and -B3LYP with def2-QZVPPD on GMTKN55 database s6

Figure S7 Numerical problem in dual-basis method for a single case in GMTKN55 database s6

Figure S8 Basis set dependence in G21EA dataset s7

Figure S9 Comparison between the absolute reference energy and the absolute basis deviation on MB16-43 dataset .. s7

Figure S10 Absolute errors relative to the CBS limit on S66 dataset s8

Table S1 Number of basis functions in def2-series for C_{60} s9

Table S2 Mean absolute reference energy and corresponding weights for THERMO, LARGE, and BARRIER in GMTKN55 database s10

Table S3 Mean absolute reference energy and corresponding weights for INTERMOL and CONFOR in GMTKN55 database s11

Table S4 Numerical problem in dual-basis method for a single case in GMTKN55 .... s11

Table S5 WTMAD-2 for DFT, DFT-D4, and D^{2}C-DFT s12

Table S6 Averaged DC4 corrections for HF-r^{2}SCAN in small and large non-covalently bound systems s12

Table S7 Optimal XDM parameters for HF-r^{2}SCAN with respect to the def2-QZVPPD basis set s12

[*] esim@yonsei.ac.kr

[†] equal contribution

[‡] current address; Ai Research Center, SAMSUNG Advanced Institute of Technology, Suwon 16678, Korea

PySCF Code for Dual-Basis HF-DFT

A simple example of PySCF code for performing dual-basis HF-DFT.

```
from pyscf import gto, scf, dft, lib
from scipy.linalg import eigh

b1 = ’def2-svpd’ ## primary basis set (Pri.)
b2 = ’def2-qzvpd’ ## secondary basis set (Sec.)
coord = ’H 0 0 0; H 0 0 0.9’

### Pri. SCF ###
mol1 = gto.M(
    atom = coord,
    verbose = 6,
    basis = b1,
    output = f’./simple.output’,
)
myhf = scf.RHF(mol1)
myhf.max_cycle = 300
Esc = myhf.kernel()
print(f’E_SCF(HF/{b1}) {Esc:.5f}’)

Dsc = myhf.make_rdm1() ## Pri. density matrix
################

### dual-basis HF (HF/Pri./Sec.) ###
mol2 = gto.M(
      atom = coord,
      verbose = 0,
      basis = b2,
      )

mf = scf.RHF(mol2)
Nocc     = mol2.nelectron//2
S        = mf.get_ovlp()
H        = mf.get_hcore()
J        = scf.jk.get_jk((mol2,mol2,mol1,mol1),Dsc,scripts=’ijkl,lk->ij’,aosym=’s4’)
K        = scf.jk.get_jk((mol2,mol1,mol1,mol2),Dsc,scripts=’ijkl,jk->il’)
F        = H + J - .5*K
e,C      = eigh(F,S)
Dhfpc    = 2*lib.einsum(’ik,jk->ij’,C[:,:Nocc],C[:,:Nocc],optimize=True) ## density matrix Pri./Sec.
############

### dual-basis HF-DFT ###
mydft = dft.RKS(mol2)
mydft.xc  = ’r2scan’
E_dbhfdft = mydft.energy_tot(Dhfpc)
print(f’HF-r2SCAN with {b1}-{b2} {E_dbhfdft:.5f}’)
```

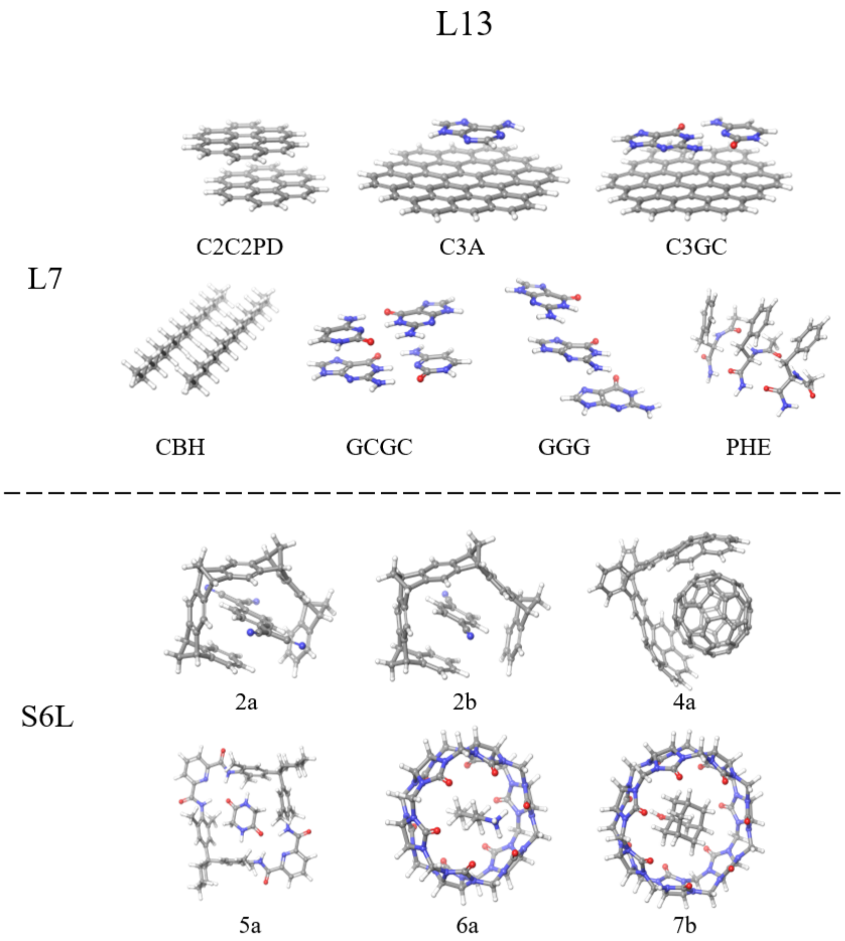

Figure S1: L13 database: Structures of the 7 complexes included in the L7 dataset and 6 supramolecular complexes included in the S6L dataset. The geometries of L7 and S6L are from Refs. [1] and [2], and reference energies are taken from Ref. [3]. The database is referred to as L13.

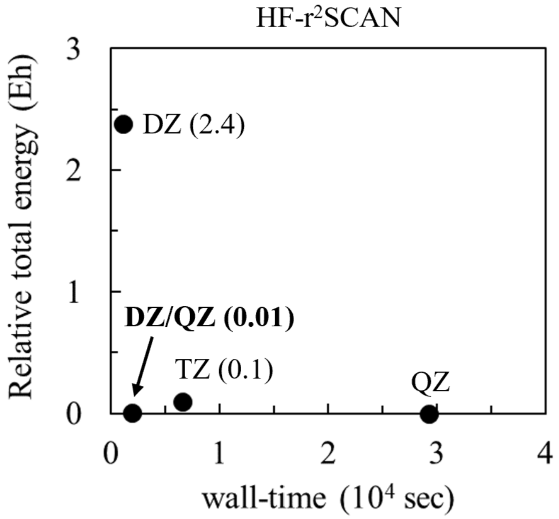

Figure S2: C_{60} relative total energy and corresponding wall-time for HF-r^{2}SCAN with different basis sets. The total energy of QZ is set to 0, and the relative energy is represented in parentheses. The DZ, TZ, and QZ represents a double (def2-SVPD), triple (def2-TZVPPD), and quadruple (def2-QZVPPD) zeta basis sets, respectively. Wall times were measured on 64 cores of an Intel(R) Xeon(R) CPU Platinum 8358 @ 2.60GHz.

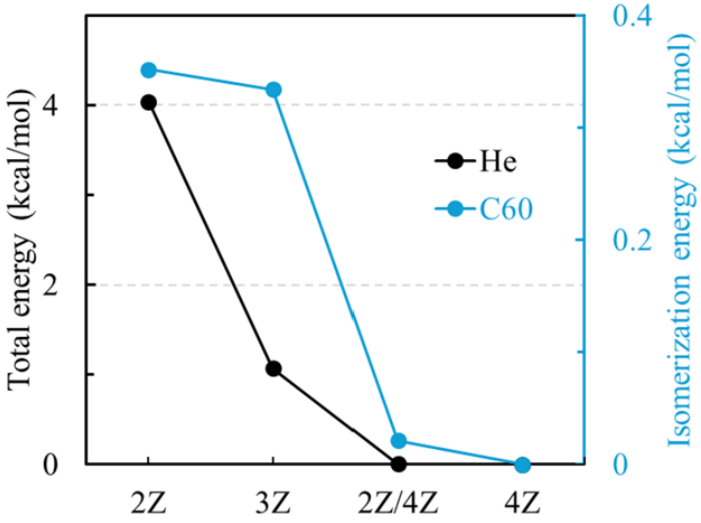

Figure S3: The HF total energy of the He atom (black) and the HF isomerization energies of the C_{60} molecule (blue) relative to HF with 4Z. The 2Z, 3Z, and 4Z correspond to def2-SVPD, def2-TZVPPD, and def2-QZVPPD basis sets, respectively. The 2Z/4Z denotes a dual-basis result with 2Z as the primary and 4Z as the secondary basis sets.

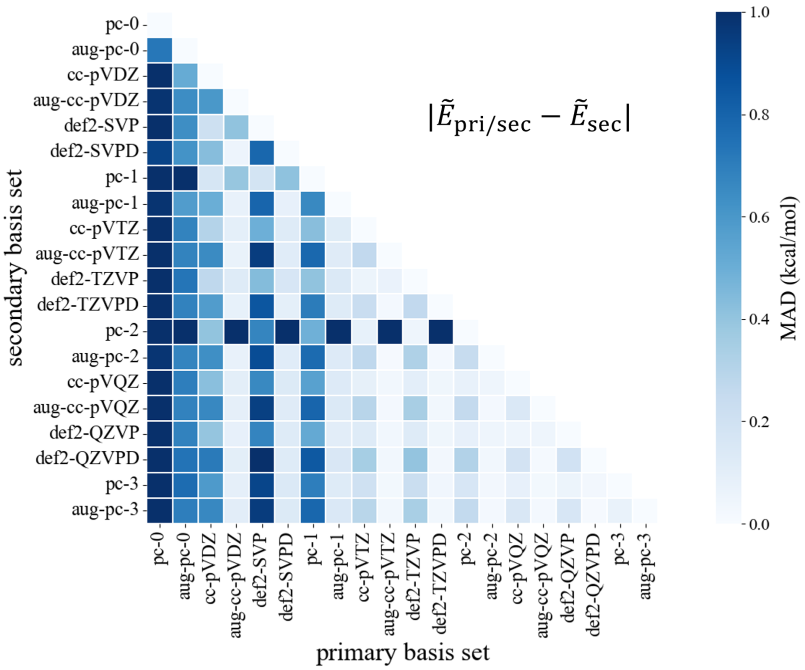

Figure S4: Mean absolute basis deviation (MABD) from energies calculated with secondary basis set. The binding energies were computed using D^{2}C-r^{2}SCAN with various basis set pairings for the Bauza30[4] dataset. Lighter colors indicate less deviation.

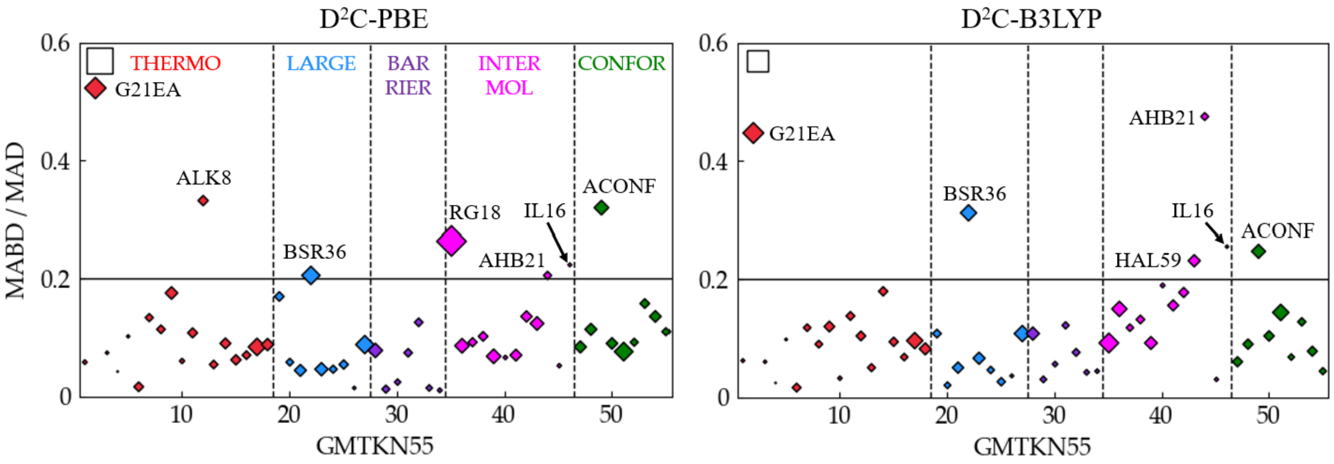

Figure S5: The ratio of MABD (2Z/4Z) to MAD (4Z) with D^{2}C-PBE (left) and D^{2}C-B3LYP (right) where 2Z and 4Z are def2-SVPD and def2-QZVPPD, respectively. Marker size is proportional to the weighted MABD for each datasets. The solid horizontal line indicates an arbitrary value of 0.2 used as a baseline for outliers. The WTMAD-2 for each D^{2}C-DFT is shown with the square in the upper left corner denoting 6.53 kcal/mol for D^{2}C-PBE and 4.67 kcal/mol for D^{2}C-B3LYP. The sub-datasets are ordered as in Ref. [5] and classified according to chemical properties as in Ref. [6]. The few datasets that exceed the 20% line exhibit either have MABDs much smaller than WTMAD-2 or have 4Z MADs less than 1 kcal/mol. This indicates that an extremely small MAD can result in an unusually large ratio.

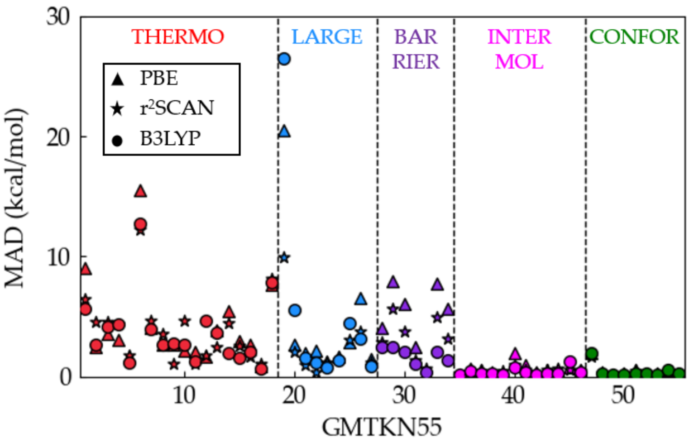

Figure S6: The MAD of D^{2}C-DFTs/def2-QZVPPD on the GMTKN55 database. D^{2}C-PBE, -r^{2}SCAN, and -B3LYP are represented by diamonds, asterisks, and circles, respectively. The datasets are categorized and colored in the same manner as illustrated in Fig. S5.

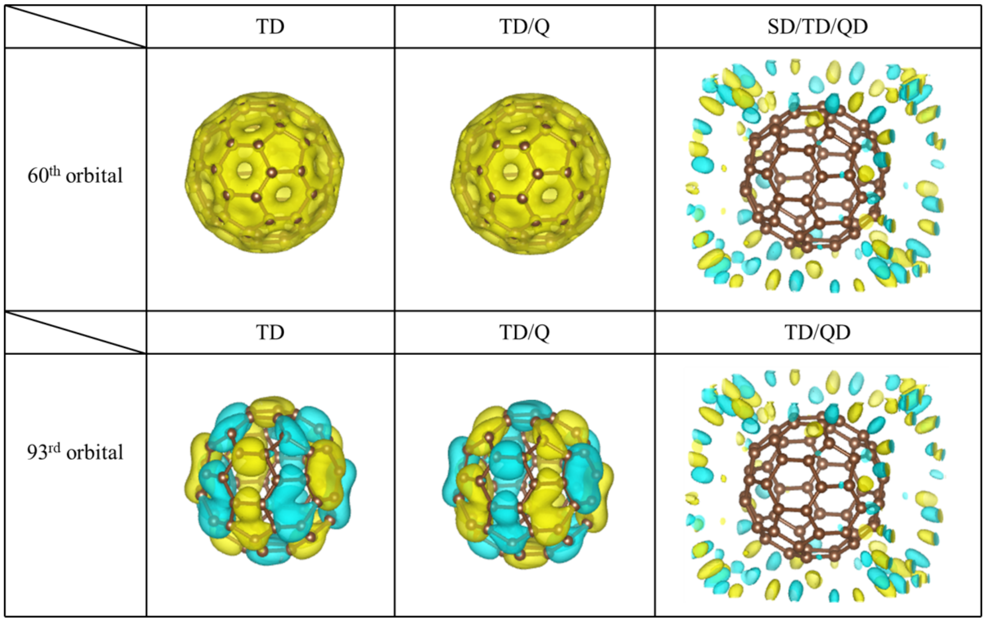

Figure S7: The orbitals that exhibit problematic behavior as obtained through diagonalization within the dual-basis method. TD is the self-consistent HF results with def2-TZVPD basis set. TD/Q is the dual-basis HF result with def2-TZVPD as a primary and def2-QZVP as a secondary basis set. Similarly, TD/QD is def2-TZVPD/def2-QZVPD. TD/Q/QD is evaluated in two steps. Initially, the dual-basis TD/Q is calculated. Subsequently, the density matrix of TD/Q is assumed to be the result of the primary basis set, and def2-QZVPD basis set is utilized as a secondary basis set. The same process is employed for the evaluation of SV/TD/QD. Similar numerical issues are observed in both TD/QD and SD/TD/QD.

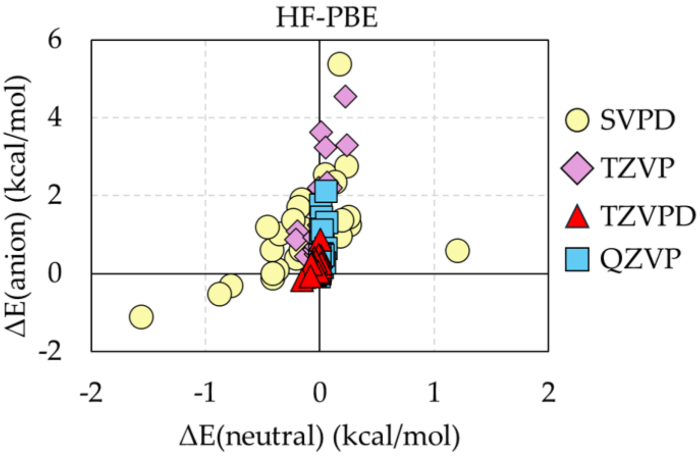

Figure S8: In the G21EA dataset, the total energies difference between HF-PBE/def2-QZVPD and the dual-basis HF-PBE where the secondary basis set is def2-QZVPD and each marker represents a primary basis set: def2-SVPD (SVPD), def2-TZVP (TZVP), def2-TZVPD (TZVPD), and def2-QZVP (QZVP).

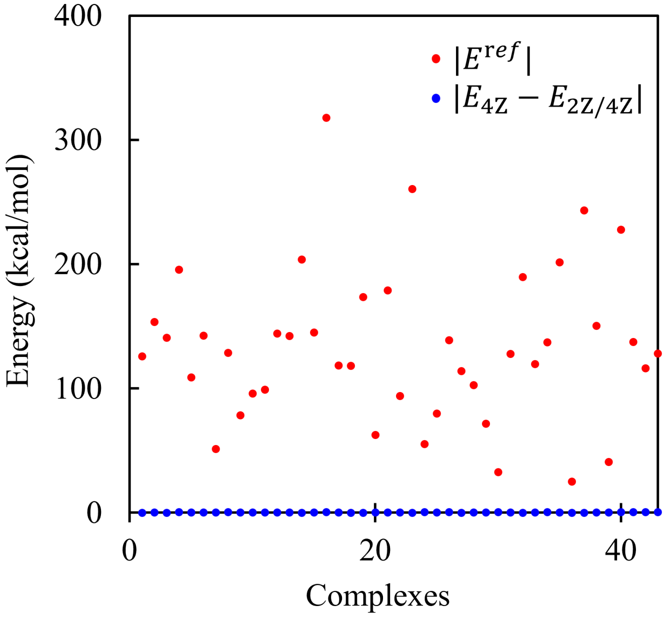

Figure S9: The absolute reference energy for the MB16-43 dataset (red) and the absolute basis deviation of D^2C-r^{2}SCAN (blue) are presented.

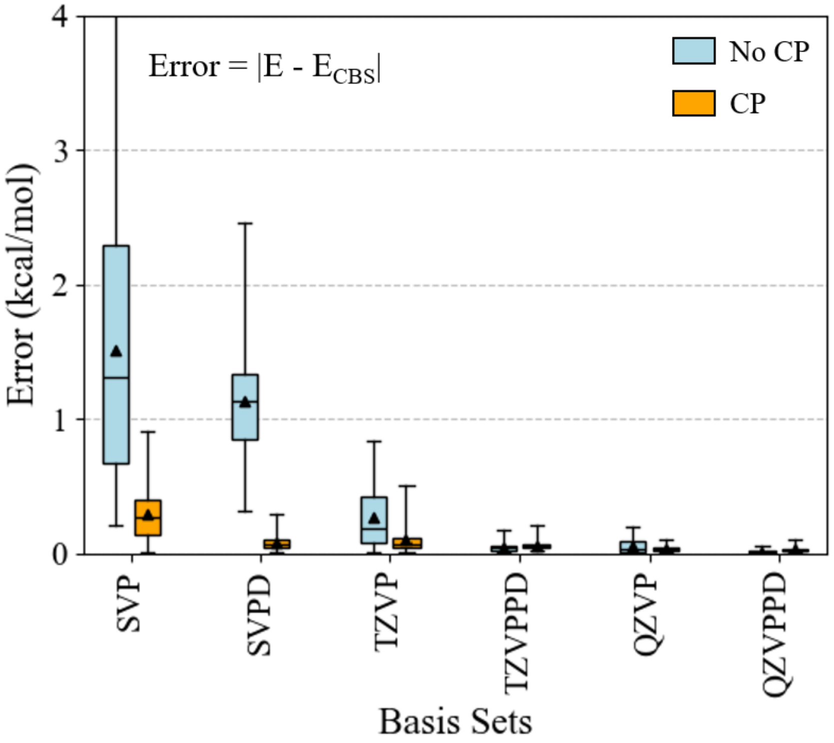

Figure S10: Absolute errors relative to the CBS limit for each basis set in the S66 dataset. The CBS limits are the half CP corrected aug-cc-PVQZ results. Blue denotes cases without CP correction, and orange denotes cases with CP correction. The black triangle and vertical line signify the mean absolute error and median values. The def2-series of basis set are used and 'def2-' is omitted for simplicity. Calculations have been performed with the Pyscf v2.3.0, numerical grid level 6. The convergence threshold has been set to 1e-8 a.u.

Table S1: Number of basis functions (contracted Gaussian type orbital, CGTO) in Ahlrichs def2- series[7] for C_{60}. For carbon, the basis functions of def2-TZVPD (def2-QZVPD) are identical to those of def2-TZVPPD (def2-QZVPPD).

def2-	$\mathbf{N_{CGTO}}$
def2-SVPD	1,200
def2-TZVP	1,860
def2-TZVPD	2,220
def2-QZVP	3,420
def2-QZVPD	3,780

Table S2: Mean absolute reference energy and corresponding weights for THERMO, LARGE, and BARRIER in GMTKN55. The values are taken from Ref. [5].

GMTKN55	Category	Dataset	$\overline{\lvert E^{\text{ref}} \rvert}$	Weight (w)
1	THERMO	W4-11	306.91	0.19
2	THERMO	G21EA	33.62	1.69
3	THERMO	G21IP	257.61	0.22
4	THERMO	DIPCS10	654.26	0.09
5	THERMO	PA26	189.05	0.30
6	THERMO	SIE4x4	33.72	1.69
7	THERMO	ALKBDE10	100.69	0.56
8	THERMO	YBDE18	49.28	1.15
9	THERMO	AL2X6	35.88	1.58
10	THERMO	HEAVYSB11	58.02	0.98
11	THERMO	NBPRC	27.71	2.05
12	THERMO	ALK8	62.60	0.91
13	THERMO	RC21	35.70	1.59
14	THERMO	G2RC	51.26	1.11
15	THERMO	BH76RC	21.39	2.66
16	THERMO	FH51	31.01	1.83
17	THERMO	TAUT15	3.05	18.66
18	THERMO	DC13	54.98	1.03
19	LARGE	MB16-43	468.39	0.12
20	LARGE	DARC	32.47	1.75
21	LARGE	RSE43	7.60	7.48
22	LARGE	BSR36	16.20	3.51
23	LARGE	CDIE20	4.06	14.02
24	LARGE	ISO34	14.57	3.90
25	LARGE	ISOL24	21.92	2.59
26	LARGE	C60ISO	98.25	0.58
27	LARGE	PArel	4.63	12.28
28	BARRIER	BH76	18.61	3.05
29	BARRIER	BHPERI	20.87	2.72
30	BARRIER	BHDIV10	45.33	1.25
31	BARRIER	INV24	31.85	1.78
32	BARRIER	BHROT27	6.27	9.06
33	BARRIER	PX13	33.36	1.70
34	BARRIER	WCPT18	34.99	1.62

Table S3: Mean absolute reference energy and corresponding weights for INTERMOL and CONFOR in GMTKN55. The values are taken from Ref. [5].

GMTKN55	Category	Dataset	$\overline{\lvert E^{\mathrm{ref}} \rvert}$	Weight (w)
35	INTERMOL	RG18	0.58	98.00
36	INTERMOL	ADIM6	3.36	16.93
37	INTERMOL	S22	7.30	7.78
38	INTERMOL	S66	5.47	10.40
39	INTERMOL	HEAVY28	1.24	45.79
40	INTERMOL	WATER27	81.14	0.70
41	INTERMOL	CARBHB12	6.04	9.42
42	INTERMOL	PNICO23	4.27	13.30
43	INTERMOL	HAL59	4.59	12.38
44	INTERMOL	AHB21	22.49	2.53
45	INTERMOL	CHB6	26.79	2.12
46	INTERMOL	IL16	109.04	0.52
47	CONFOR	IDISP	14.22	4.00
48	CONFOR	ICONF	3.27	17.40
49	CONFOR	ACONF	1.83	30.99
50	CONFOR	Amino20x4	2.44	23.31
51	CONFOR	PCONF21	1.62	35.05
52	CONFOR	MCONF	4.97	11.43
53	CONFOR	SCONF	4.60	12.36
54	CONFOR	UPU23	5.72	9.93
55	CONFOR	BUT14DIOL	2.80	20.30

Table S4: Numerical problem in the optimal structure of C_{60}, the first system in the C60ISO dataset, with def2-TZVPD/def2-QZVPD (TD/QD). The energy components are decomposed for the different basis sets. The problematic orbital is shown in Fig. S7. The basis sets are named in accordance with the Fig. S7.

	TD	**TD/Q**	**TD/QD**	**TD/Q/QD**	**SD/TD/QD**
e1	-19657	-19658	-19614	-19657	-19614
$\mathbf{E}_J$	9375	9375	-2487518743	9375	9350
$\mathbf{E}_K$	-312	-312	1246657962	-312	-311
$\mathbf{E}_{nn}$	8322	8322	8322	8322	8322
$\mathbf{E}_{tot}$	-2272	-2273	-1240872073	-2272	-2253
$\mathbf{E}_{\mathrm{HF-r^2SCAN}}$	-2286	-2286	-2218	-2286	-2220
$\mathbf{E}_{rel}$	0	-0.1	67.5	-0.1	65.8

Table S5: Comparison of the WTMAD-2 values. For DFT, DFT-D4, and D^{2}C-DFT, def2-QZVPPD is employed. The D^{2}C-DFT/2Z/4Z denotes D^{2}C-DFT with def2-SVPD (2Z) and def2-QZVPPD(4Z) as the primary and secondary basis set, respectively. 2Z/4Z can change 1∼3 percentage of WTMAD-2 of D^{2}C-DFT.

WTMAD-2	DFT[a]	DFT-D4[a]	D^{2}C-DFT[a] (A)	D^{2}C-DFT/2Z/4Z (B)	|A-B|/A×100
PBE	13.89	10.12	6.53	6.72	2.9
r^{2}SCAN	8.66	7.11	5.36	5.43	1.3
B3LYP	16.15	6.15	4.67	4.82	3.2

[a] Ref. [8].

Table S6: Average DC4 corrections in small and large non-covalently bound systems and their relative contributions to total interaction energies.

	Dataset	E^{ref}	$E^{\text{DC4}}_{\text{disp}}$	$\|E^{\text{DC4}}_{\text{disp}}/E^{\text{ref}}\|$	$\|E^{(9),\text{DC4}}_{\text{disp}}/E^{\text{DC4}}_{\text{disp}}\|$
		(kcal/mol)		(%)	
Small	HAL59[a]	-4.3	-0.8	27.9	4.6
	S66[b]	-5.5	-1.5	38.1	4.3
Large	L7[c]	-17.1	-9.8	78.0	18.8
	S6L[c]	-43.6	-15.7	43.5	24.9

[a] Ref. [9, 10]. [b] Ref. [11]. [c] Ref. [3].

Table S7: Optimal XDM parameters for HF-r^{2}SCAN with def2-QZVPPD (4Z). The optimization is performed as the same sequence in Ref. [8]. By focusing on the DIET set[12], a condensed version of GMTKN55 derived through genetic algorithms, we minimized the mean absolute error for density-insensitive reactions following the DC-DFT[13] principle. Furthermore, our selected parameters were validated against the WATER27 set, emphasizing accurate water interactions. We used postg program[14] for all XDM calculations in the setting 'hf' atomic volumes.

Basis set	a_1	a_2
def2-QZVPPD	1.0682	1.1821

References

[1] Robert Sedlak, Tomasz Janowski, Michal Pitonak, Jan Rezac, Peter Pulay, and Pavel Hobza. Accuracy of quantum chemical methods for large noncovalent complexes. Journal of chemical theory and computation, 9(8):3364–3374, 2013.

[2] Tobias Risthaus and Stefan Grimme. Benchmarking of london dispersion-accounting density functional theory methods on very large molecular complexes. Journal of chemical theory and computation, 9(3):1580–1591, 2013.

[3] Corentin Villot, Francisco Ballesteros, Danyang Wang, and Ka Un Lao. Coupled cluster benchmarking of large noncovalent complexes in l7 and s12l as well as the c60 dimer, dna–ellipticine, and hiv–indinavir. The Journal of Physical Chemistry A, 126(27):4326–4341, 2022. PMID: 35766331.

[4] Antonio Bauzá, Ibon Alkorta, Antonio Frontera, and José Elguero. On the reliability of pure and hybrid dft methods for the evaluation of halogen, chalcogen, and pnicogen bonds involving anionic and neutral electron donors. Journal of Chemical Theory and Computation, 9(11):5201–5210, 2013. PMID: 26583427.

[5] Lars Goerigk, Andreas Hansen, Christoph Bauer, Stephan Ehrlich, Asim Najibi, and Stefan Grimme. A look at the density functional theory zoo with the advanced gmtkn55 database for general main group thermochemistry, kinetics and noncovalent interactions. Phys. Chem. Chem. Phys., 19:32184–32215, 2017.

[6] Axel D Becke, Golokesh Santra, and Jan ML Martin. A double-hybrid density functional based on good local physics with outstanding performance on the gmtkn55 database. The Journal of Chemical Physics, 158(15), 2023.

[7] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. Phys. Chem. Chem. Phys., 7:3297–3305, 2005.

[8] Minhyeok Lee, Byeongjae Kim, Mingyu Sim, Mihira Sogal, Youngsam Kim, Hayoung Yu, Kieron Burke, and Eunji Sim. Correcting dispersion corrections with density-corrected dft. Journal of Chemical Theory and Computation, 2024.

[9] Sebastian Kozuch and Jan ML Martin. Halogen bonds: Benchmarks and theoretical analysis. Journal of chemical theory and computation, 9(4):1918–1931, 2013.

[10] Jan Rezac, Kevin E Riley, and Pavel Hobza. Benchmark calculations of noncovalent interactions of halogenated molecules. Journal of chemical theory and computation, 8(11):4285–4292, 2012.

[11] Jan Rezác, Kevin E Riley, and Pavel Hobza. S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures. Journal of chemical theory and computation, 7(8):2427–2438, 2011.

[12] Tim Gould. 'diet gmtkn55'offers accelerated benchmarking through a representative subset approach. Physical Chemistry Chemical Physics, 20(44):27735–27739, 2018.

[13] Suhwan Song, Stefan Vuckovic, Eunji Sim, and Kieron Burke. Density sensitivity of empirical functionals. The journal of physical chemistry letters, 12(2):800–807, 2021.

[14] https://github.com/aoterodelaroza/postg.