Supplementary Information for

Efficient yet accurate density-based description of macromolecular polarizabilities

Dongbo Zhao,[a,+] Yilin Zhao,[b,+] Xin He,[c] Paul W. Ayers[b,*] and Shubin Liu[d,e,*]

[a]Institute of Biomedical Research, Yunnan University, Kunming 650500, Yunnan, P.R. China

[b]Department of Chemistry and Chemical Biology, McMaster University, Hamilton, ONL8S 4M1 Canada

[c]Qingdao Institute for Theoretical and Computational Sciences, Shandong University, Qingdao 266237, Shandong, P.R. China

[d]Research Computing Center, University of North Carolina, Chapel Hill, North Carolina 27599-3420, USA

[e]Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599-3290, USA

*Corresponding authors. Emails: ayers@mcmaster.ca; shubin@email.unc.edu.

[+]These authors contributed equally.

The Tkatchenko-Scheffler (TS) formula and its variant. The original Tkatchenko-Scheffler (TS) formula (eq S1) is given as follows:

$$\alpha_{mol}^{TS-old} = \sum_A \alpha_A^{eff} = \sum_A \alpha_A^{free} \left(\frac{V_A^{eff}}{V_A^{free}} \right) \quad \text{(S1)}$$

where the sum runs over all atoms in a molecule. The weights (V_n^{eff}/V_n^{free}) measure the volume ratio for atom in a molecule to the free atom in vacuum are obtained by the Hirshfeld partitioning of the electron density.

Very recently, an improved TS formula is given in eq S2:

$$\alpha_{mol}^{TS} = \sum_A \alpha_A^{eff} = \sum_A \alpha_A^{free} \left(\frac{L_A^{eff}}{L_A^{free}} \right)^4 \quad \text{(S2)}$$

and its equivalent form can be expressed as:

$$\alpha_{mol}^{TS-new} = \sum_A \alpha_A^{eff} = \sum_A \alpha_A^{free} \left(\frac{V_A^{eff}}{V_A^{free}} \right)^{\frac{4}{3}} \quad \text{(S3)}$$

which allows us to keep the simplicity of the original TS method but make it consistent with the L^4 scaling law. The characteristic length L is defined via the L^2 norm of the position operator.

Based on the GEBF method, we propose to use the equations (GEBF-TS) proposed in this work for a macromolecule as given in eqs S4 and S5:

$$\alpha_{mol}^{GEBF-TS} = \sum_m C_m \sum_A \frac{\alpha_{m,A}^{free} V_A^{eff}}{V_A^{free}} \quad \text{(S4)}$$

and

$$\alpha_{mol}^{GEBF-TS} = \sum_m C_m \sum_A \alpha_{m,A}^{free} \left(\frac{V_A^{eff}}{V_A^{free}} \right)^{\frac{4}{3}} \quad \text{(S5)}$$

Table S1. The slope (a) and intercept (b) data of linear equations for inorganic systems in **Table 1** in text.

systems	S_S		I_F		S_{GBP}		E_2		E_3		$^rR_2$		$^rR_3$	
	a	b	a	b	a	b	a	b	a	b	a	b	a	b
Be	2.842	53.569	0.163	69.142	0.620	68.738	2.052	70.621	0.207	69.891	4.153	69.203	4.032	65.854
Mg	-15.918	212.158	0.048	109.179	0.622	105.706	0.163	109.425	0.001	109.032	4.209	110.992	4.157	113.617
Na	-29.625	501.736	0.075	352.646	0.881	349.837	0.285	352.995	0.001	352.906	6.041	349.914	6.017	344.027
S	-0.348	-0.947	0.002	-1.406	0.031	-1.440	0.004	-1.401	6×10^{-6}	-1.401	0.209	-1.390	0.207	-1.370
Au	-0.149	-22.954	4×10^{-4}	-23.471	0.337	-24.667	1×10^{-4}	-23.467	4×10^{-10}	-23.467	0.539	-23.514	0.539	-23.612
HF	1.756	-1.912	0.008	-1.426	0.064	-1.426	0.037	-1.408	3×10^{-4}	-1.413	0.433	-1.446	0.424	-1.497
H_2O	1.139	-3.882	0.018	-2.901	0.121	-2.895	0.102	-2.799	0.001	-2.814	0.810	-3.003	0.782	-3.312
$H^+\cdot(H_2O)_n$	1.133	-4.737	0.018	-4.558	0.120	-4.510	0.101	-4.433	0.001	-4.409	0.805	-4.138	0.777	-4.008
CO_2	1.802	-5.037	0.014	-2.630	0.104	-2.654	0.080	-2.516	0.001	-2.521	0.703	-2.562	0.690	-2.616
$Cl^-\cdot(NaCl)_n$	-1.469	16.720	0.007	8.099	0.120	9.532	0.019	5.929	3×10^{-5}	2.973	0.808	8.491	0.801	6.969
fullerene	2.235	-33.268	0.040	-22.488	0.240	-23.125	0.308	-22.128	0.009	-22.352	1.589	-21.812	1.534	-20.947
BN	1.747	-19.541	0.028	-14.981	0.172	-15.049	0.201	-14.720	0.004	-14.899	1.144	-14.797	1.102	-14.737
ZnS	-0.639	-78.052	0.005	-79.858	0.139	-80.192	0.006	-79.840	1×10^{-6}	-79.853	0.933	-80.02	0.927	-80.319

systems	G_1		G_2		G_3		I_G		Vol		Θ_{iso}	
	a	b	a	b	a	b	a	b	a	b	a	b
Be	23.626	107.173	-19.590	89.781	4.713	105.987	373.071	25.617	0.160	16.343	-2.990	67.785
Mg	-218.581	331.026	638.612	232.314	20.212	284.287	1291.7	354.051	0.212	7.970	-4.623	80.092
Na	-271.792	564.302	307.67	666.133	51.071	586.072	5879.188	-1200.717	0.275	336.336	-7.660	330.870
S	-4.651	0.521	8.685	0.512	0.668	1.157	95.812	1.108	0.018	-2.389	-0.261	-1.322
Au	-198.533	27.289	233.489	84.277	8.564	67.808	8600.3	-350.028	0.204	-76.94	-1.862	-16.797
HF	15.328	7.625	-5.462	-1.679	0.650	0.817	48.57	1.084	0.036	-2.953	-0.724	0.114
H_2O	-149.091	-59.811	-8.969	0.694	0.771	0.721	72.072	-15.033	0.051	-1.673	-1.318	-2.570
$H^+\cdot(H_2O)_n$	-68.874	-174.427	-9.157	15.159	0.763	-3.115	72.404	35.460	0.052	-5.900	-1.756	13.151
CO_2	-5.217	-3.567	10.243	-8.850	0.748	3.335	128.955	-7.999	0.059	-5.109	-0.971	-2.545
$Cl^-\cdot(NaCl)_n$	-92.142	346.011	-180.45	-659.29	4.284	43.539	187.072	-240.835	0.073	-10.377	-1.107	-78.332
fullerene	-3.791	-21.469	5.002	-28.39	1.167	-12.911	178.98	3.470	0.140	-42.282	-1.575	-27.901
BN	7.250	-0.516	-4.482	1.738	0.902	-7.141	103.669	-24.402	0.108	-47.547	-1.279	-15.974
ZnS	-84.901	-128.41	817.019	277.001	4.517	-51.717	536.141	-137.18	0.133	-177.29	-1.496	-69.243

Table S2. The slope (a) and intercept (b) data of linear equations for organic and biological systems in **Table 2** in text.

systems	S_S		I_F		S_{GBP}		E_2		E_3		$^rR_2$		$^rR_3$	
	a	b	a	b	a	b	a	b	a	b	a	b	a	b
polyene	5.965	-286.538	0.183	-248.663	0.960	-260.283	1.440	-247.035	0.041	-246.799	6.322	-260.891	5.974	-262.735
polyyne	9.309	-331.732	0.192	-270.763	1.150	-284.868	1.498	-269.095	0.042	-268.923	7.745	-286.387	7.475	-289.501
acene	3.716	-183.391	0.090	-148.446	0.507	-157.695	0.706	-147.281	0.020	-147.103	3.350	-158.252	3.197	-159.851
all-*trans*	4.131	-119.832	0.079	-93.213	0.486	-99.229	0.551	-92.482	0.011	-92.473	3.244	-99.722	3.118	-101.010
push-pull	7.278	-160.745	0.036	-96.366	0.368	-119.745	0.111	-68.914	2×10^{-4}	-42.888	2.454	-119.010	2.380	-118.958
octane	6.853	-505.377	1.451	-2841.300	6.425	-2763.602	6.781	-1644.672	0.024	-119.630	29.883	-1951.981	11.146	-723.865
base pairs	2.284	-61.598	0.029	-0.406	0.211	-25.122	0.170	24.510	0.001	139.209	1.411	-26.528	1.384	-30.272
$(C_6H_6)_n$	1.441	1.539	0.044	3.914	0.229	4.078	0.343	4.217	0.010	3.841	1.503	4.371	1.422	3.927
amino acids	1.290	10.357	0.027	-5.749	0.221	-45.850	0.032	106.644	-9×10^{-7}	133.308	1.462	-45.721	1.396	-45.878
dipeptides	1.276	19.363	13.502	669.760	0.263	800.745	2.212	117.472	0.285	2297.81	1.456	-46.452	14.676	105.931
tripeptides	1.285	29.492	0.027	19.339	0.222	-63.175	0.032	231.670	-1×10^{-6}	282.430	1.469	-64.451	1.115	5.841
α-helix	1.406	-8.606	0.029	-3.102	0.177	-5.067	0.198	-2.892	0.003	-1.516	1.169	-6.196	1.116	-7.184
3_{10}-helix	1.404	-5.790	0.029	-1.335	0.177	-3.268	0.198	-1.223	0.003	0.206	1.170	-4.476	1.117	-5.382
β-sheet	1.454	-14.461	0.031	-12.155	0.184	-12.263	0.207	-13.037	0.004	-15.584	1.217	-12.194	1.160	-12.333

systems	G_1		G_2		G_3		I_G		Vol		Θ_{iso}	
	a	b	a	b	a	b	a	b	a	b	a	b
polyene	-13.545	-267.687	18.052	-265.26	5.113	-250.311	429.681	-304.233	0.335	-283.439	-7.765	-258.650
polyyne	-18.976	-297.677	28.783	-337.798	7.425	-275.075	913.626	-414.168	0.444	-318.315	-12.849	-357.770
acene	-7.516	-164.657	9.910	-162.507	2.596	-150.950	277.713	-199.705	0.230	-204.044	-4.265	-162.370
all-*trans*	-13.633	-110.659	23.967	-106.599	2.720	-92.385	318.904	-141.212	0.208	-119.272	-2.171	21.968
push-pull	-8.615	-43.567	11.473	-10.900	2.676	-128.060	313.697	-137.247	0.179	-124.434	-2.791	-90.512
octane	-13.117	-345.001	1.396	62.909	-1.613	245.143	42.257	40.916	0.007	87.895	-0.886	48.131
base pairs	0.595	187.221	-2.782	204.831	1.093	-14.576	0.089	-3.365	0.595	187.221	-0.862	86.547
$(C_6H_6)_n$	-3.242	30.994	4.244	81.139	1.155	16.161	106.005	-9.943	0.090	-33.578	-1.849	7.548
amino acids	-3.373	25.857	4.16	56.666	1.085	-24.730	98.166	-26.206	0.078	-10.247	-1.951	-39.870
dipeptides	14.070	88.583	-31.038	1142.760	38.555	1501.210	10.606	398.525	290.952	1931.1	-0.296	165.580
tripeptides	0.009	1.028	0.007	1.394	-0.009	2.575	0.012	2.719	0.005	1.765	5×10^{-3}	282.960
α-helix	-3.521	170.222	4.746	305.087	0.917	-2.416	94.612	-47.349	0.089	-68.006	-1.514	9.112
3_{10}-helix	-3.530	172.497	4.758	309.932	0.926	-2.834	93.593	-43.945	0.084	-54.842	-1.544	4.281
β-sheet	-4.616	-5.806	7.946	-3.567	0.978	-8.001	90.514	-17.983	0.083	-11.191	-1.691	-16.091

Table S3. Molecular polarizabilities (in Bohr^3) of fullerene cages (C_{100} and C_{180}), 15 tetrapeptides and 10 proteins, predicted by the original TS approach[a] with the molecular wavefunction obtained at the M06-2X/def2-TZVP level.

systems	eq S1			eq S3	conv.
	Becke	Hirshfeld	avg.	Hirshfeld	
fullerenes					
C_{100}	626.5	1019.2	822.9	984.5	993.7
C_{180}	1119.4	1827.0	1473.2	1762.2	1932.5
tetrapeptides					
Ala_Ser_Tyr_Asp	265.8	400.8	333.3	381.4	323.0
Cys_Ala_Val_Thr	247.0	365.3	306.2	344.2	289.7
Asp_Tyr_Cys_Ala	279.7	416.7	348.2	397.6	336.2
Glu_Trp_Ser_Gln	321.1	485.0	403.1	461.5	399.3
Phe_Val_Tyr_Ser	319.8	482.6	401.2	455.6	390.6
Gly_Thr_Gln_His	260.7	389.7	325.2	368.1	314.4
His_Ser_Cys_Trp	319.0	472.8	395.9	448.7	396.2
Leu_Arg_Pro_Gln	314.7	479.2	397.0	448.8	371.2
Thr_Gln_Asp_Ala	252.6	379.6	316.1	360.6	301.2
Val_Pro_Ala_Phe	281.1	423.2	352.2	398.0	340.0
Arg_Cys_Glu_Gly	273.3	406.1	339.7	384.3	328.0
Cys_Ile_Arg_Asn	305.4	457.1	381.3	429.6	361.3
Glu_Trp_Tyr_Glu	371.9	555.8	463.9	531.1	467.6
Phe_Gly_Asn_Asn	267.1	400.6	333.9	378.9	325.9
Pro_Tyr_Pro_Ser	281.3	427.8	354.6	403.4	337.7
proteins					
4TTK	819.8	1259.9	1039.8	1192.1	985.4
2LDJ	1192.4	1864.3	1528.3	1766.2	1435.2
2LL5	1105.3	1718.6	1411.9	1634.3	1335.1
2LYE	1401.9	2140.6	1771.3	1569.6	1355.1
5JI4	1113.3	1665.5	1389.4	3102.2	2513.6
1SP7	1151.6	1807.5	1479.6	2028.6	1657.3
1L2Y	2214.4	3465.7	2840.0	1712.7	1375.0
2F4K	2154.4	3881.2	3017.8	3290.1	2647.0
6PHM	1762.6	2745.2	2253.9	2610.7	2168.6
1CBN	2502.2	3881.1	3191.6	3690.1	3010.1
MAE(%)[b]	22.0	30.0	26.0	18.2	

[a]Two partition schemes of Becke and Hirshfeld are employed.

[b]**MAE**: mean absolute error.

Table S4. Molecular polarizabilities (in $Bohr^3$) based on the subsystem wavefunctions obtained at the M06-2X/def2-TZVP level of 12 protein systems.

systems	GEBF-TS		GEBF
	eq S4	eq S5	
4TTK	819.8	1192.2	982.7
2LDJ	1864.4	1766.1	1438.8
2LL5	1718.6	1633.9	1332.2
2LYE	1665.5	1569.5	1345.4
5JI4	3255.2	3102.2	2513.6
1SP7	2140.7	2028.8	1661.3
1L2Y	1807.5	1712.7	1379.7
2F4K	3465.3	3289.3	2663.7
6PHM	2745.2	2610.7	2143.7
1CBN	3881.3	3690.3	3007.2
1UBQ	7182.6	6801.3	5574.8
1K6F	6368.4	6018.6	4833.3
MAE(%)[a]	29.0	22.3	

[a]**MAE**: mean absolute error.

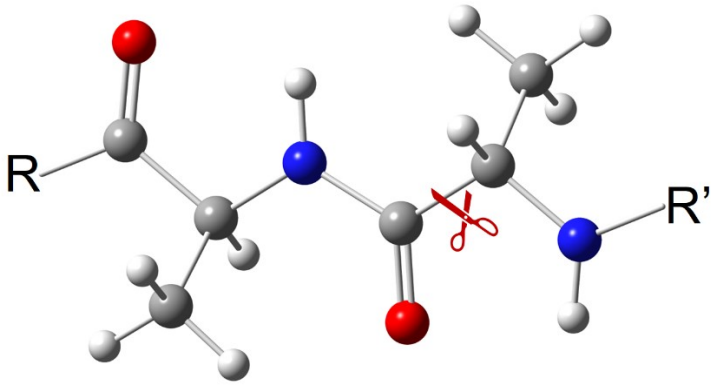

Fig. S1 Fragmentation details of proteins. For clarity, we only employ a dipeptide (R-Ala-Ala-R', R and R' are natural amino acids) where the single C–C bond is cut. Color code: H in white, C in grey, N in blue and O in red.

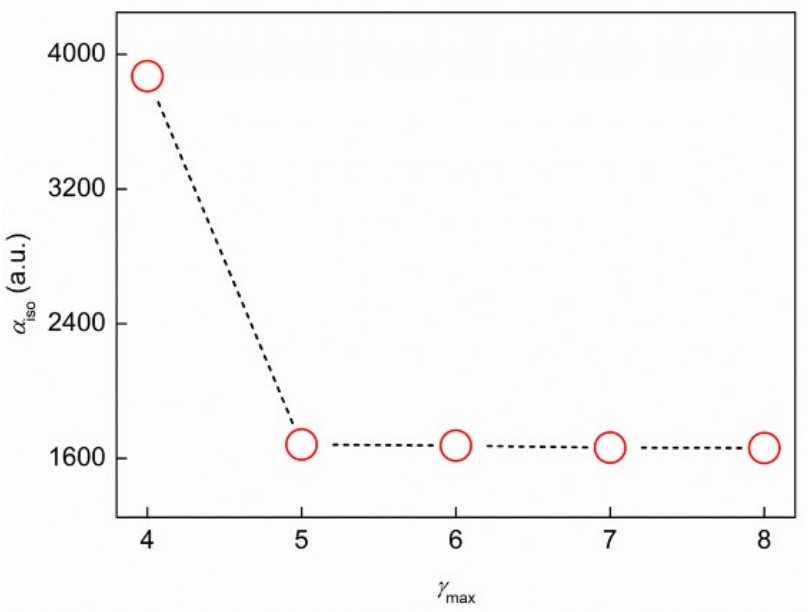

Fig. S2 Convergence check of molecular polarizabilities of 1SP7 with respect to the maximum number of fragments in a subsystem (γ_{max}).

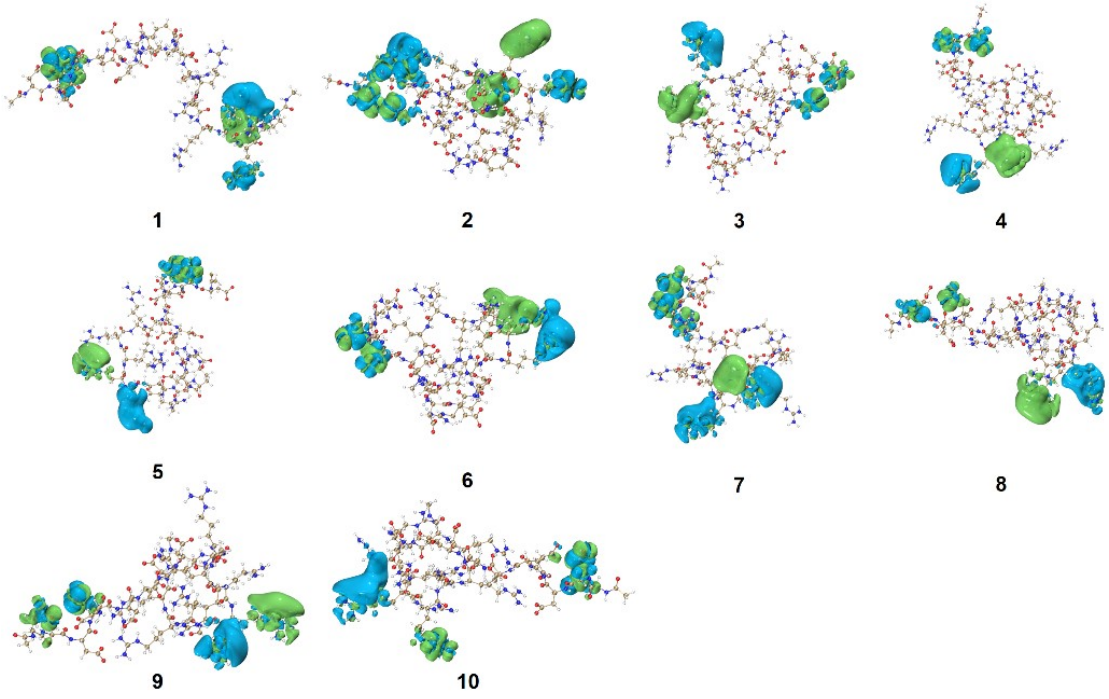

Fig. S3 Spin density maps (at the UHF/cc-pVDZ level) of 10 representative $D_{10}R_{10}$ structures from classical MD simulations with the ff19SB force fields for $D_{10}R_{10}$ and TIP3P for water molecules (not shown here). The isovalue is 0.0004 a.u.