Supporting Information for

Efficient calculation of protein-ligand binding free energy

with GFN methods: the power of cluster model

Yuan-qiang Chen,[1] Yan-jing Sheng,[1] Yu-qiang Ma,[2] and Hong-ming Ding[1*]

[1] Center for Soft Condensed Matter Physics and Interdisciplinary Research, School of Physical Science and Technology, Soochow University, Suzhou 215006, China.

[2] National Laboratory of Solid State Microstructures and Department of Physics, Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing 210093, China.

*Email: dinghm@suda.edu.cn

Table S1. The PDB entries, the net charge of protein and ligand, the experimental pK_d ($pK_d = -\log K_d$) and the descriptions of the complexes.

ID	PDB	pK_d	Charge(Pro/Lig)	Descriptions
1	186L	4.85	+8/0	Enterobacteria phage T4 LYSOZYME / N4B
2	188L	3.33	+8/0	Enterobacteria phage T4 LYSOZYME / OXE
3	1A28	8.29	+1/0	ligand-binding domain of the human progesterone receptor / STR
4	1A52	9.86	-7/0	human estrogen receptor alpha ligand-binding domain / EST
5	1AU0	7.66	+7/0	cathepsin k / SDK
6	1BCU	3.28	+7/0	thrombin alpha / PRL
7	1CEA	4.96	0/0	Plasminogen / ACA
8	1CEB	6	0/0	Plasminogen / AMH
9	1CTU	11.92	-9/0	cytidine deaminase / ZEB
10	1D2S	9.74	-6/0	N-terminal laminin G-like domain of human sex hormone-binding globulin / DHT
11	1D7I	3.6	+1/0	FK506 binding protein / DSS
12	1D7J	3.3	+1/0	FK506 binding protein / BUQ
13	1DRK	6.82	-2/0	rous sarcoma virus protease / RIP
14	1E02	6.15	-16/0	odorant-binding protein / UNA
15	1E3G	9	+5/0	Ligand-binding Domain (LBD) of human Androgen Receptor / R18
16	1ERB	8.1	-4/0	retinoic acid receptor rxr-alpha / ETR
17	1FKH	8.15	+1/0	FK506 binding protein / SBX
18	1G50	6.28	-6/0	wildtype human estrogen receptor alpha ligand-binding domain (hERalpha LBD) / EST
19	1G85	5.48	-10/0	odorant-binding protein / 3OL
20	1GX8	6.37	-8/0	bovine beta-lactoglobulin / RTL
21	1J01	6.47	-8/0	xylanase beta-1,4 / XIL
22	1JAQ	4.48	-12/0	matrix metalloproteinase-8 / 01S

23	1KDK	9.05	-5/0	sex hormone-binding globulin / DHT
24	1KMY	5.1	-10/0	2,3-dihydroxybiphenyl-1,2-dioxygenase / BPY
25	1L83	3.4	+8/0	lysozyme C54T C97A L99A / BNZ
26	1LHU	8.83	-4/0	N-terminal laminin G-like domain of human sex hormone-binding globulin / EST
27	1LI2	4.04	+8/0	Lysozyme / IPH
28	1LKE	7.52	-4/0	diga16 / DOG
29	1OGZ	5.85	-4/0	STEROID DELTA-ISOMERASE(P39A) / EQU
30	1QKU	9.59	-6/0	ESTRADIOL RECEPTOR/ EST
31	1RBP	6.72	-4/0	retinoic acid receptor rxr-alpha / RTL
32	1RDL	2.24	-1/0	mannose-binding protein-c / MMA
33	1W6Y	5.36	-7/0	steroid delta-isomerase / EQU
34	2A2G	7	-5/0	alpha-2u-globulin / LEO
35	2E2R	8.25964	-11/0	human estrogen-related receptor c (ERRγ) / 2OH
36	2I0G	9.72	-4/0	estrogen receptor beta / I0G
37	2QE4	7.96	-6/0	Estrogen receptor / JJ3
38	2XAB	9.27	-8/0	HSP90 / VHD
39	2XHT	5.96	-11/0	HSP90 / C0Y
40	3EKO	6.69897	-11/0	Hsp90 Thr184 / PYU
41	3KGT	5.51	-5/0	V30M mutant human transthyretin (TTR) / GEN
42	4HBV	4.64	+2/0	Bromodomain-containing protein 4 / 15E
43	4MRE	2.19	-4/0	murine cd44 hyaluronan binding domain / 2C9
44	4O09	7.7	-7/0	Heat shock protein HSP 90-alpha / 2R6
45	4QEV	7.66	+1/0	Bromodomain-containing protein 2, BRD2(BD2) L383A mutant / 31O
46	4QF7	8.6	+4/0	Phospholipase A2 VRV-PL-VIIIa / C0R
47	4STD	10.33	-4/0	scytalone dehydratase / BFS

48	5G2G	6.05	-5/0	ketosteroid isomerase mutant M112K / EQU
49	6ELN	6.05	-7/0	Heat shock protein HSP 90-alpha / P4A
50	7STD	10.72	-4/0	scytalone dehydratase / CRP
51	1C4V	10.8	+6/+1	thrombin alpha / IH2
52	1C5P	4.68	+6/+1	trypsin beta / BEN
53	1CBR	9.4	-4/-1	mouse cellular retinol-binding protein (CRBP-1) / REA
54	1CE5	4.74	+6/+1	trypsin beta / BEN
55	1DB5	6.82	+15/-1	human non-pancreatic secretory phospholipase A2 / 6IN
56	1DF8	9.7	-1/-1	Streptavidin / BTN
57	1EZQ	9.05	+1/+2	coagulation factor xa / RPR
58	1F0U	7.16	+6/+2	trypsin beta / RPR
59	1FCY	8.52	-3/-1	retinoic acid receptor gamma-1 / 564
60	1FCZ	9.22	-3/-1	retinoic acid receptor gamma-1 / 156
61	1FIG	6.22	+4/-2	antibody fab fragment / TSA
62	1I2S	3.31	-8/-3	beta-lactamase from Bacillus licheniformis BS3 / CIT
63	1LAH	7.52	-3/+1	diga16 / ORN
64	1LPG	7.09	+1/+2	coagulation factor xa / IMA
65	1NF8	7.82	-4/-2	phenazine biosynthesis protein phzd / ISC
66	1O4M	1.4	+4/-2	proto-oncogene tyrosine-protein kinase src/ MLA
67	1O4P	2.6	+4/-2	proto-oncogene tyrosine-protein kinase src / 791
68	1QB6	6.06	+6/+2	trypsin beta / 623
69	1TNG	2.93	+6/+1	trypsin beta / AMC
70	1TOM	8.3	+7/+2	thrombin alpha / MIN
71	1V2N	5.9	+5/+2	trypsin beta / BBA
72	1XUG	7.05	+6/+3	Trypsin / BAB
73	2BOK	6.55	-1/+3	coagulation factor xa / 784
74	2GKL	5.35	0/-2	zinc carbapenemase CPHA / PD2

| 75	2GTV	5.15	+1/-2	chorismate mutase / TSA
76	2J4A	10.6	-9/-1	HUMAN THYROID HORMONE RECEPTOR BETA LIGAND BINDING DOMAIN / OEF
77	2V77	5.06048	-6/-4	human carboxypeptidase A1 (hCPA1) / PAY
78	2YG2	6.05	-5/-3	APOLIPOPROTEIN M / FLC
79	2ZFS	6.54	+6/+2	Cationic trypsin / 12U
80	2ZFT	6.52	+6/+2	Cationic trypsin / 10U
81	3B3X	3.6	-8/-2	class A beta-lactamase of Bacillus licheniformis BS3 / A33
82	3BU1	8.14874	-3/+1	AM-10, monomine / HSM
83	3TVC	7.74	-10/-1	Collagenase 3 / E3P
84	4BAQ	9.01	+7/+2	THROMBIN / M4Z
85	4GGZ	10.18	+5/-1	Bradavidin 2 / BTN
86	4M8H	7.6	-7/-1	human Retinoic acid receptor RXR-alpha ligand binding domain / R4M
87	4O9V	8	-6/+3	Suppressor of tumorigenicity 14 protein Matriptase catalytic domain complex / NT4
88	4QF8	7.51	+4/+3	Phospholipase A2 VRV-PL-VIIIa / SPD
89	4RGD	3.94	+8/-3	Bacteriocin AS-48 G13K/L40K mutant / CIT
90	6GGD	5.59	+2/+1	Cellular tumor antigen p53 / EYB
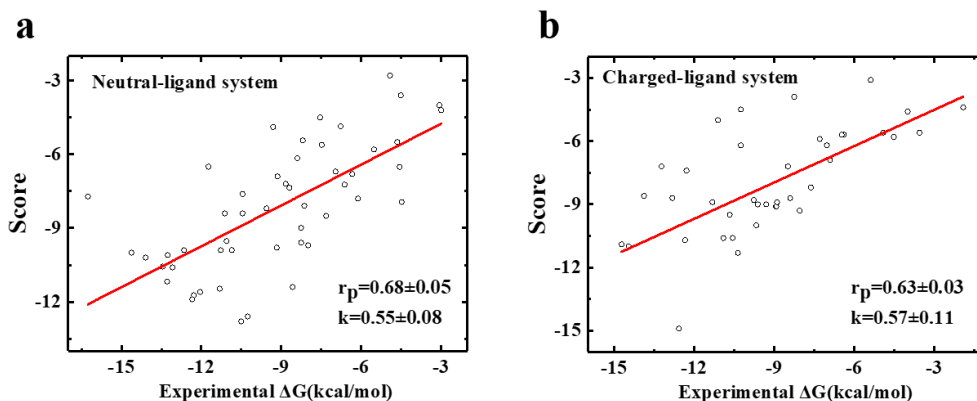
Figure S1. Correlation between experimental measurements and the docking score calculated by using the Autodock vina in (a) neutral-ligand system and (b) charged-ligand system. The uncertainties of r_p and k is estimated by the bootstrapping algorithm (i.e., randomly selecting 80% of each dataset with the repeat times equaling to the number of complexes in each dataset).
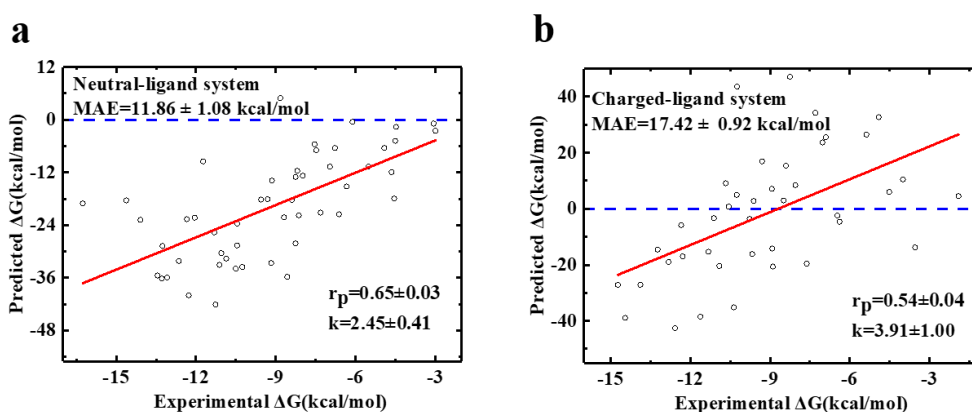

Figure S2. MAE and correlation between experimental and the predicted binding free energy with the MM/PBSA method in (a) neutral-ligand system and (b) charged-ligand system. The uncertainties of r_p and k is estimated by the bootstrapping algorithm (i.e., randomly selecting 80% of each dataset with the repeat times equaling to the number of complexes in each dataset).
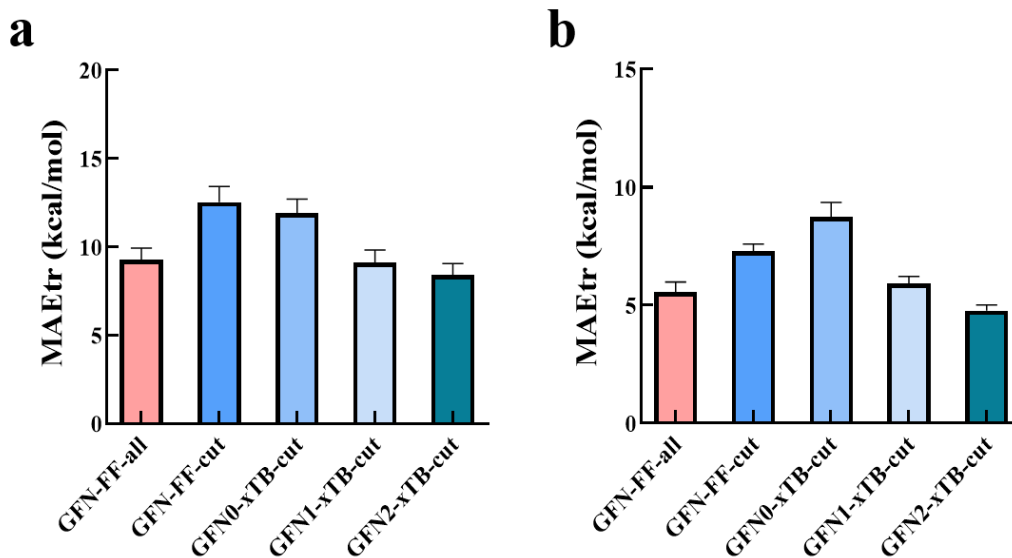
Figure S3. MAEtr for the binding free energy of the whole protein-ligand (calculated using GFN-FF method) and its cluster model (calculated using GFN0, GFN1, GFN2-xTB and GFN-FF methods), after the removal of the systematic error (mean signed error, MSE). (a) MAEtr value of the charged-ligand system; (b) MAEtr value of the neutral-ligand system.
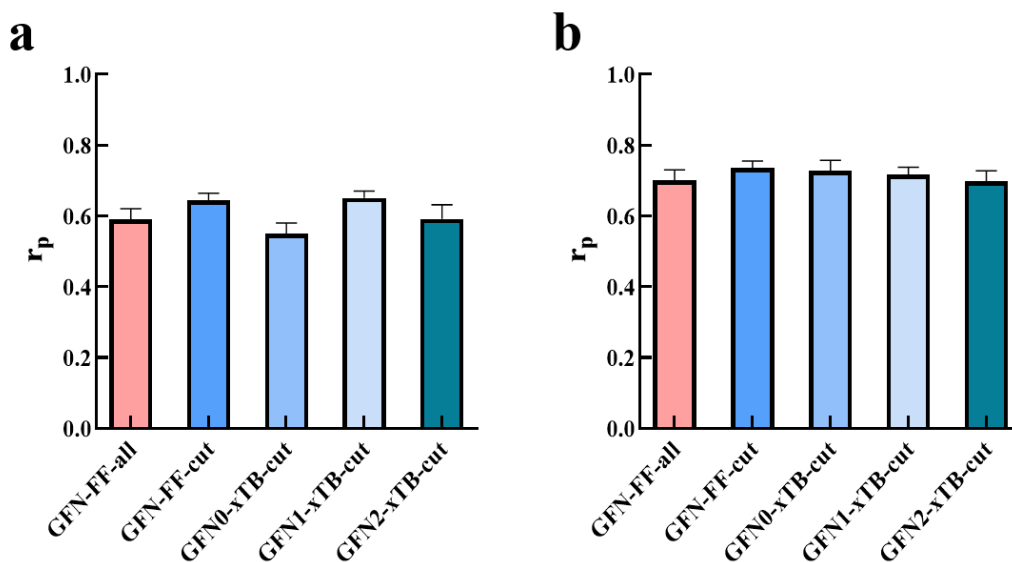

Figure S4. The Pearson correlation coefficients r_P of (a) the charged-ligand truncated systems and (b) the neutral-ligand truncated systems under different GFN methods.
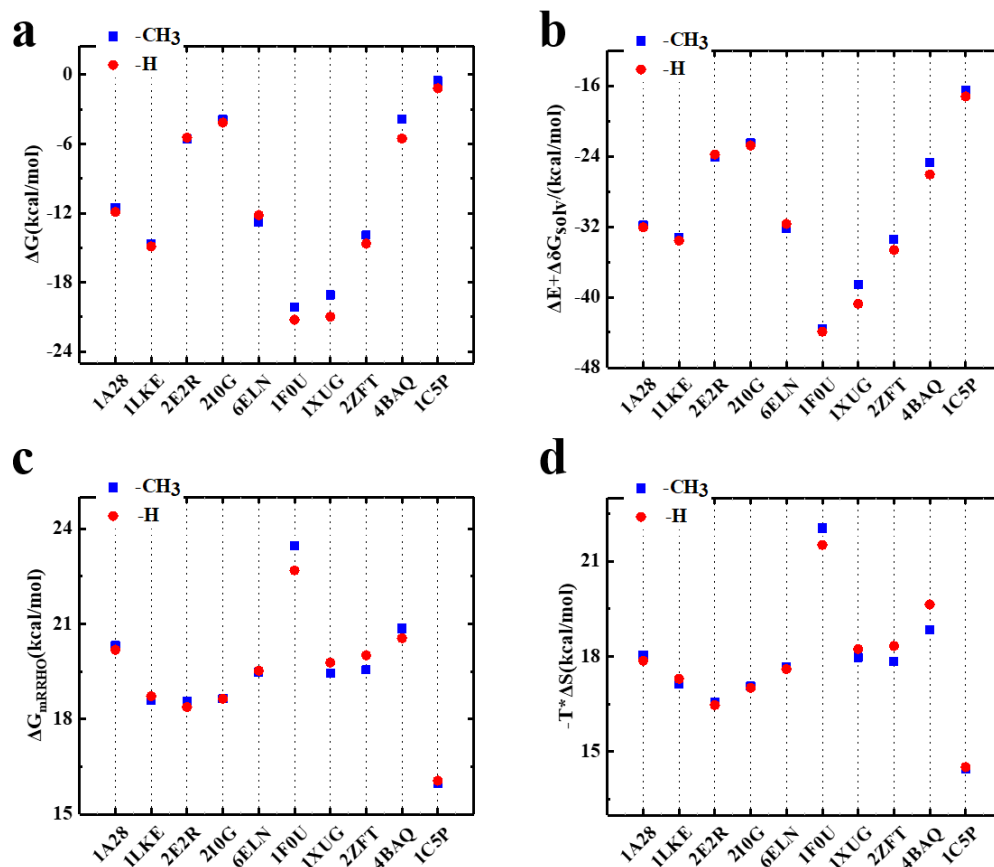
Figure S5. In the cluster model with the GFN2-xTB method, (a) association free energy (ΔG), (b) enthalpy changes ($\Delta E + \Delta \delta G_{solv}$), (c) thermostatistical corrections of free energy (ΔG_{mRRHO}) and (d) entropy changes ($-T*\Delta S$) of ten protein-ligand systems using different truncated methods (hydrogen atoms and methyl groups).
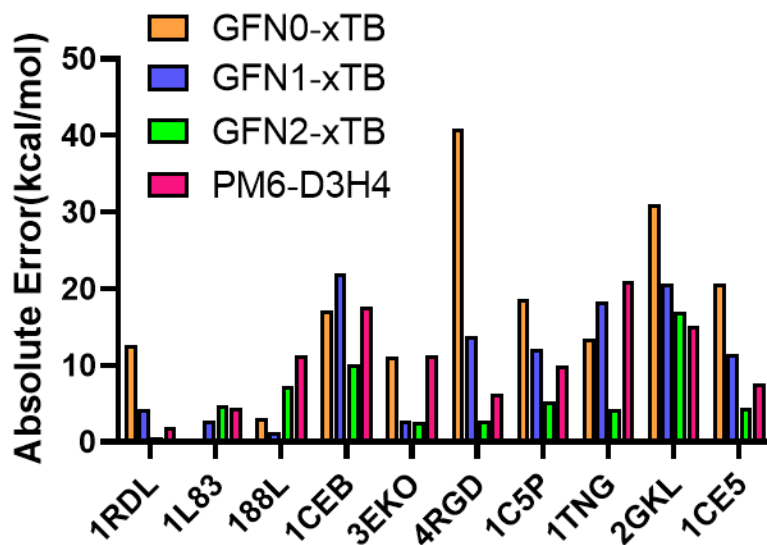
Figure S6. Absolute error (i.e., $|\Delta G_{exp} - \Delta G_{cal}|$) for the binding free energy of the truncated systems with the PM6-D3H4 and GFNn-xTB methods.