Electronic Supplementary Material (ESI) for New Journal of Chemistry.

Supporting information: A machine learning approach using frequency descriptor for molecular property predictions

Jialu Chen,[1] Wenjun Xu,[1] Ruiqin Zhang*[1,2]

[1]Department of Physics, City University of Hong Kong, Hong Kong SAR, People's Republic of China

[2]Beijing Computational Science Research Center, Beijing 100193, People's Republic of China

*Corresponding author.

Email address: aprqz@cityu.edu.hk

Table S1 Hyperparameters of different levels of calculations in training KRR with RBF kernel function.

Method	γ	λ
PM6	0.02	0.001
PM7	0.02	0.001
XTB	0.02	0.001
B3LYP/6-31G(2df,p)	0.025	0.001

Table S2 Hyperparameters of different levels of calculations in Δ-ML training of KRR with RBF kernel function. For the Mix1 method, the energy of GFN2-xTB and frequencies of PM7 method were used. while the energy of PM7 and frequencies of GFN2-xTB method were used in the Mix2 method.

Method	γ	λ
PM6	0.03	0.001
PM7	0.03	0.001
XTB	0.02	0.001
Mix1	0.03	0.001
Mix2	0.04	0.001
PBE	0.015	0.001
GFN2-PBE	0.015	0.001
GFN2-M06-2X	0.015	0.001
GFN2-PBE2	0.015	0.001

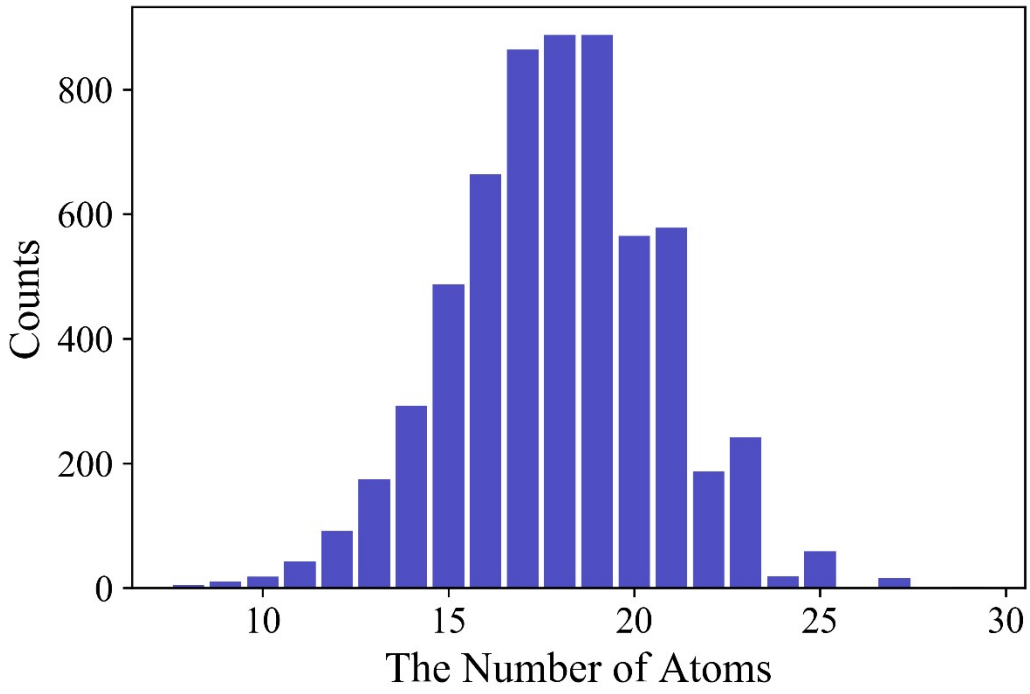

Fig. S1 The number of atoms of the randomly chosen 6095 molecules from the QM9 dataset.

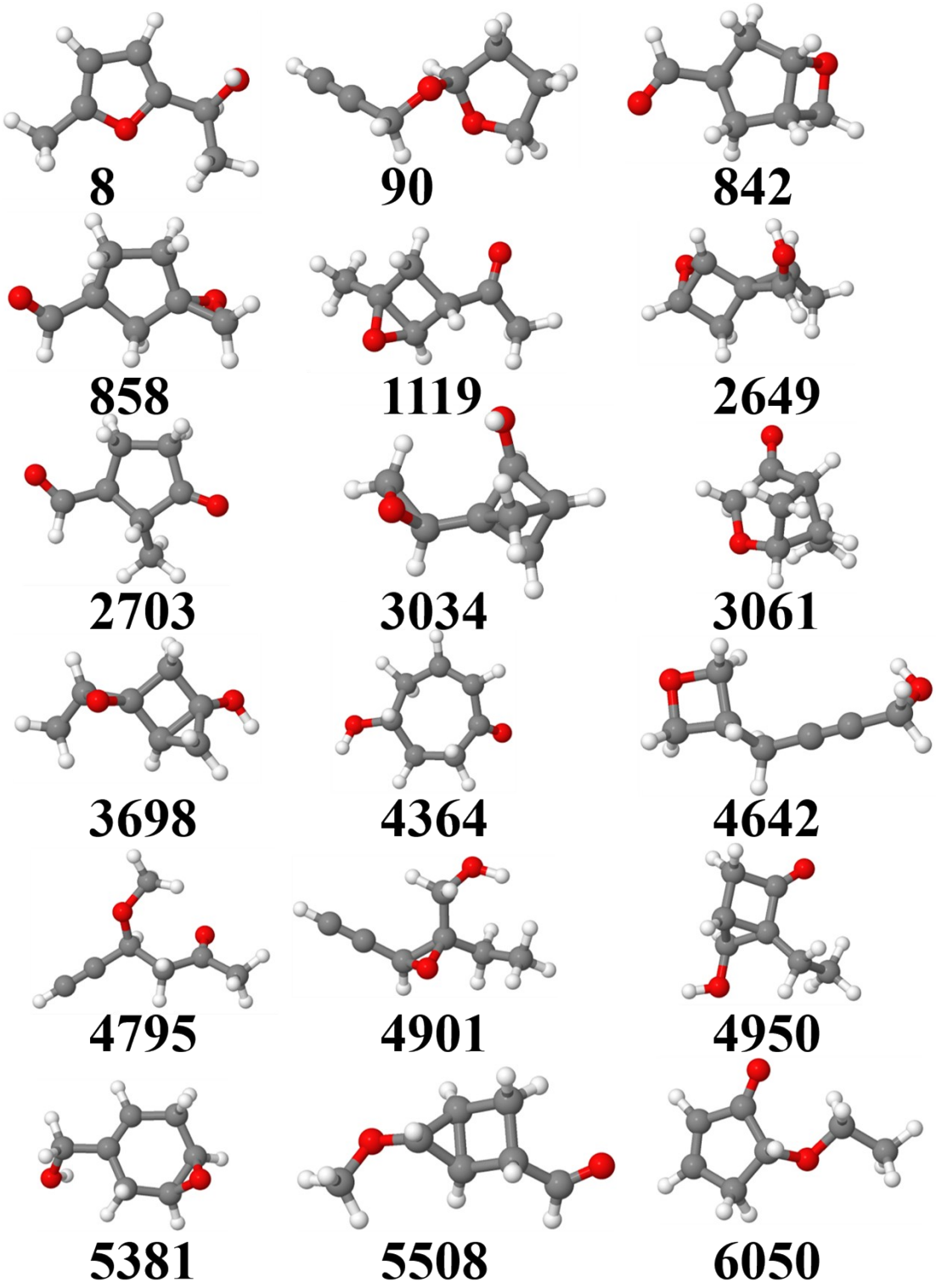

Fig. S2 Configurations were not converged within default steps of PM7 calculations using Gaussian 16.

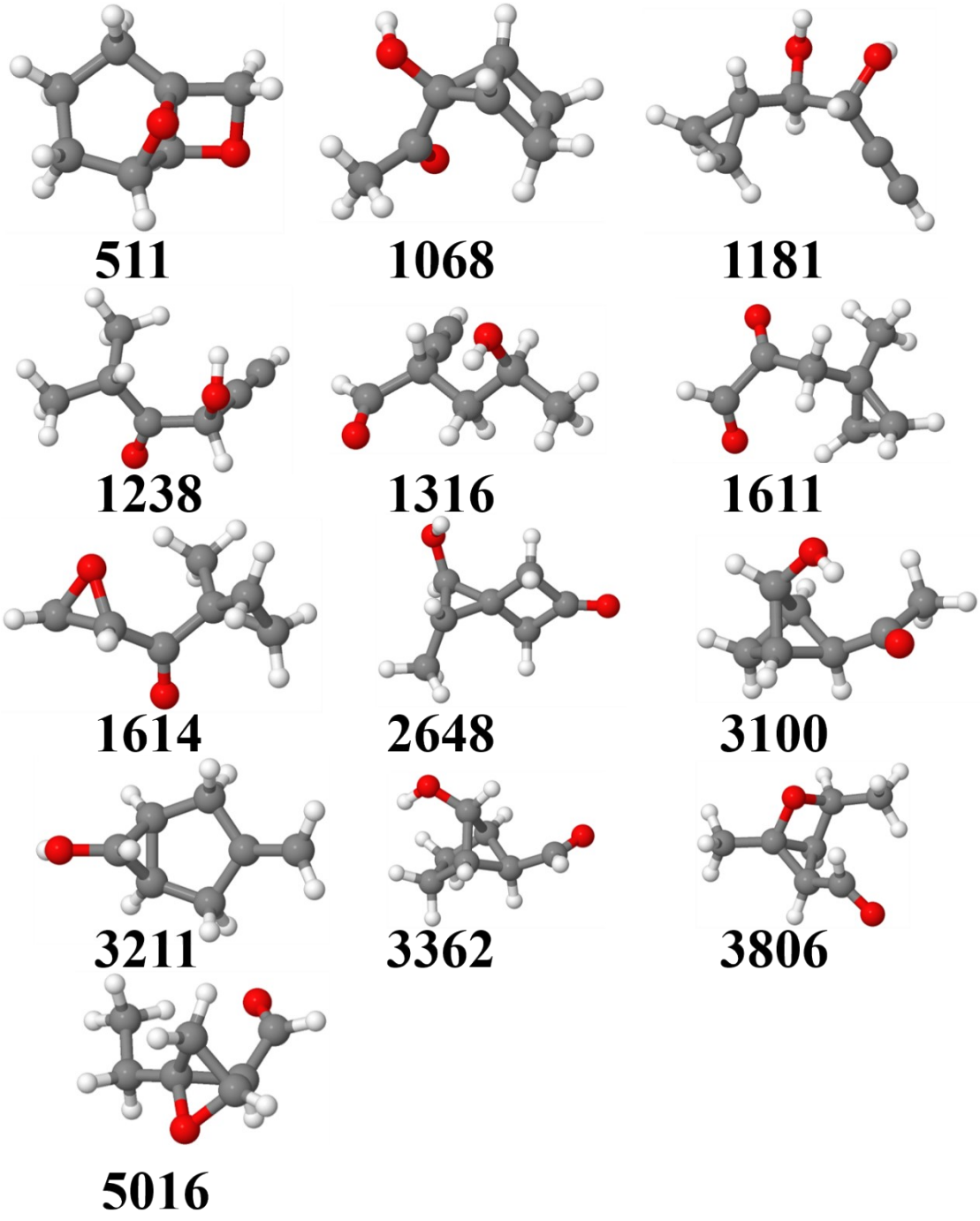

Fig. S3 Configurations were not converged within default steps of PM6 calculations using Gaussian 16.

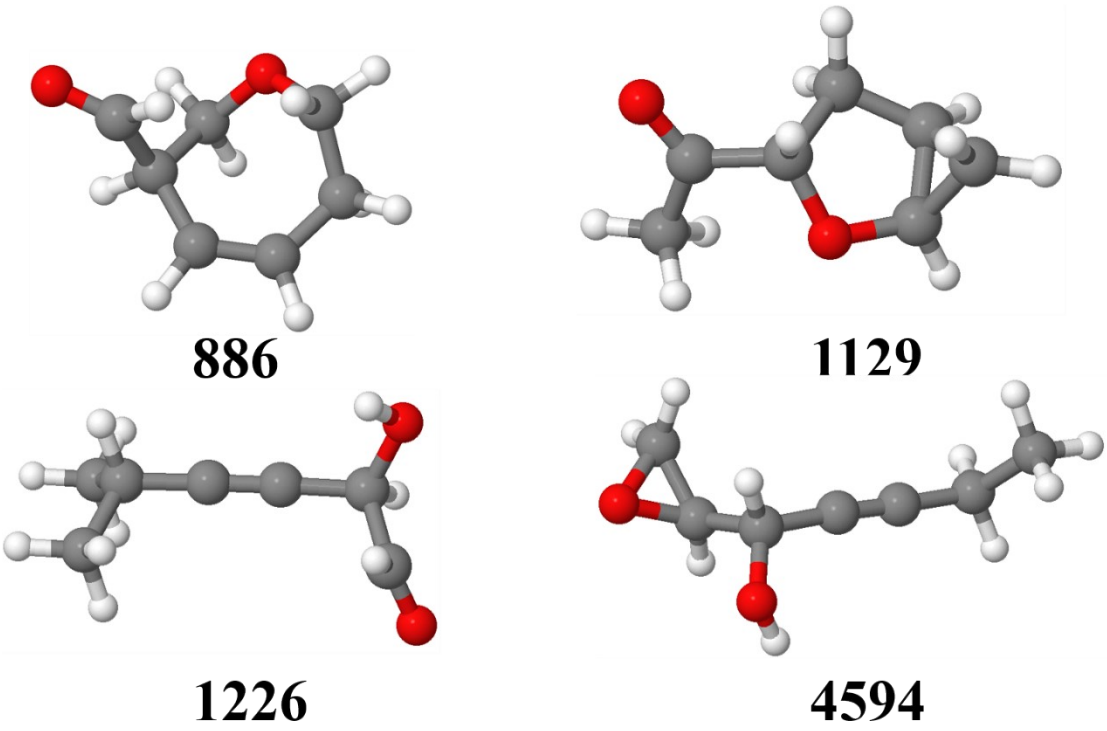

Fig. S4 Configurations were not converged within default steps of PBE calculations using Gaussian 16.

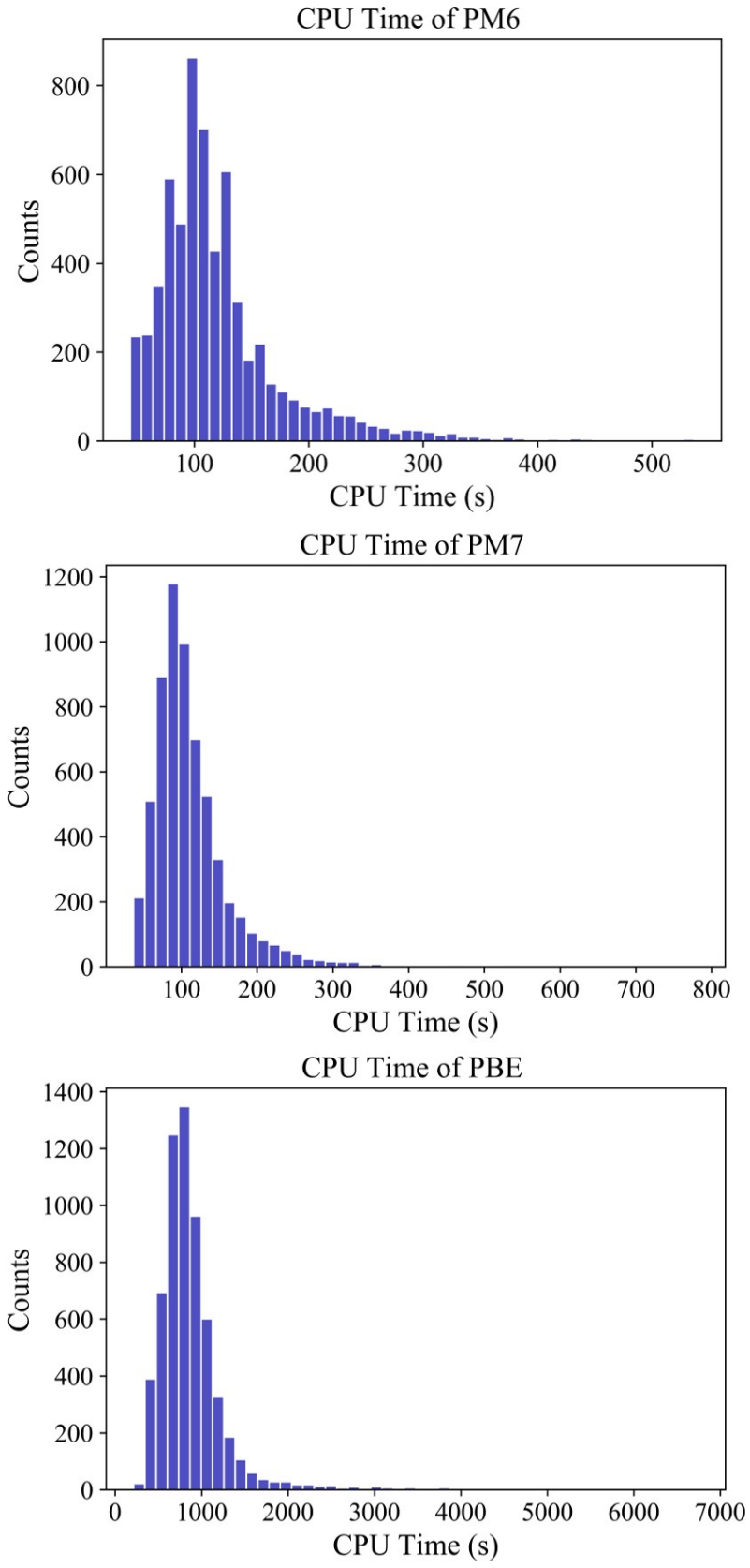

Fig. S5 The distributions of CPU time of each geometry optimization at PM6, PM7 and PBE/def2-SV(P) levels.

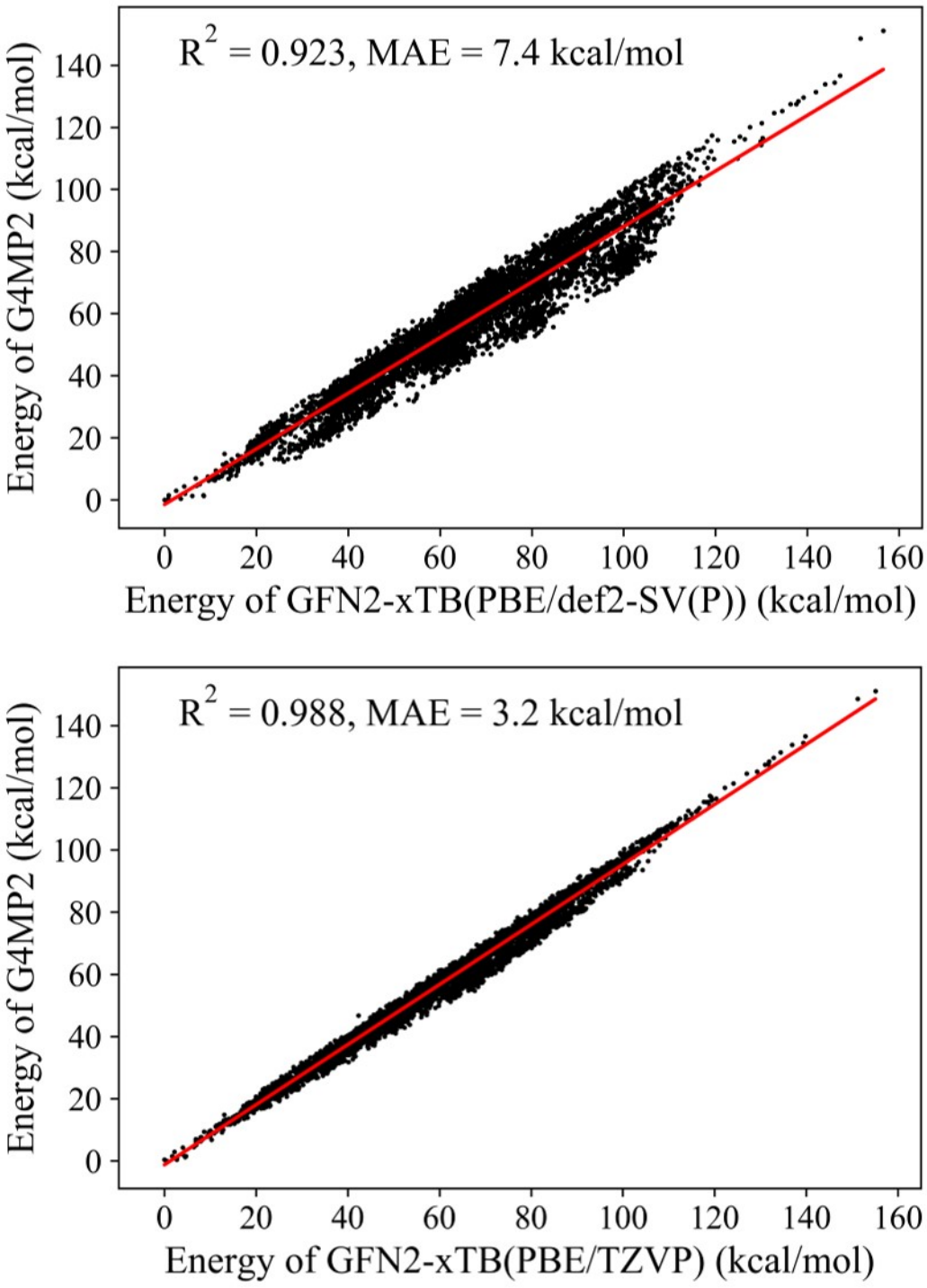

Fig. S6 Single-point energy correlations between G4MP2 and GFN2-xTB(PBE/def2-SV(P)) or GFN2-xTB(PBE/TZVP) method.

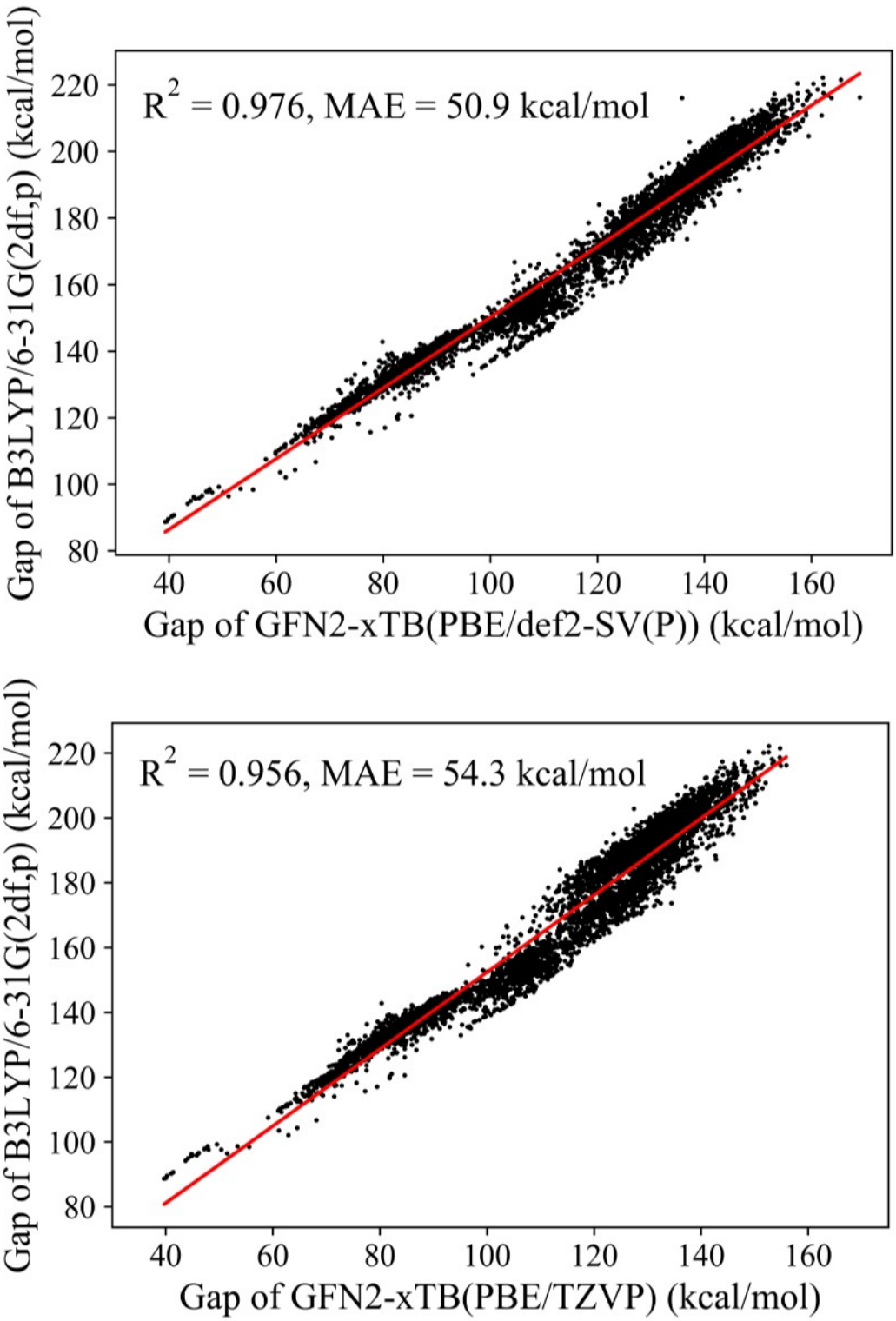

Fig. S7 Correlations of single-point gaps between B3LYP/6-31G(2df,p) and GFN2-xTB(PBE/def2-SV(P)), or GFN2-xTB(PBE/TZVP) method.

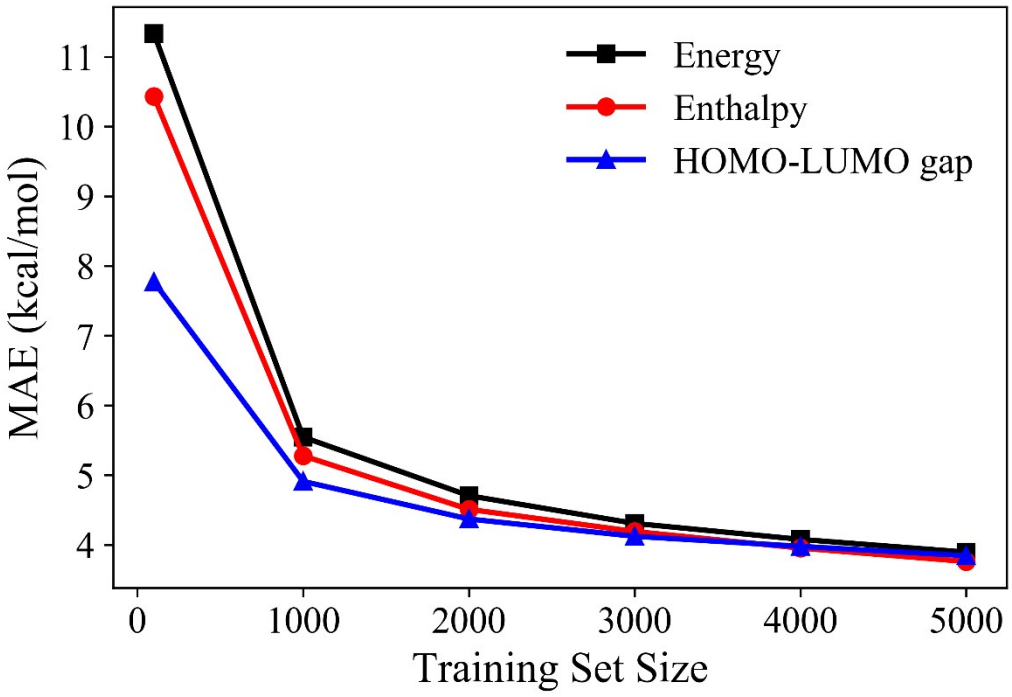

Fig. S8 The MAEs of the energies, enthalpies and HOMO-LUMO gaps of the dataset with different numbers of atoms.