# Supporting Information

# Dataset Construction to Explore Chemical Space with 3D Geometry and Deep Learning

*Jianing Lu[1,3] , Song Xia[1,3] , Jieyu Lu[1] , and Yingkai Zhang[1,2*]*

[1]Department of Chemistry, New York University, New York, New York 10003, United States

[2]NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China

[3]These authors contributed equally

*To whom correspondence should be addressed.
E-mail: yingkai.zhang@nyu.edu

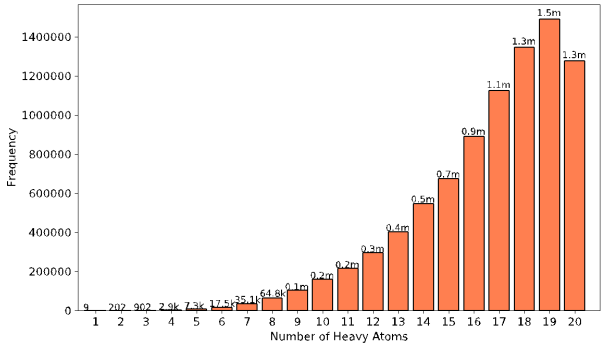


**Figure S1**. The Distribution of Molecules after Murcko Fragmentation for Mol20.

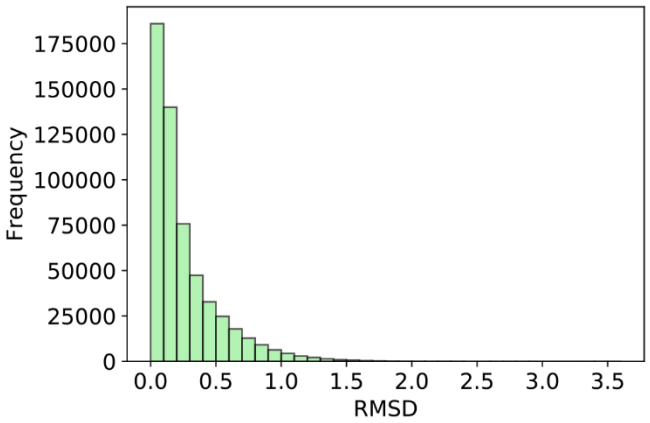


**Figure S2**. The Distribution of RMSD between MMFF Optimized Geometry and DFT Optimized Geometry for Frag20.

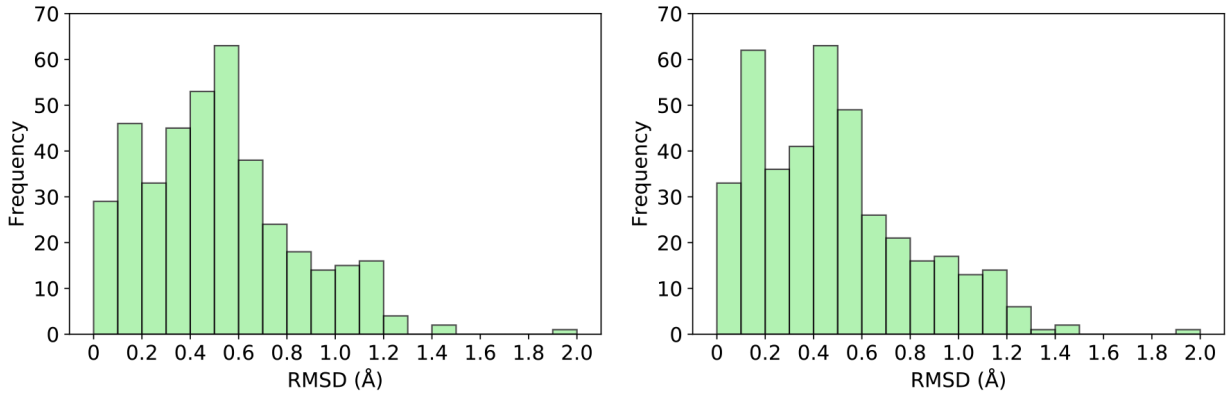


**Figure S3.** The Distribution of Smallest RMSD between Crystal Structure of Protein-Bound Ligand Conformation and (Left) MMFF Optimized Geometry and (Right) DFT Optimized Geometry of Generated Conformations for Molecules in Plati20.

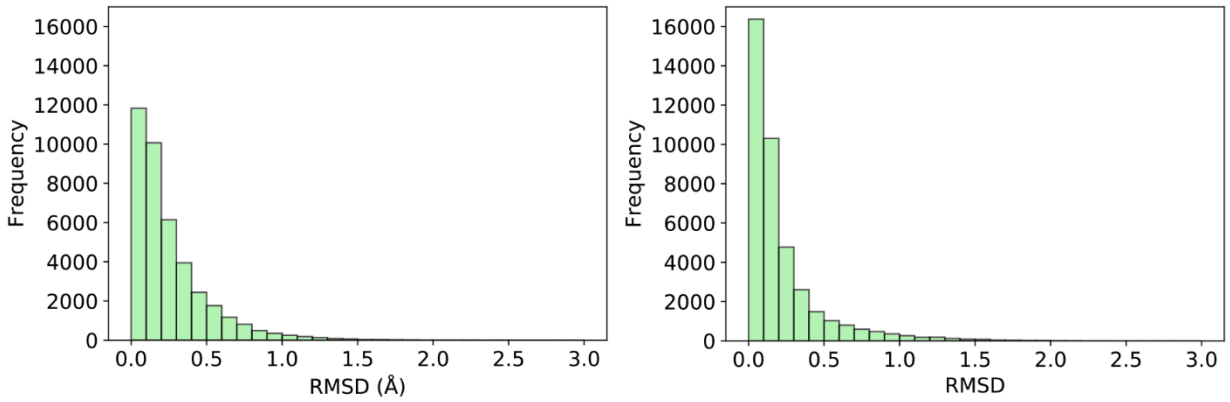


**Figure S4.** The Distribution of RMSD between Crystal Structure and (Left) MMFF Optimized Geometry and (Right) DFT Optimized Geometry in CSD20.

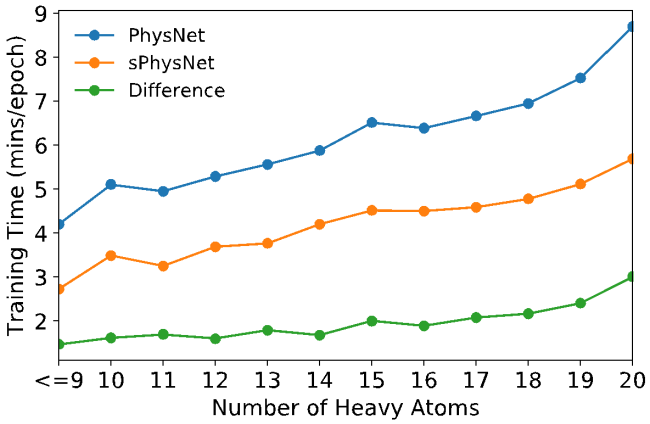


**Figure S5**. Training Time per Epoch for PhysNet and sPhysNet Model. Here, the training time is the average of 10 epochs for both models using same dataset on Tesla K80 GPU. The training time of PhysNet, sPhysNet, and the difference between PhysNet and sPhysNet has been colored as blue, orange and green, respectively.

**Table S1**. Frag20 Dataset.

| #Heavy Atoms | #Initial[a] | #Fragments[b] | #Selection[c] | #Final[d] |
|---|---|---|---|---|
| <=9 | 301051 | 232185 | 232185 | 158535 |
| 10 | 310649 | 161237 | 161237 | 143180 |
| 11 | 527364 | 216179 | 21542 | 17269 |
| 12 | 890659 | 295573 | 26663 | 21502 |
| 13 | 1537542 | 402662 | 32168 | 25793 |
| 14 | 2591635 | 547207 | 38100 | 30331 |
| 15 | 4119711 | 673725 | 40098 | 31719 |
| 16 | 6718088 | 890190 | 44565 | 35584 |
| 17 | 10563669 | 1125931 | 45306 | 36201 |
| 18 | 16033276 | 1346477 | 40814 | 32501 |
| 19 | 23085220 | 1489930 | 29459 | 23474 |
| 20 | 31770341 | 1277732 | 13125 | 10207 |
| <= 20 | 98449205 | 8659028 | 725262 | 566296 |

[a]The number of molecules in Mol20 dataset. [b]The number of molecules after molecule fragmentation. [c]The number of molecules after molecule selection. [d]The number of molecules in Frag20.

**Table S2**. Details of Hyperparameter Tuning. Hyperparameters that are different from the original PhysNet model is labeled in red. Models 1-15 were only trained 400 epochs for faster speed.

| Hyper-params\No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| # modules | 5 | 3 | 1 | 5 | 5 | 5 | 5 | 5 |
| # residuals (atomic) | 2 | 2 | 2 | 0 | 1 | 2 | 2 | 2 |
| # residuals (interaction) | 3 | 3 | 3 | 0 | 1 | 3 | 3 | 3 |
| # residuals (output) | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| # output blocks | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Activations | SSF[a] | SSF[a] | SSF[a] | SSF[a] | SSF[a] | SSF[a] | SSF[a] | SSF[a] |
| Number of features (F) | 128 | 128 | 128 | 128 | 128 | 128 | 128 | 128 |
| Number of expansions (K) | 64 | 64 | 64 | 64 | 64 | 64 | 64 | 64 |
| Calculate coulomb | True | True | True | True | True | False | True | False |
| Calculate dispersion | True | True | True | True | True | True | False | False |
| Cutoff (Å) | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| # epochs | 400 | 400 | 400 | 400 | 400 | 400 | 400 | 400 |
| # of params (M) | 1.29 | 0.78 | 0.27 | 0.30 | 0.80 | 1.29 | 1.29 | 1.29 |
| Performance (kcal/mol) | 0.23 | 0.23 | 0.44 | 0.25 | 0.23 | 0.22 | 0.23 | 0.22 |

[a]Shifted Soft Plus.

Cont.

| Hyper-params\No. | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|
| # modules | 5 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| # residuals (atomic) | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
| # residuals (interaction) | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 1 |
| # residuals (output) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| # output blocks | 3 | 3 | 3 | 3 | 3 | 3 | 1 | 1 |
| Activations | SSF[a] | SSF[a] | SSF[a] | SSF[a] | SSF[a] | SSF[a] | SSF[a] | SSF[a] |
| Number of features (F) | 96 | 96 | 160 | 192 | 160 | 192 | 160 | 160 |
| Number of expansions (K) | 64 | 64 | 64 | 64 | 64 | 64 | 64 | 64 |
| Calculate coulomb | True | True | True | True | True | True | True | True |
| Calculate dispersion | True | True | True | True | True | True | False | False |
| Cutoff (Å) | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| Epochs | 400 | 400 | 400 | 400 | 400 | 400 | 400 | 1000 |
| # of params (M) | 0.74 | 0.45 | 1.21 | 1.80 | 0.74 | 1.10 | 0.74 | 0.74 |
| Performance (kcal/mol) | 0.23 | 0.25 | 0.23 | 0.22 | 0.23 | 0.22 | 0.23 | **0.19** |