Towards Physically Reliable Molecular Representation Learning (Supplementary Materials)

Seunghoon Yi[1] **Youngwoo Cho**[2] **Jinhwan Sul**[1] **Seung Woo Ko**[1]
Soo Kyung Kim[3] **Jaegul Choo**[2] **Hongkee Yoon**[*2] **Joonseok Lee**[*1,4]

[1]Seoul National University, Seoul, Korea
[2]Korea Advanced Institute of Science and Technology, Daejeon, Korea
[3]Palo Alto Research Center, Stanford Research Institute, Palo Alto, CA, USA
[4]Google Research, Mountain View, CA, USA

A IMPLEMENTATION DETAILS

We try $L \in \{4, 6, 8\}$ to stack Molecule Attention Blocks after the embedding layer. We set the embedding size $d = 256$, which is same as (number of heads) × n_b. Here, n_b is the same as the dimension of the query, key, and value in the attention block. For activation, we use LeakyRELU [Nair and Hinton, 2010, Sun et al., 2015] function after f_{mol} and ELU [Clevert et al., 2016] after f_{bond}. To enforce the positive base and exponents in the parameterized LJP and to avoid numerical errors, we add $1 + \epsilon$ to β_3, β_4, where ϵ is set to be 10^{-3}. We set the cutoff threshold $\tau = 5\text{Å}$, and the number of RBFs $n_b = 16$. We use a single linear layer for f_{atom} and f_{bond}, while a two-layer MLP for the MAM task. Specifically, the MLP outputs the estimated likelihood score for 64 atoms for each masked input token. For the overall objective function, we choose weights as $\lambda_{\text{force}} = 0.3$, $\lambda_{\text{mask}} = 0.7$, and $\lambda_{\text{bound}} = 1$. The $\beta_{z_i,k}$ and $\mu_{z_i,k}$ are initialized to $(2n_b^{-1}(1 - \exp(-\tau))^{-2}$ and uniformly within $[0, 1]$, respectively.

For training, we use a learning rate of 5×10^{-4} with Adam optimizer [Kingma and Ba, 2015]. We warm-up for 10 epochs, linearly increasing the learning rate, and we decay the learning rate with the ratio of 0.6 and patience of 24. The minimum learning rate is set to 10^{-7}. We train the model for up to 900 epochs.

For transfer learning experiment on Transition1x, we pre-train a model with $L = 6$ on QM9 dataset. The cutoff thereshold is set to $\tau = 7.5\text{Å}$, while other hyperparameters are set the same as the above.

B ADDITIONAL ABLATION STUDY

We conduct an additional ablation study with varied number of layers. Tab. I shows that the $\mathbf{A}$-mask we introduce in Fig. 1 indeed helps in most cases. Also, we observe that using more MABs up to 8 tends to improve the overall performance.

Layers	4 (Base)		6 (Large)		8 (Huge)	
Method	MAE_E	MAE_F	MAE_E	MAE_F	MAE_E	MAE_F
Base	11.86	0.91	11.83	0.77	11.33	0.72
+ [CLS]	11.70	0.78	9.03	0.90	9.70	0.78
+ $\mathbf{A}$-mask	9.89	0.98	9.55	1.33	9.33	0.88
+ MAM	10.77	1.43	9.38	1.27	8.35	1.28

Table I: Ablation study on SSL methods with different number of layers

We also search the mask ratio of our MAM task in Tab. II. We observe that using a mask ratio of 0.3 is clearly better than others in terms of both energy prediction and a reasonable PES.

Masking ratio	MAE_E	MAE_F	ΔP
0.1	16.18	0.0056	0.028
0.15	15.82	0.0060	0.028
0.2	16.77	0.0057	0.029
0.3	**15.16**	**0.0050**	**0.025**
0.5	17.73	0.0066	0.032

Table II: Ablation study on masking ratio

C ADDITIONAL EXAMPLES

Reaction barrier estimation. We evaluate the entire Transition1x reaction barrier estimation task by calculating and comparing the reaction barrier task with the ground truth across 225 reaction paths. Our method shows reasonable results on 212 of them, with a mean absolute error (MAE) less than 0.2 eV on average. These results are presented in Fig. II.

Structure optimization. We report additional structural optimization results of random molecules in the QM9 dataset in Fig. III. We observe that our model and TorchMDNet (ET) mostly preserve the optimal structure, while other baselines significantly destroy structures. In addition, we present re-

[*]Corresponding authors

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

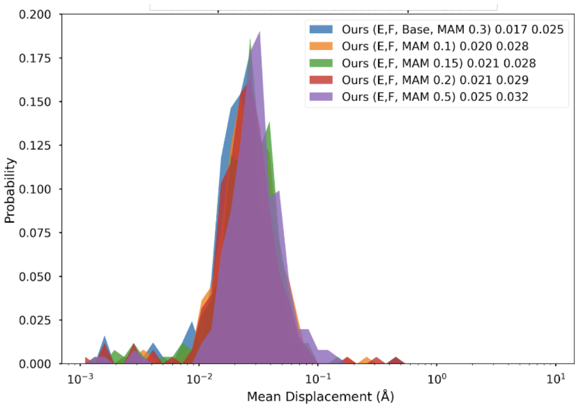

Figure I: Additional structural optimization results by different MAM making ratios.

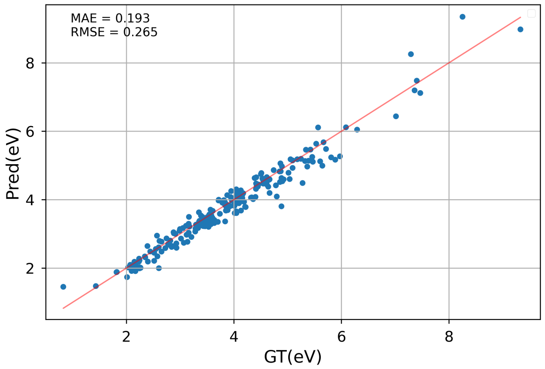

Figure II: Estimated reaction barrier along the reaction pathways of Trainsition1x dataset. The ground truth barriers are on the x-axis, and those estimated by our model are on the y-axis, in eV scale.

laxation results from 102 molecules in Fig. IV–XII. We list results from other baselines and the GT structure(Ref.). Blanks are failed results.

REFERENCES

D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *Proc. of the International Conference on Learning Representations (ICLR)*, 2016.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2015.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. of the International Conference on Machine Learning (ICML)*, 2010.

Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015.

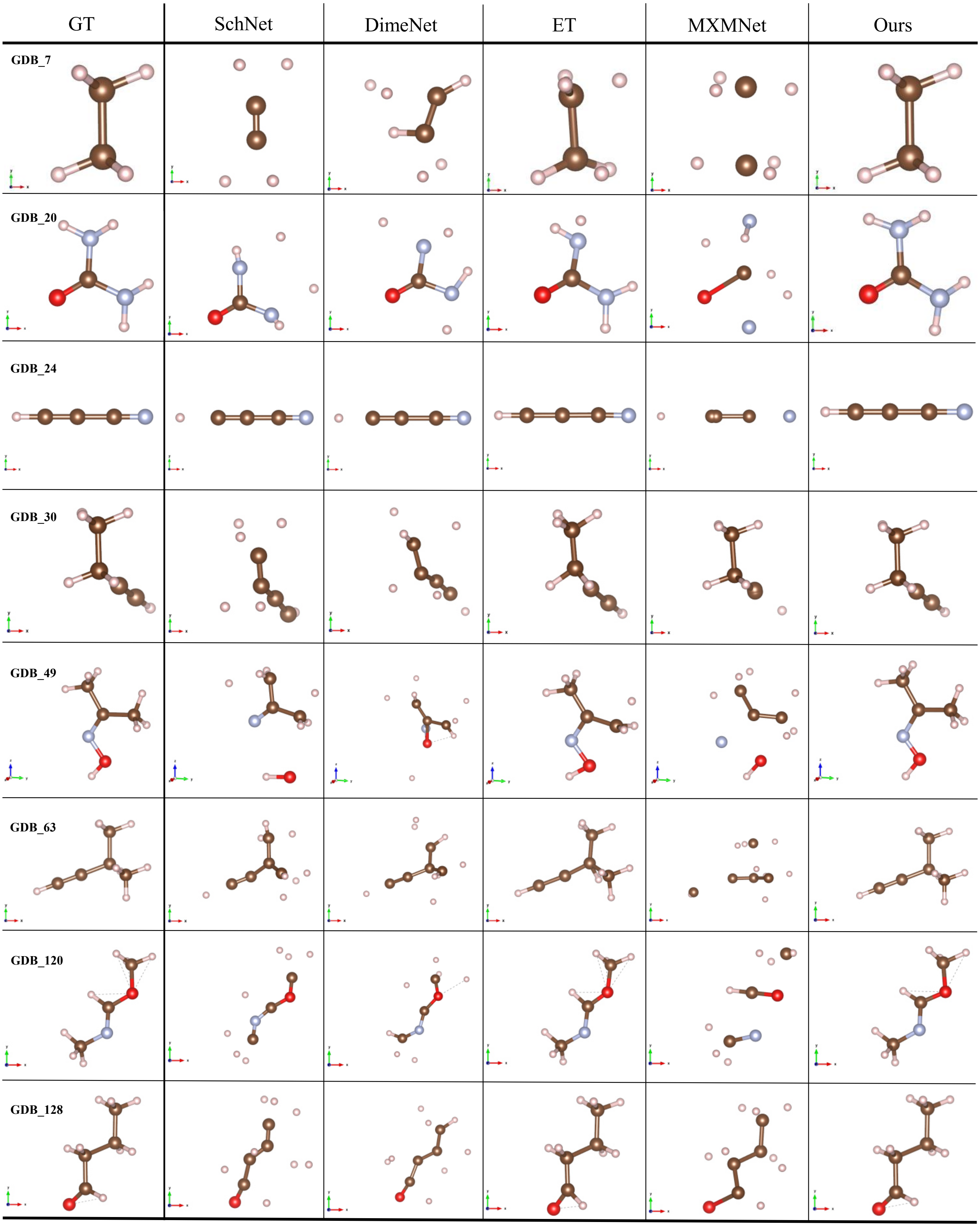

Figure III: Additional structural optimization results by ours and baselines.

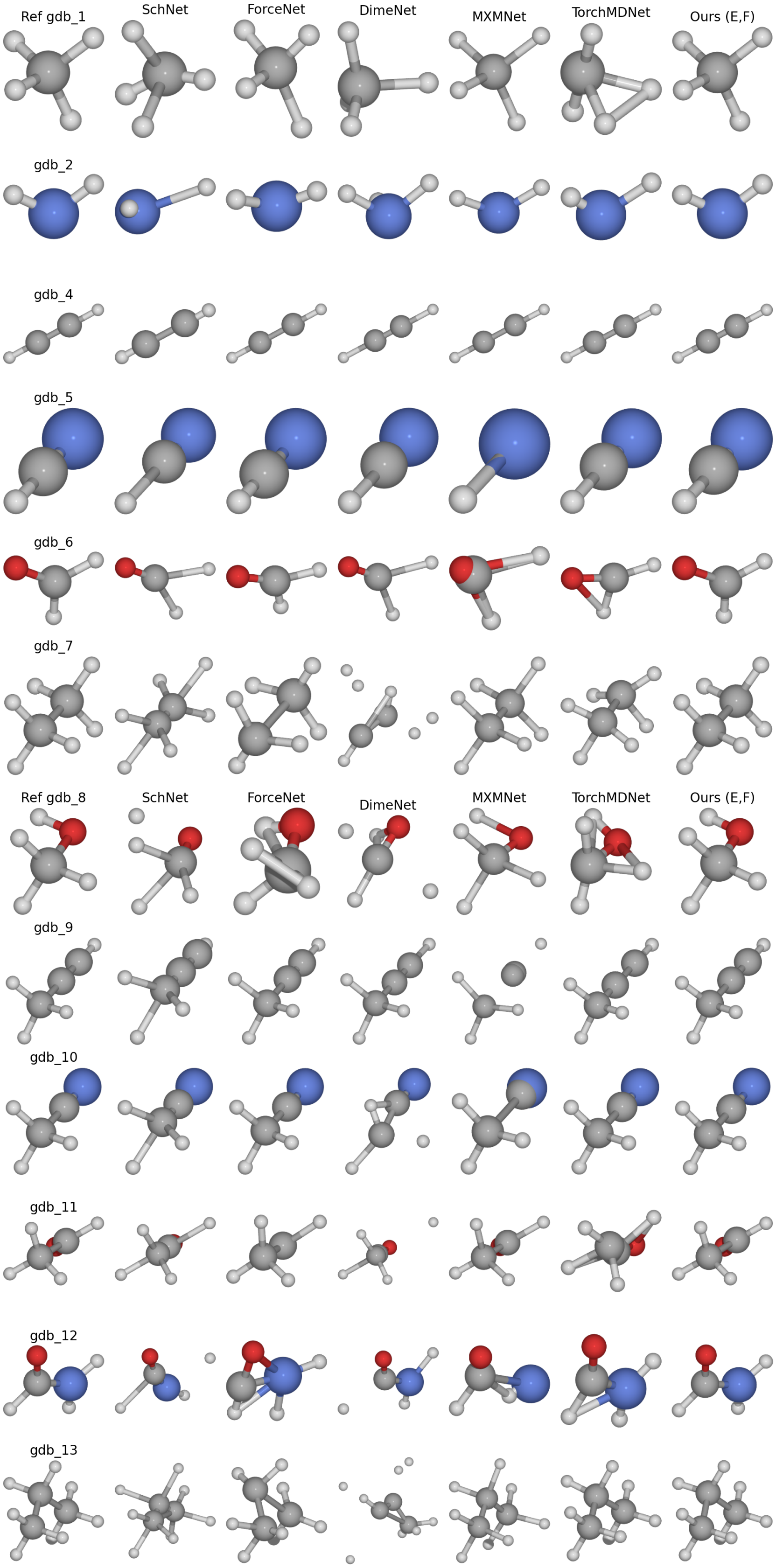

Figure IV: Additional structural optimization results (1/9)

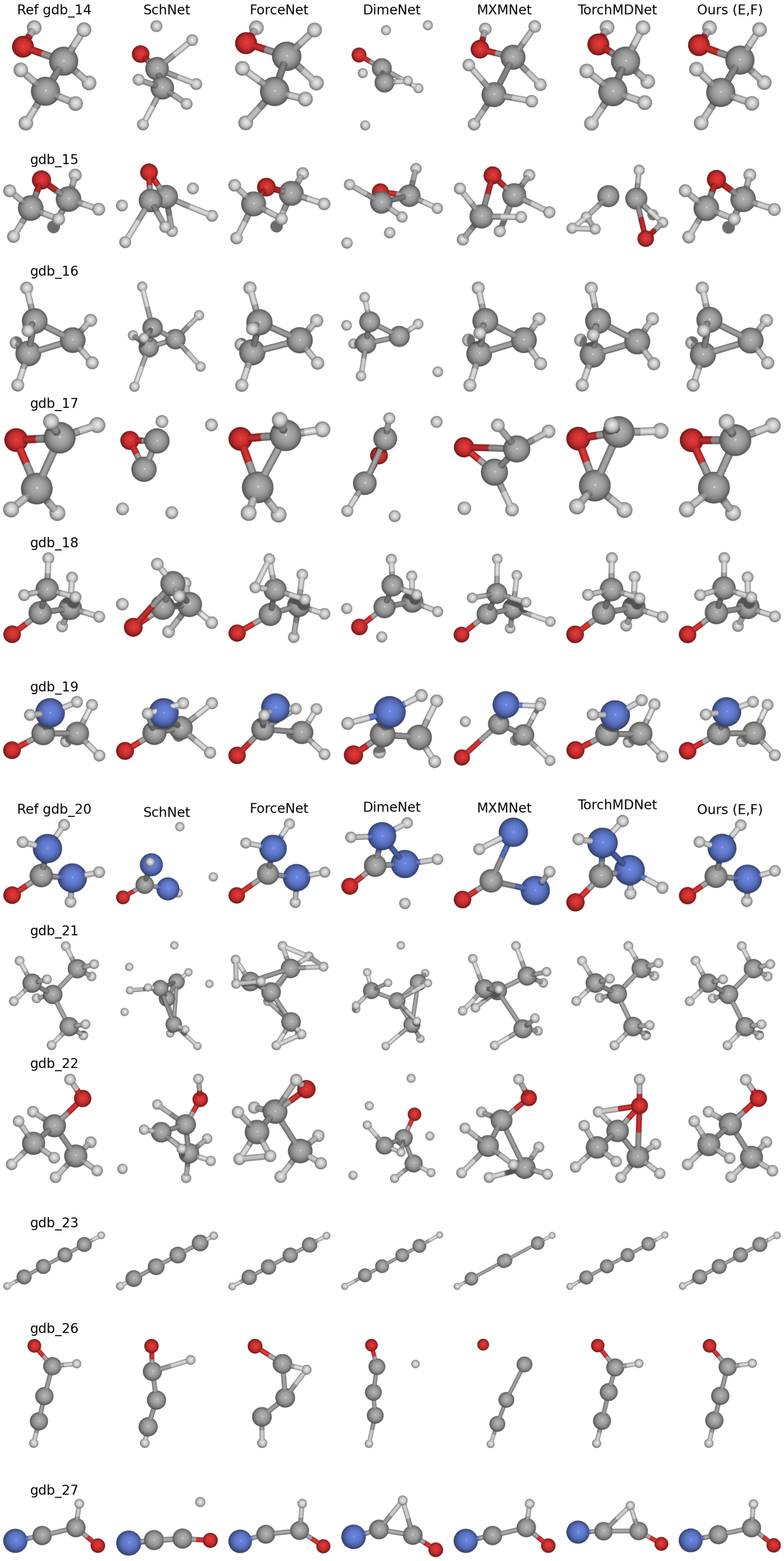

Figure V: Additional structural optimization results (2/9)

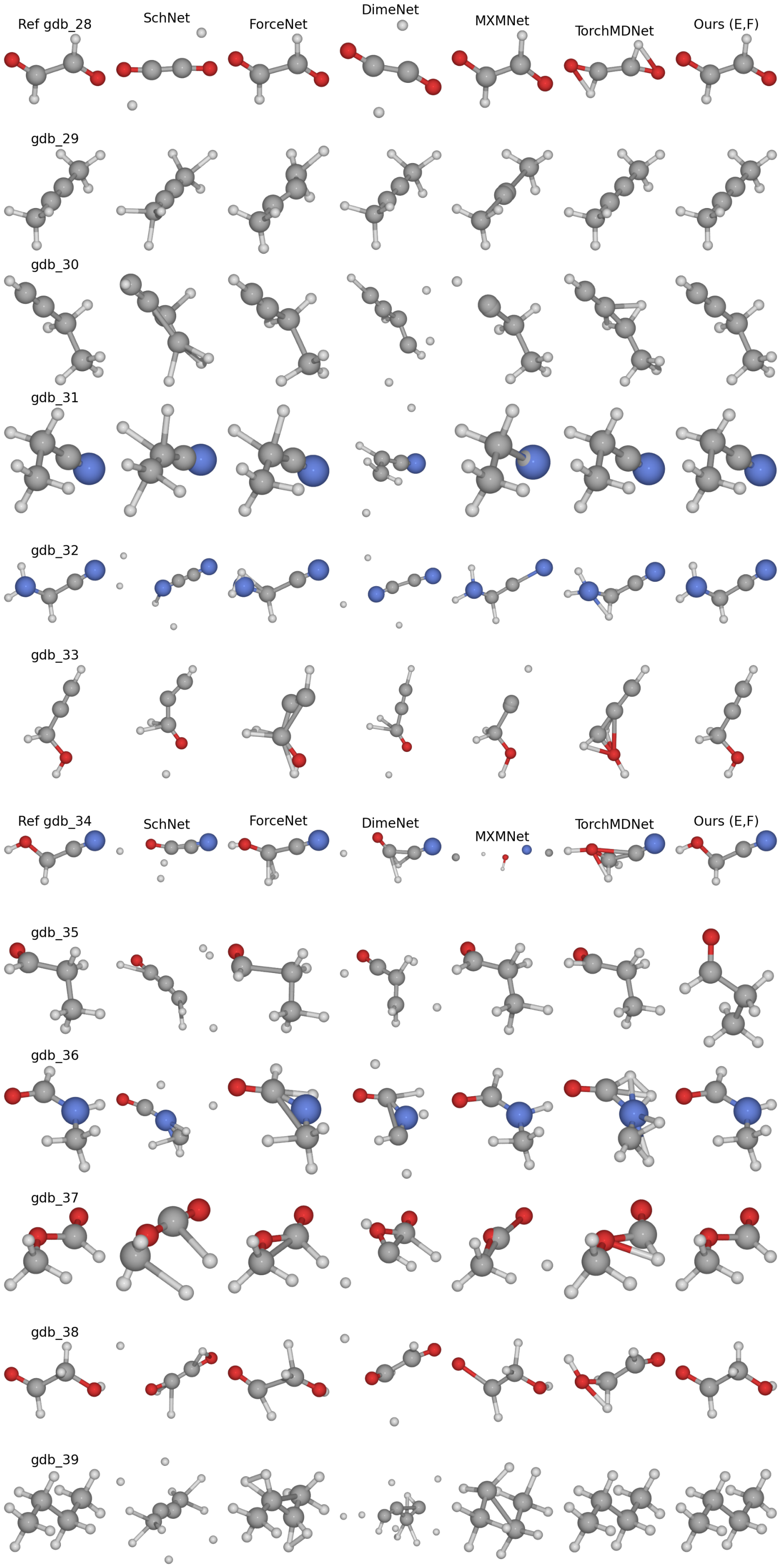

Figure VI: Additional structural optimization results (3/9)

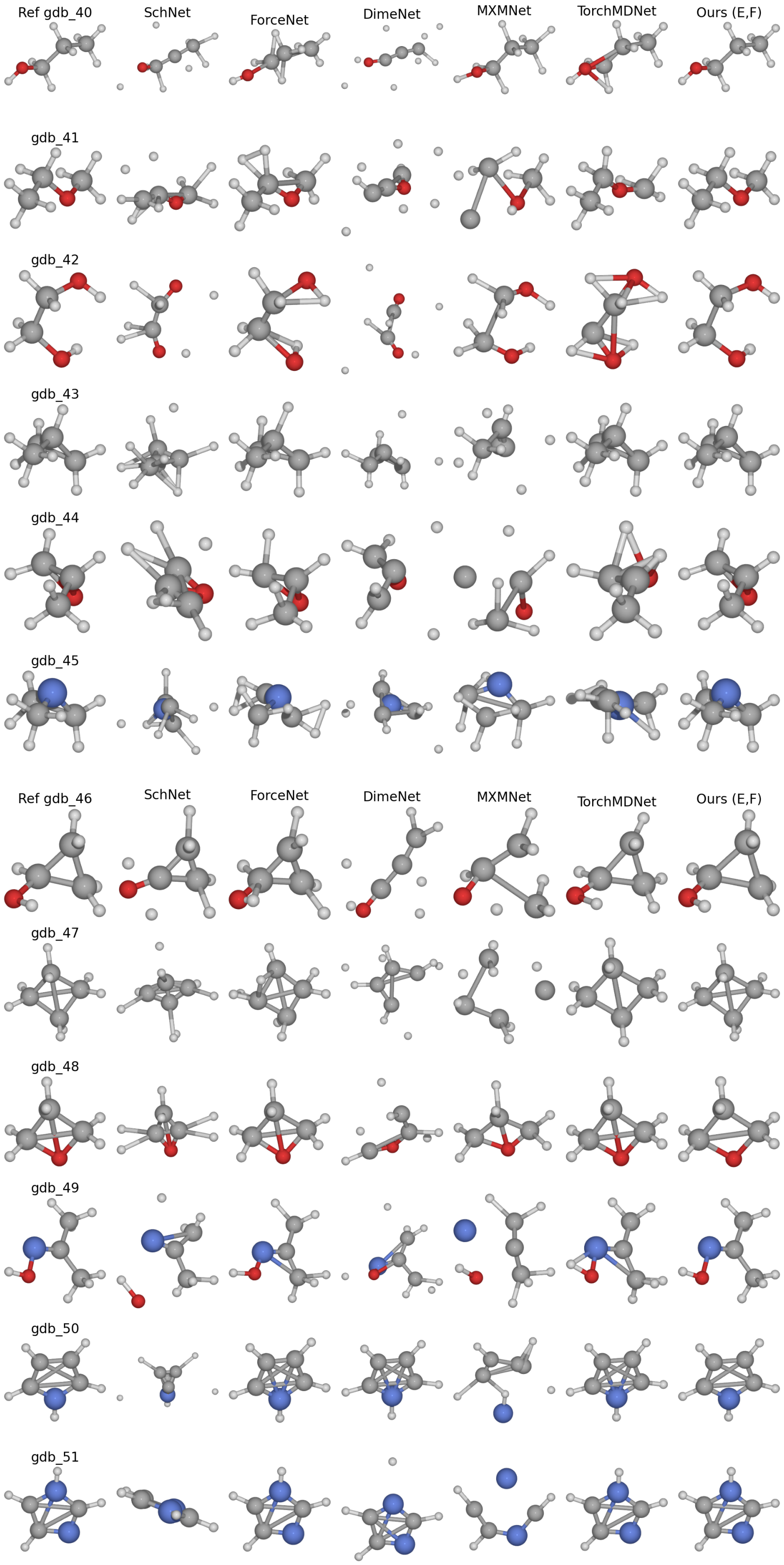

Figure VII: Additional structural optimization results (4/9)

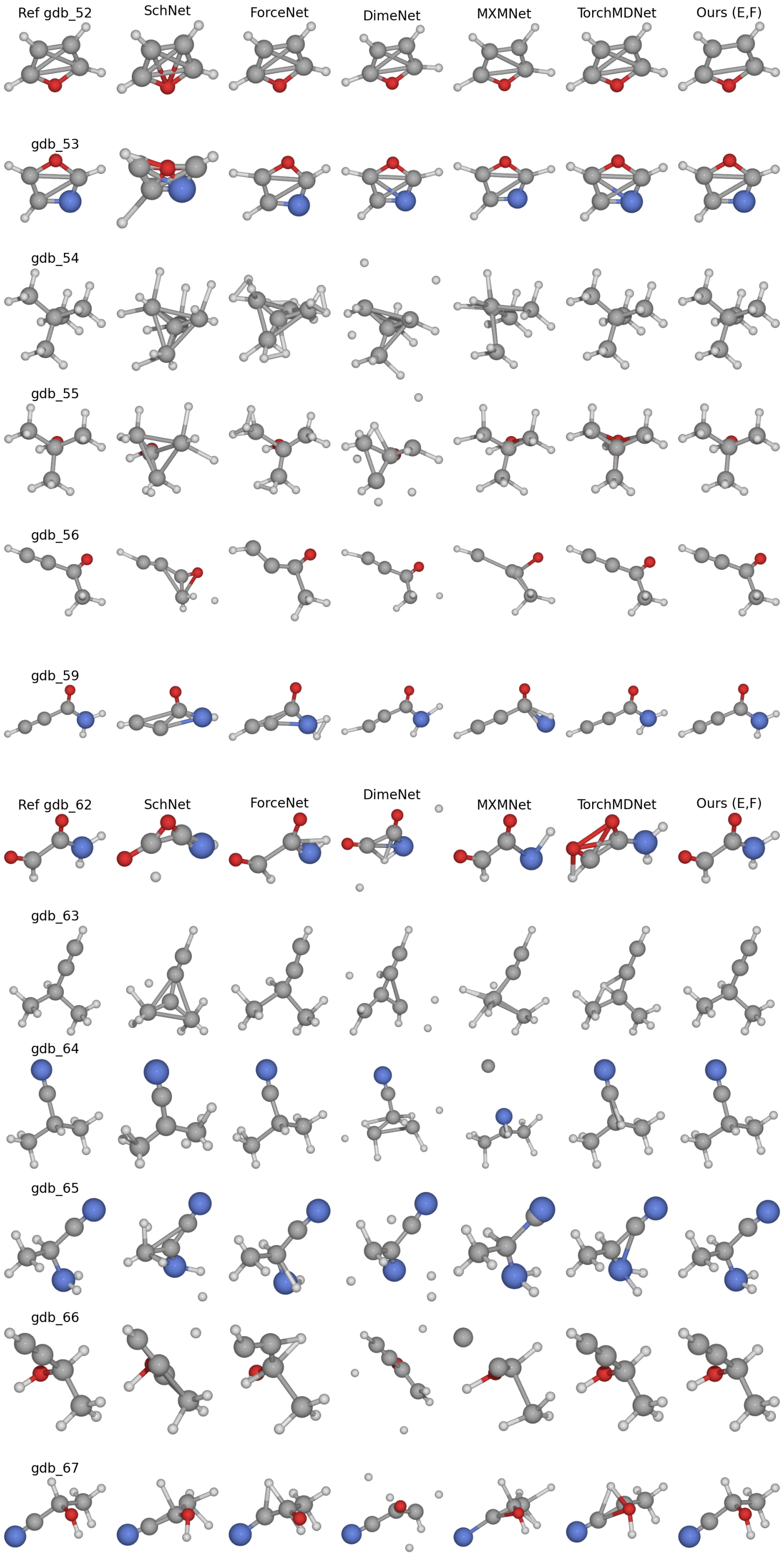

Figure VIII: Additional structural optimization results (5/9)

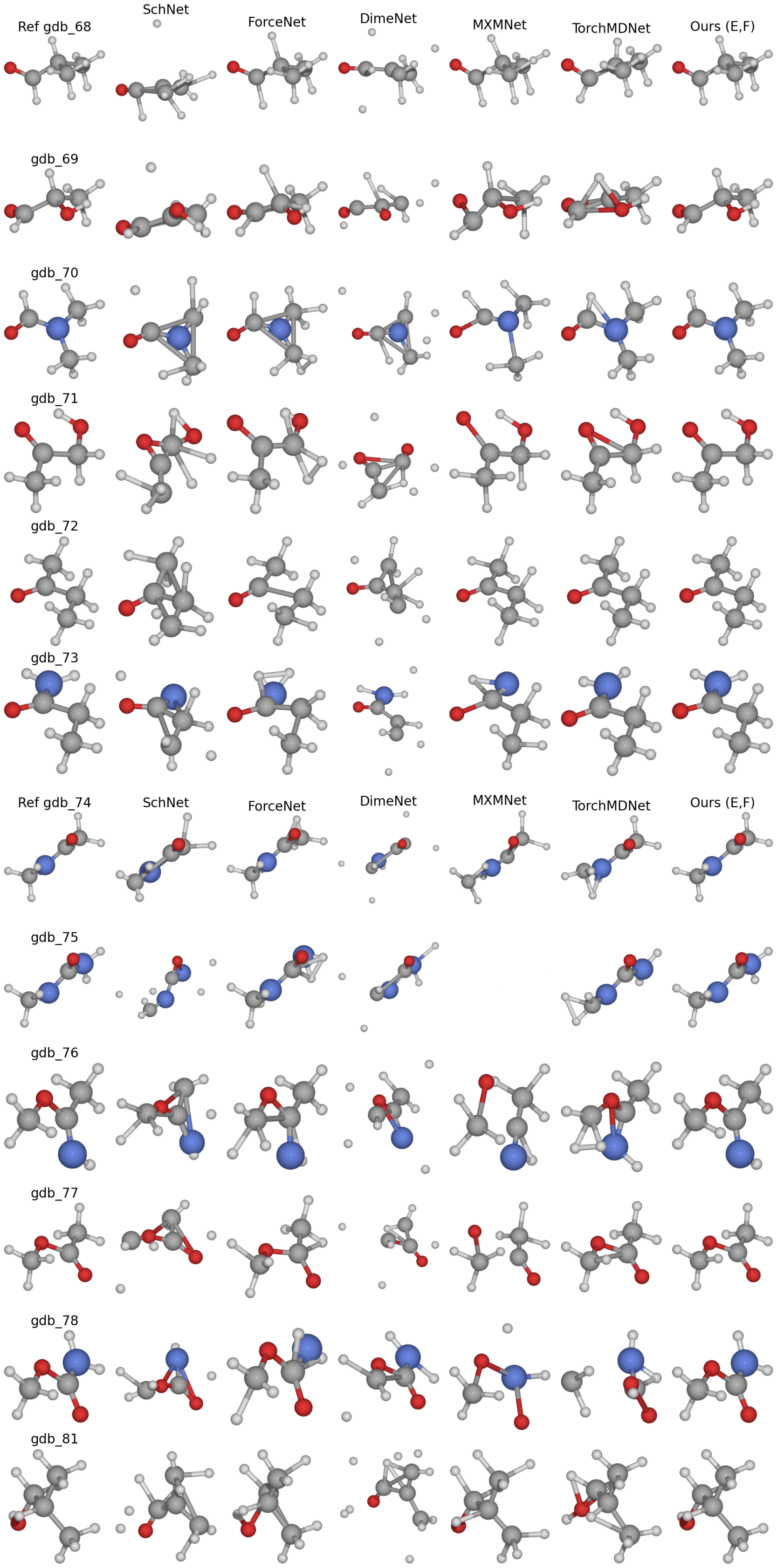

Figure IX: Additional structural optimization results (6/9)

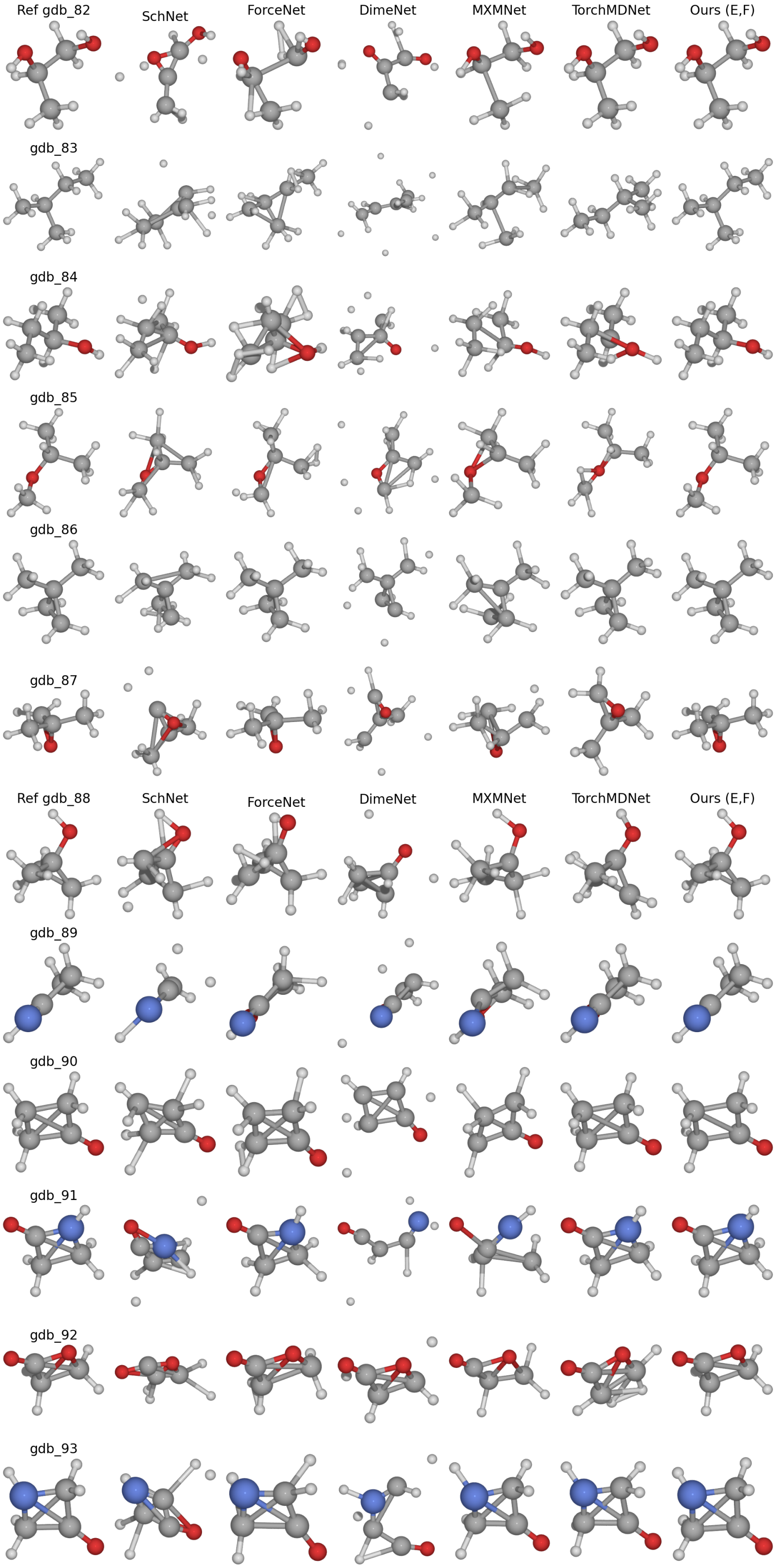

Figure X: Additional structural optimization results (7/9)

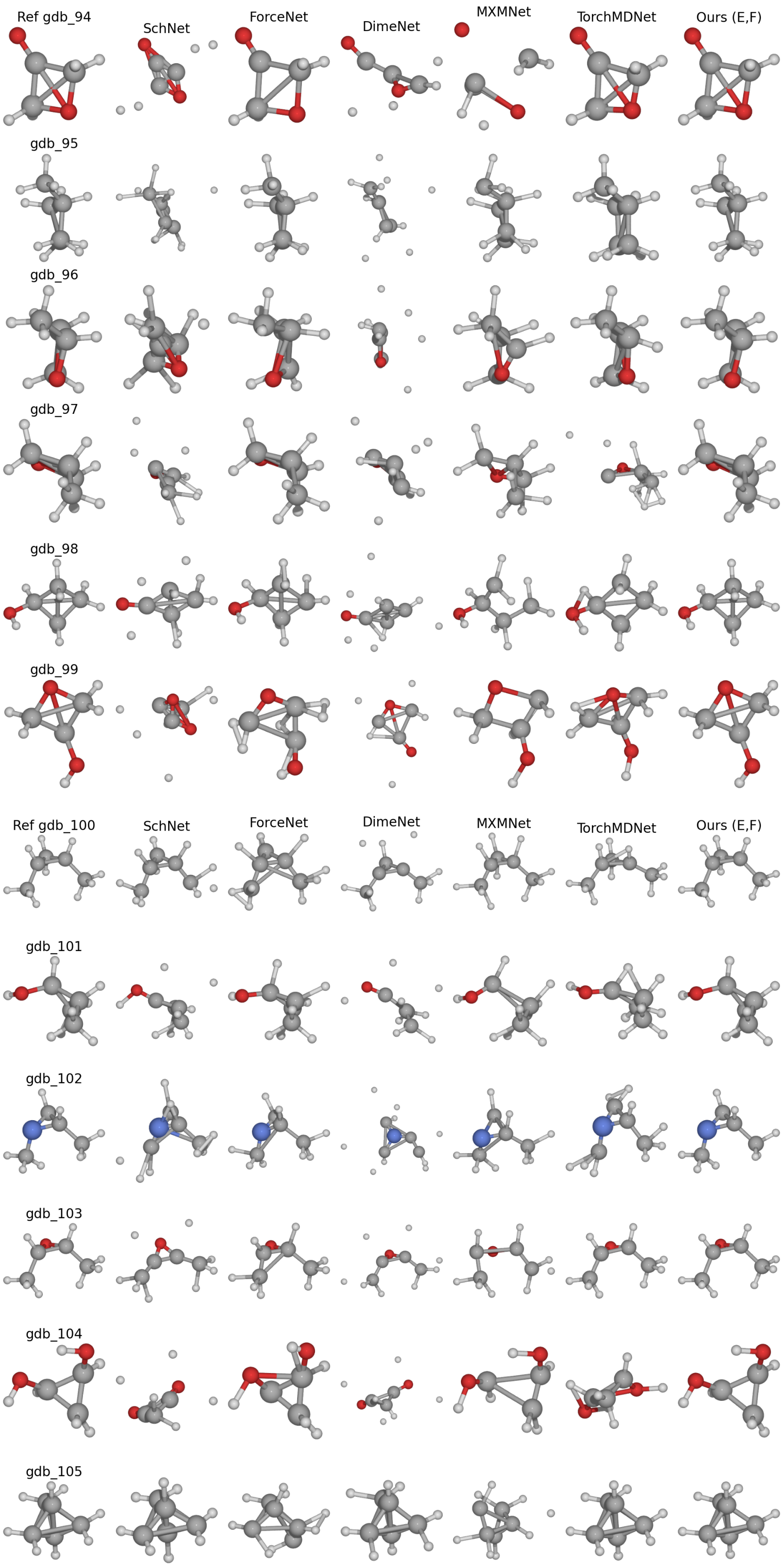

Figure XI: Additional structural optimization results (8/9)

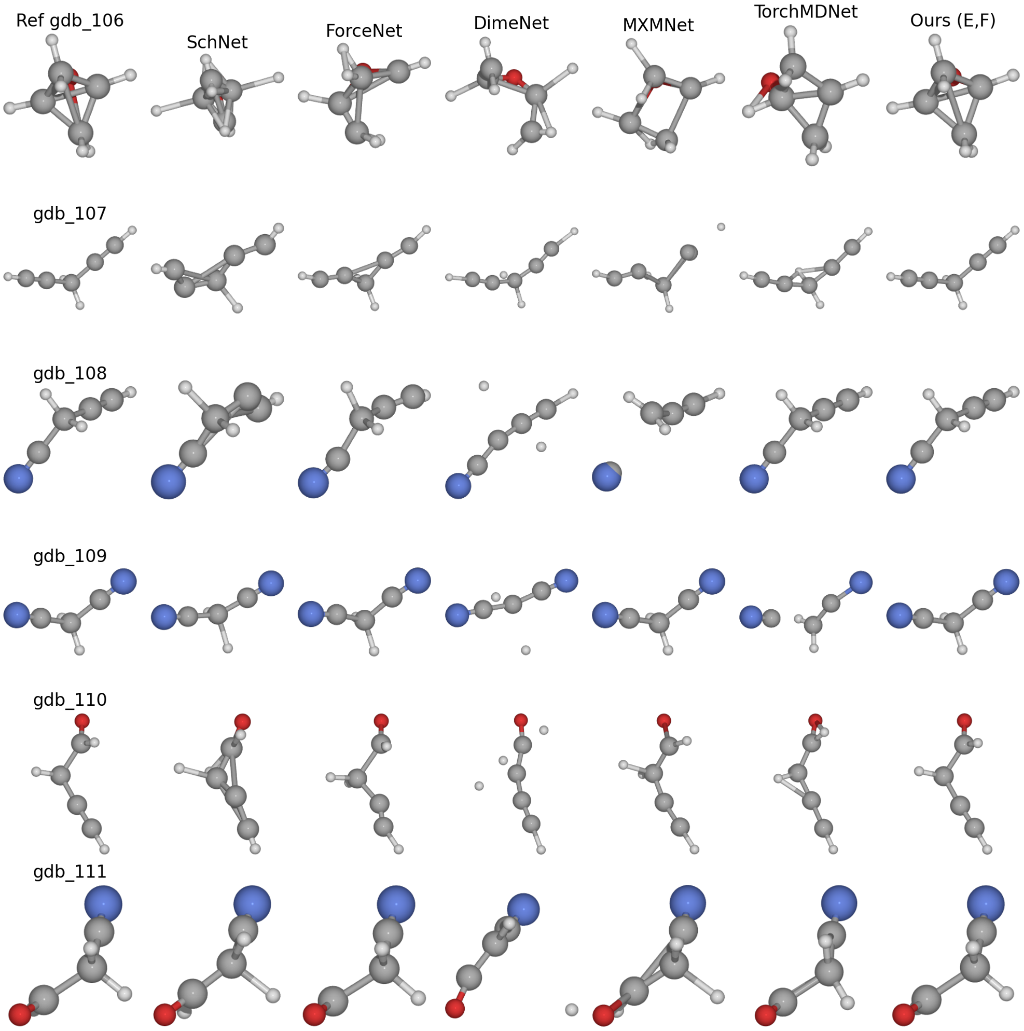

Figure XII: Additional structural optimization results (9/9)