APPENDIX

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) $\times\, n_b$. 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 & Hinton, 2010; Sun et al., 2014) function after f_{mol} and ELU (Clevert et al., 2015) 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. We set the cutoff threshold $\tau = 5$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 $\lambda_{\text{force}} = 0.2$, $\lambda_{\text{mask}} = 0.7$, 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. In the training phase, we used a learning rate of 5×10^{-4} with ADAM optimizer (Kingma & Ba, 2014). We use a warmup of 20 epochs, total training epochs of 900, a decay patience of 24 with a ratio of 0.6. The minimum learning rate was set to 10^{-7}.

B ADDITIONAL ABLATION STUDY

We conduct an additional ablation study with varied number of layers. Tab. I shows that the **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
+ **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 QUALITATIVE ANALYSIS

Self-supervised Learning with MAM. Fig. II illustrates the effect of self-supervised learning with MAM, depending on the position of atoms. For example, Fig. II (a) shows the example of CH_4, where we perform MAM inference to figure out an appropriate atom type through the vertical direction. Fig. II (b) shows the inferred atom type at each position, from atomic number 1 to 14. The atoms that the QM9 covers, H, C, N, O, and F, are marked in the figure.

Fig. II (b) shows that around ± 2Å from the center, the Carbon is strongly favored. On the other hand, Fluorine (F), which is not completely chemically favored, MAM shows a very low affinity. The Nitrogen and Carbon of C_4NH_5 also show a similar trend as shown in Fig. II (c-e). In Fig. II (e), Carbon is favored by MAM as expected, and interestingly, Nitrogen is also weakly favored, unlike

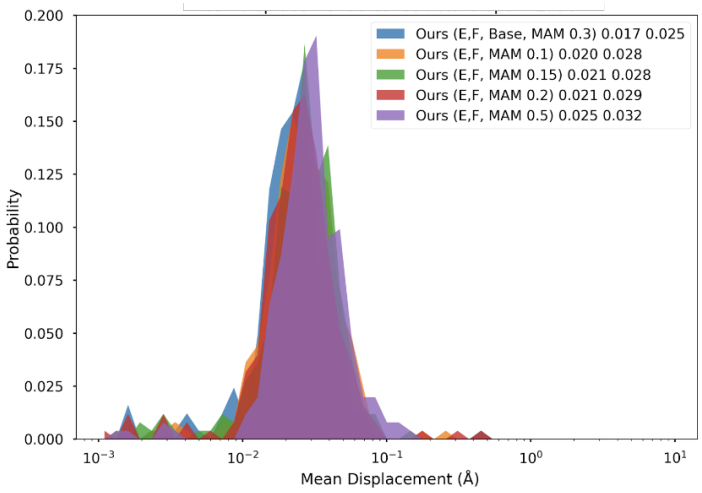

Figure I: Additional structural optimization results by different MAM making ratios.

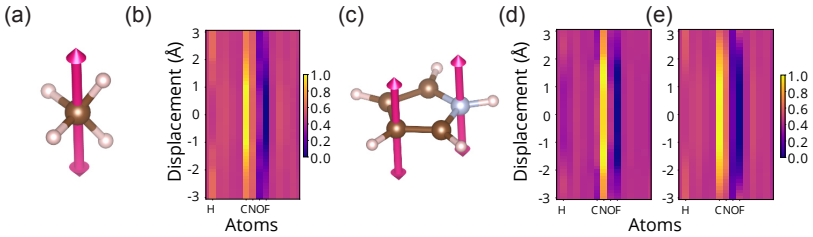

Figure II: Visualization of MAM. (a), (c) The masked atom is moved along the pink arrow (z-axis), and (b), (d-e) illustrate likelihood score along corresponding positions.

CH_4. Presumably, it is due to the shape of the C_4NH_5 molecule. Note that the amplitude of the atom recommendation through MAM is maximized at the most stable energy position. This reveals that the model self-learns the relationship between surrounding atoms from energy and the positions through MAM. In molecule generation tasks, MAM would be more efficient than randomly connecting atoms and repeating structural optimization iteratively.

Physics-driven Modeling and Regularization. We design our model to predict the parameters of a physics-inspired equation (Sec. 3.4). In Eq. (5), if both β_2 and β_3 are a finite number greater than 0, this implies that the equation is fitted to the distance. In particular, having $\beta_3 \approx 6$ indicates that it has similar behavior to the LJ potential. Since we have two freedoms of Coulomb's terms and LJP-like terms, there is no reason to converge to a single β_3; based on training, β_3 seen to be distributed between 4 and 16, which is close to the 6 of LJ potential.

D Additional Examples

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

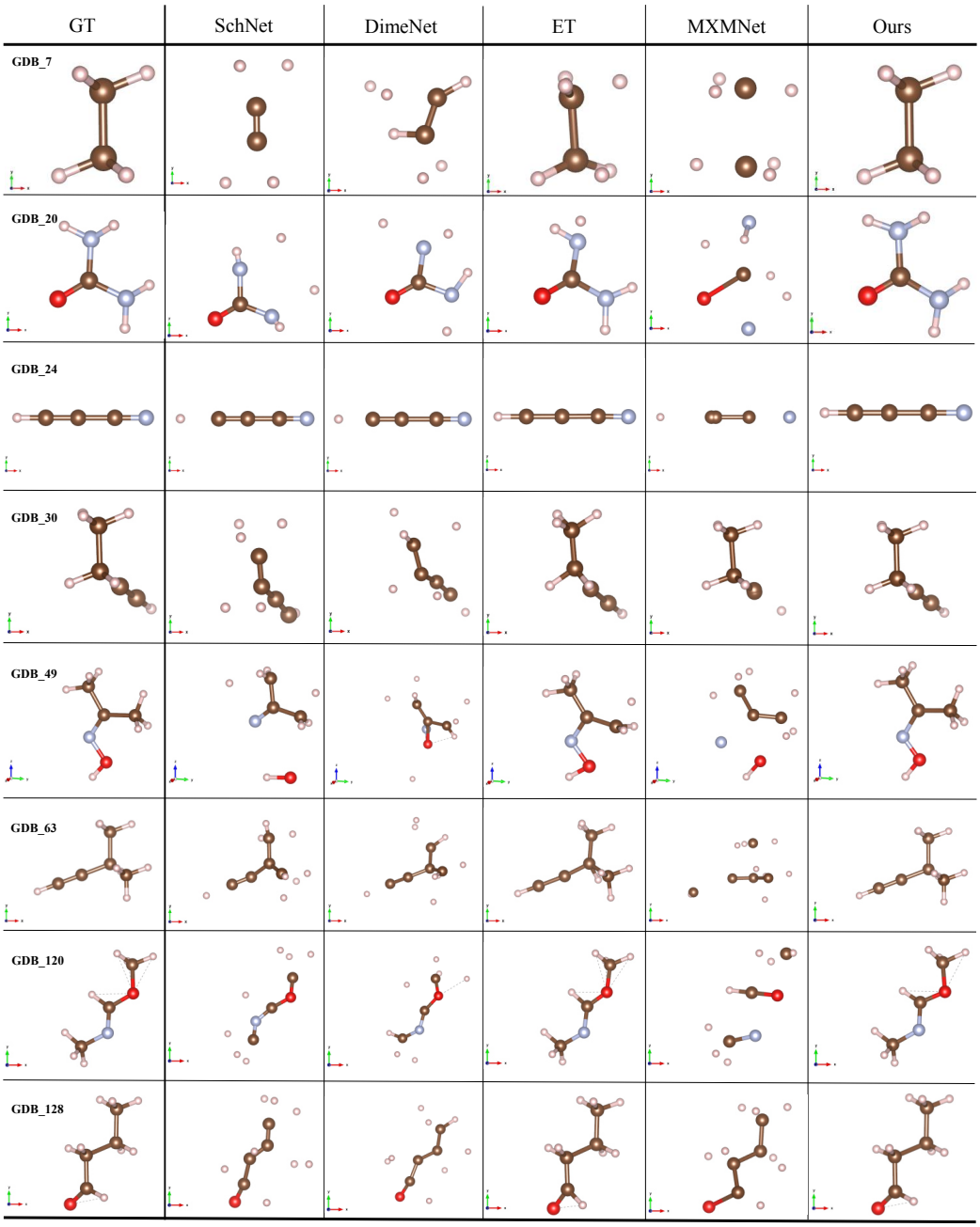

Figure III: Additional structural optimization results by ours and baselines.

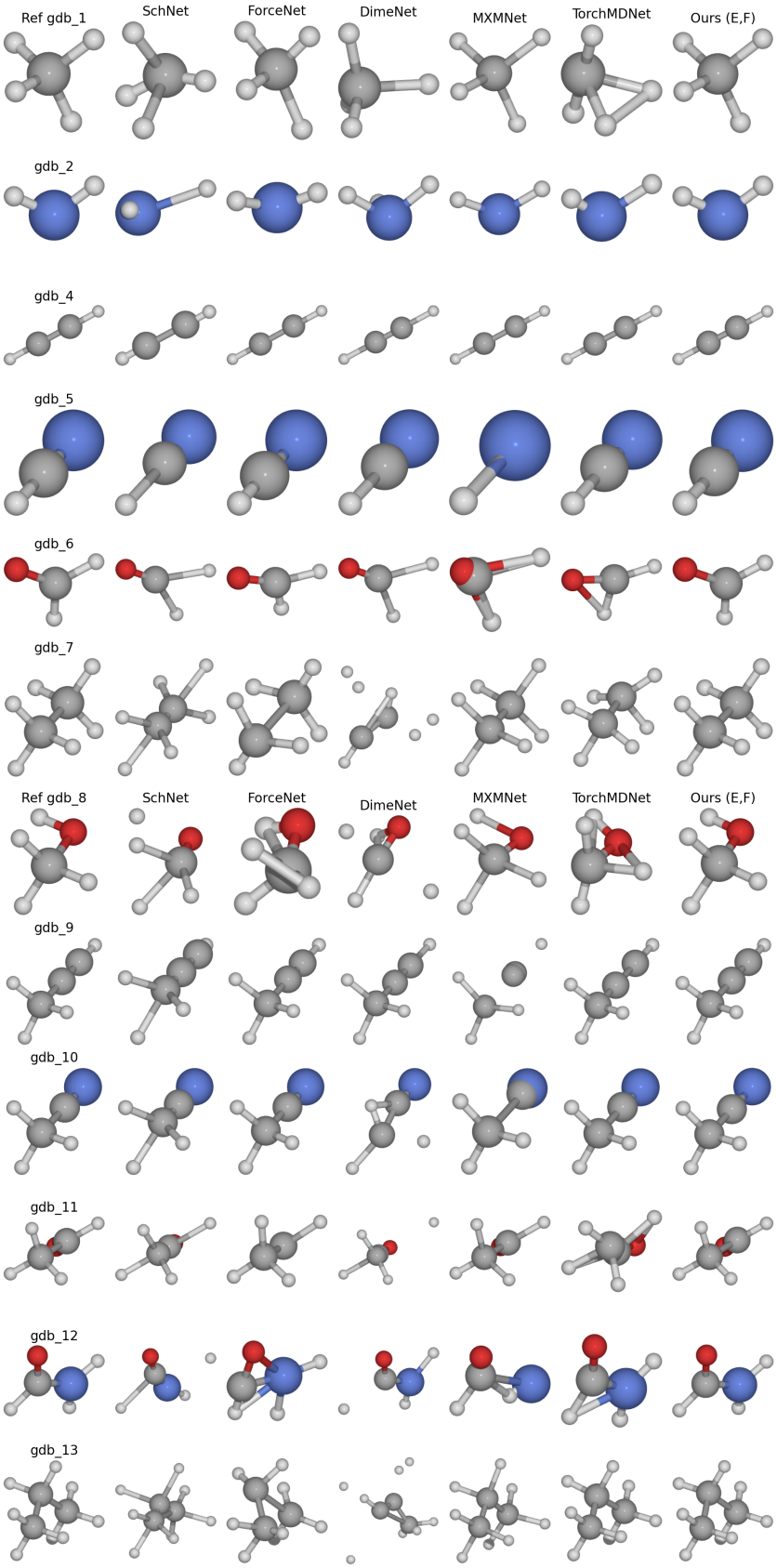

Figure IV: Additional structural optimization results (1/9)

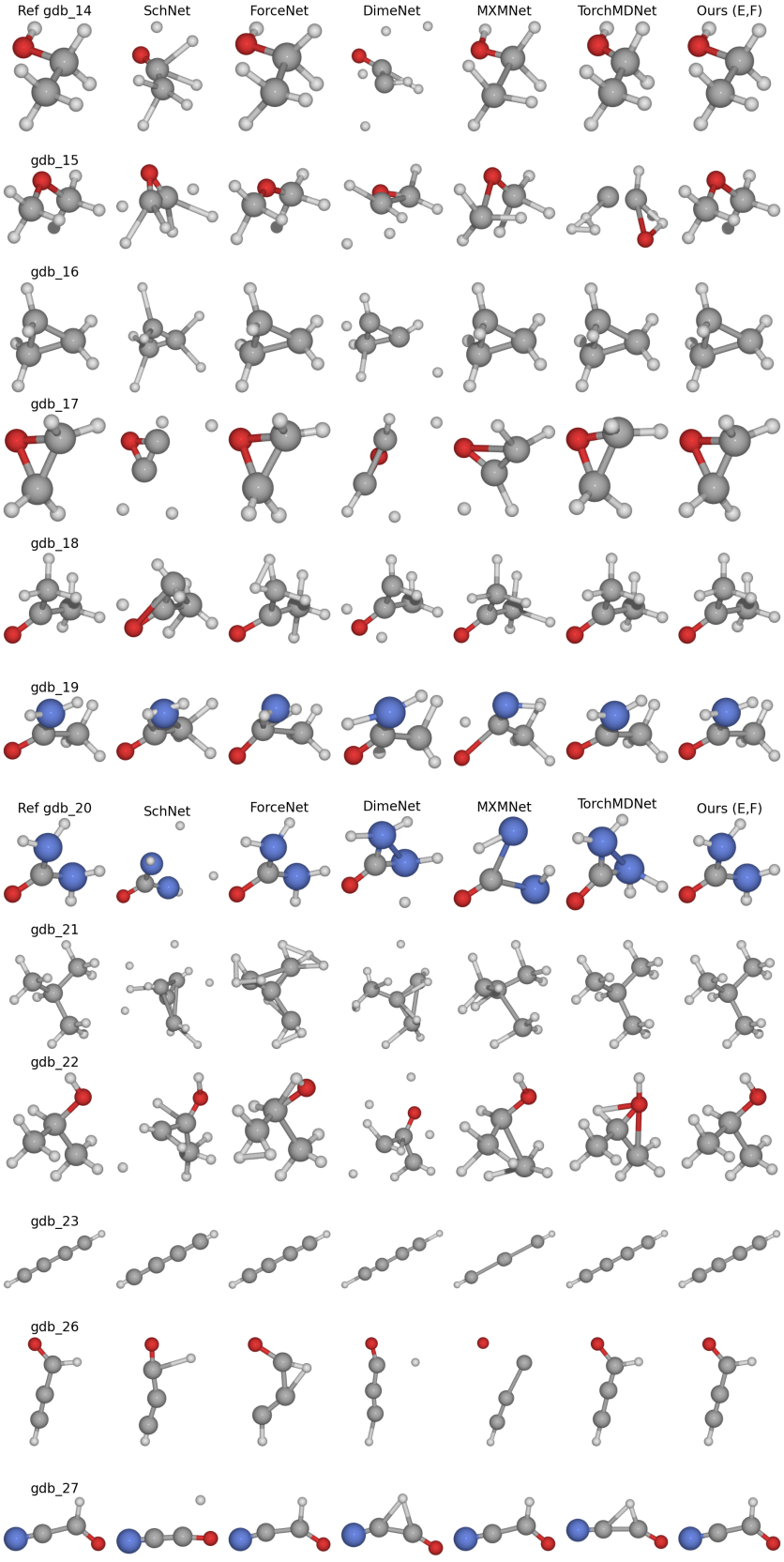

Figure V: Additional structural optimization results (2/9)

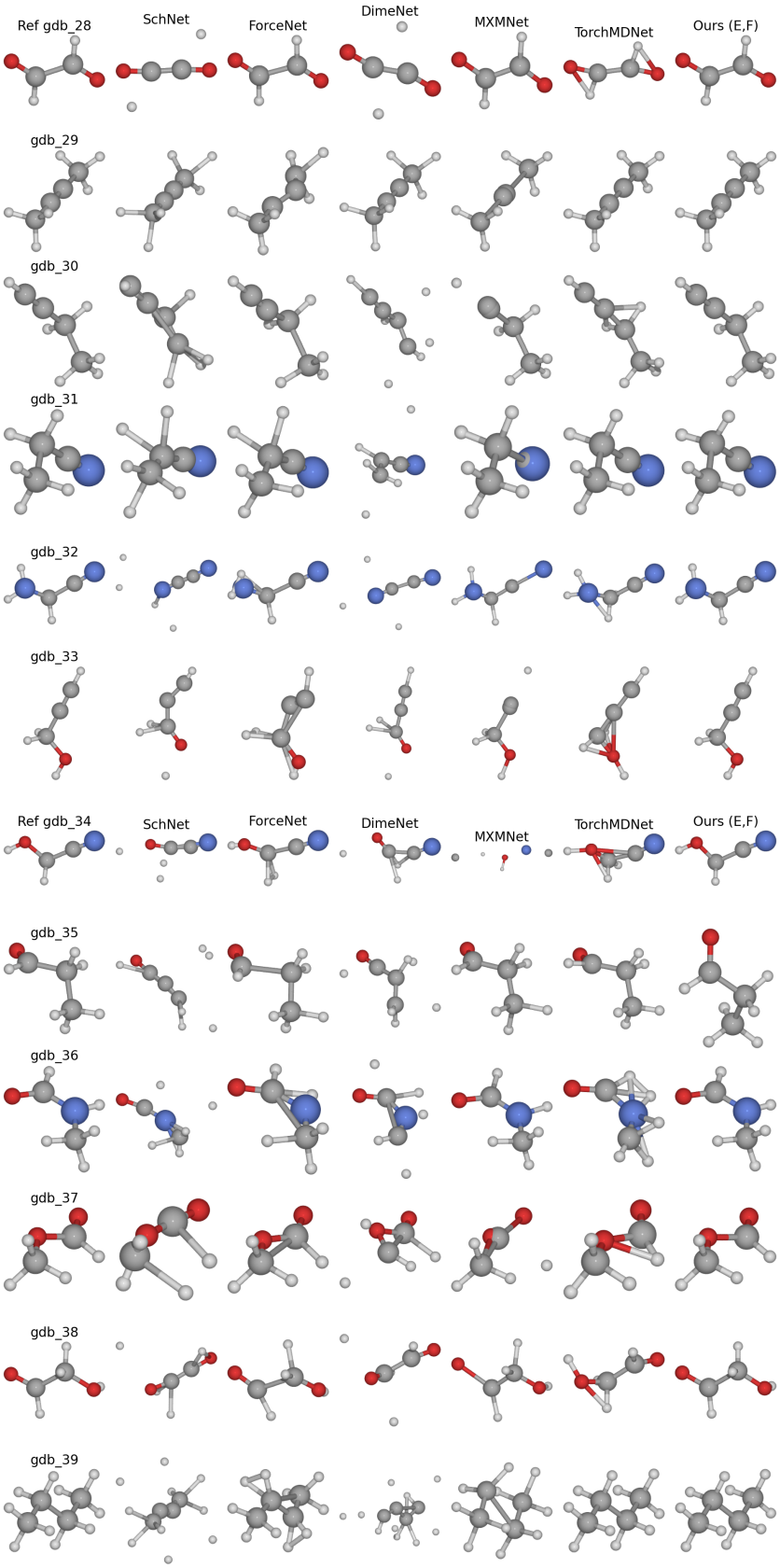

Figure VI: Additional structural optimization results (3/9)

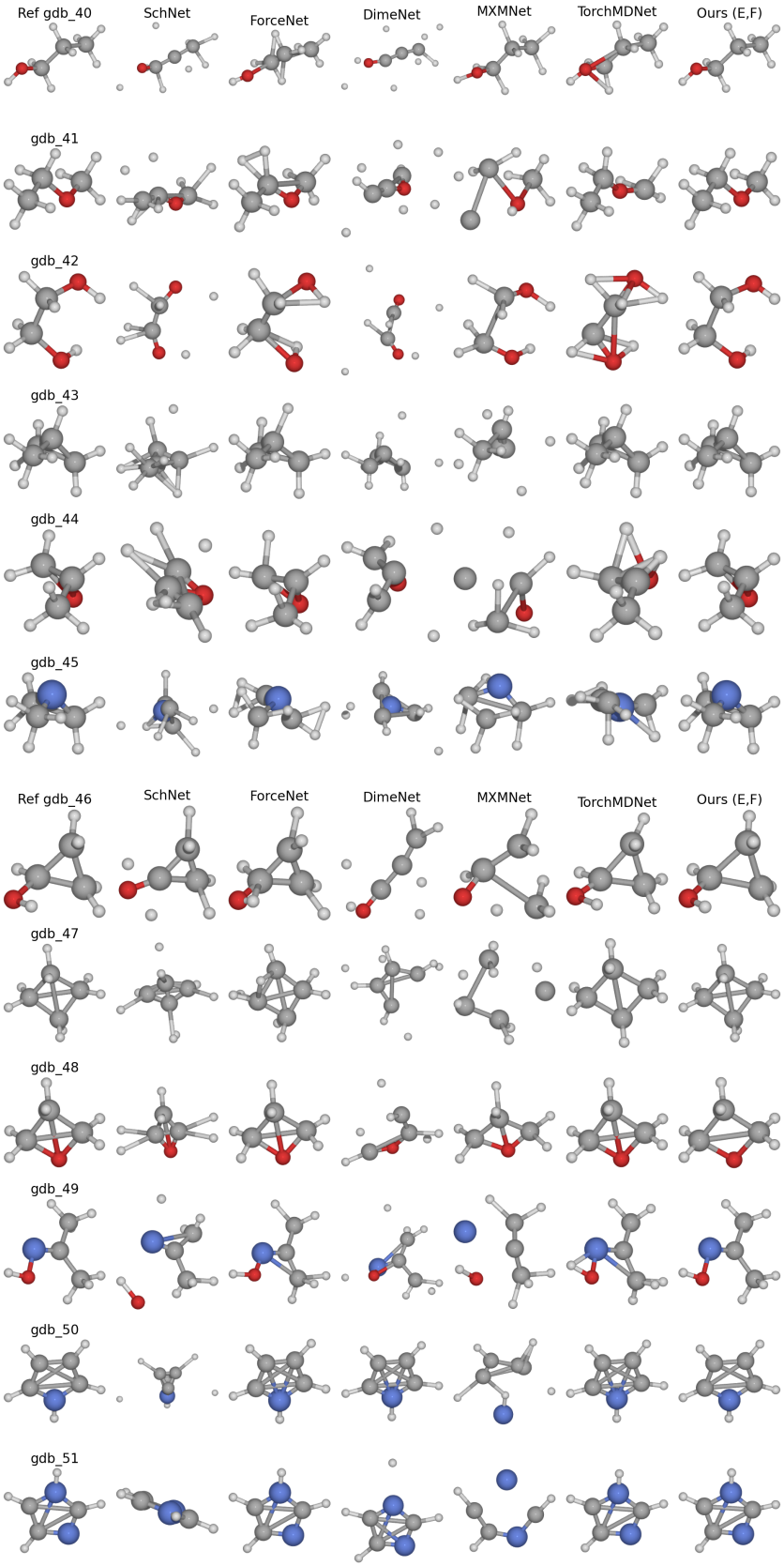

Figure VII: Additional structural optimization results (4/9)

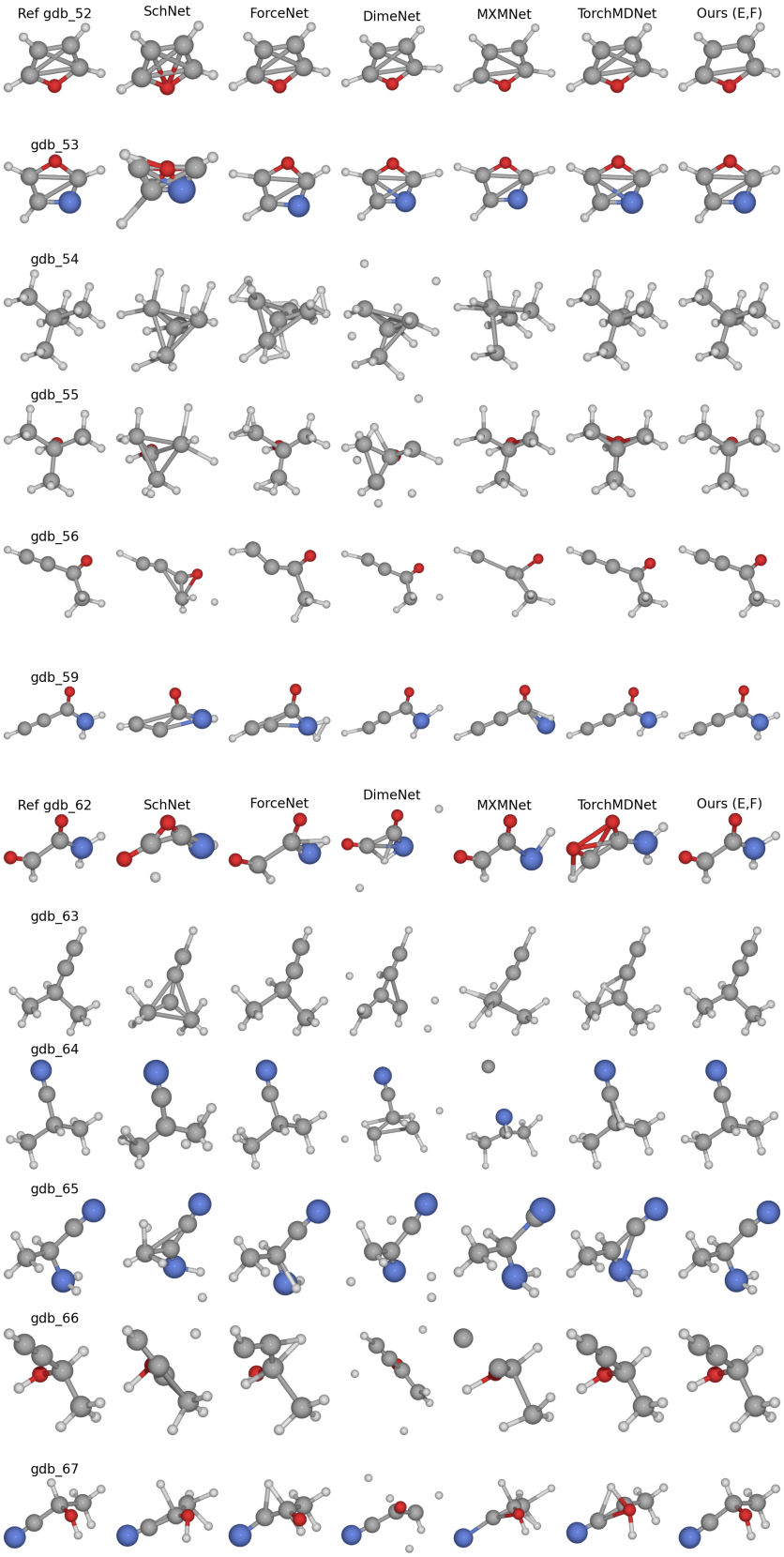

Figure VIII: Additional structural optimization results (5/9)

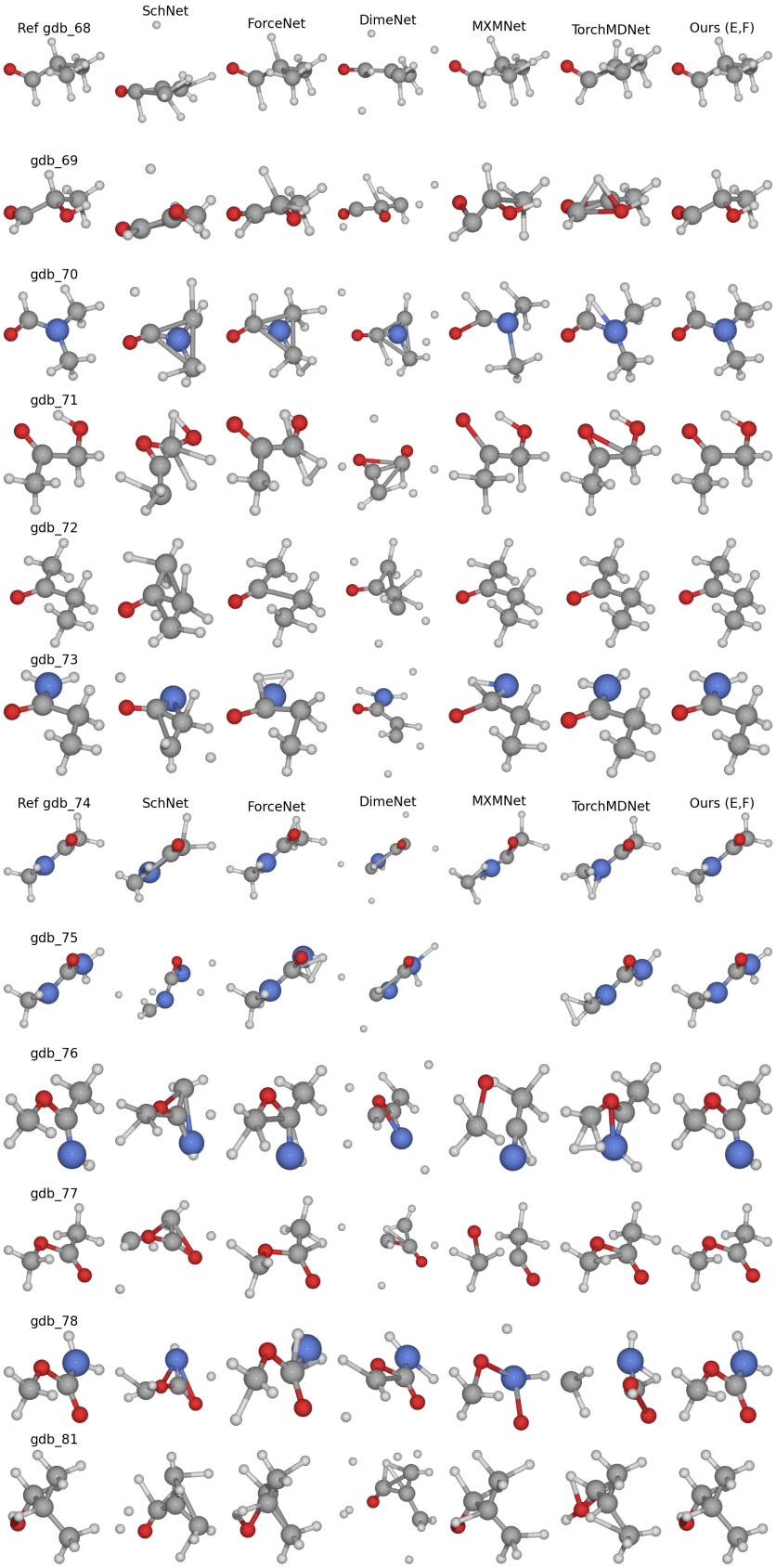

Figure IX: Additional structural optimization results (6/9)

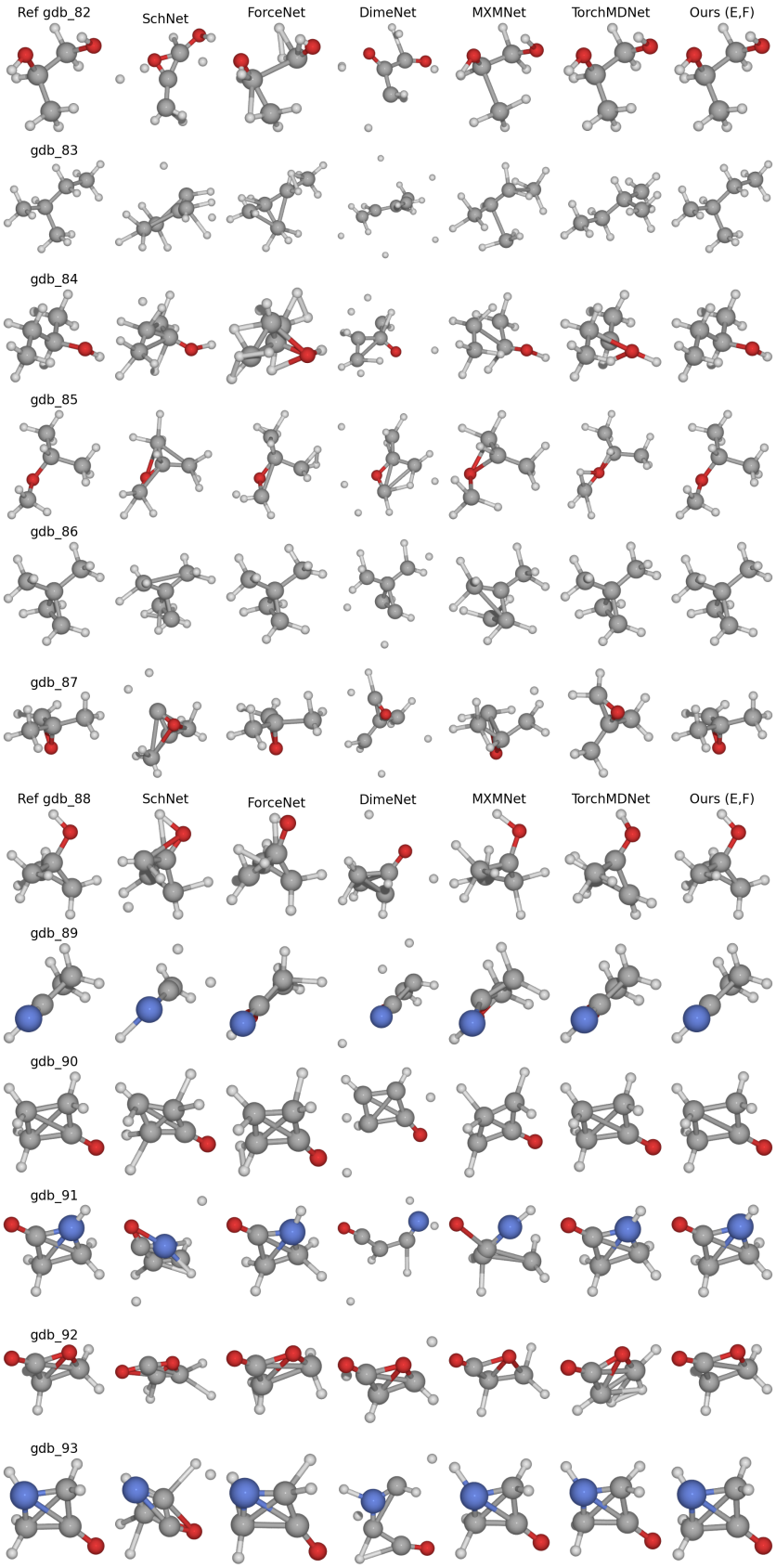

Figure X: Additional structural optimization results (7/9)

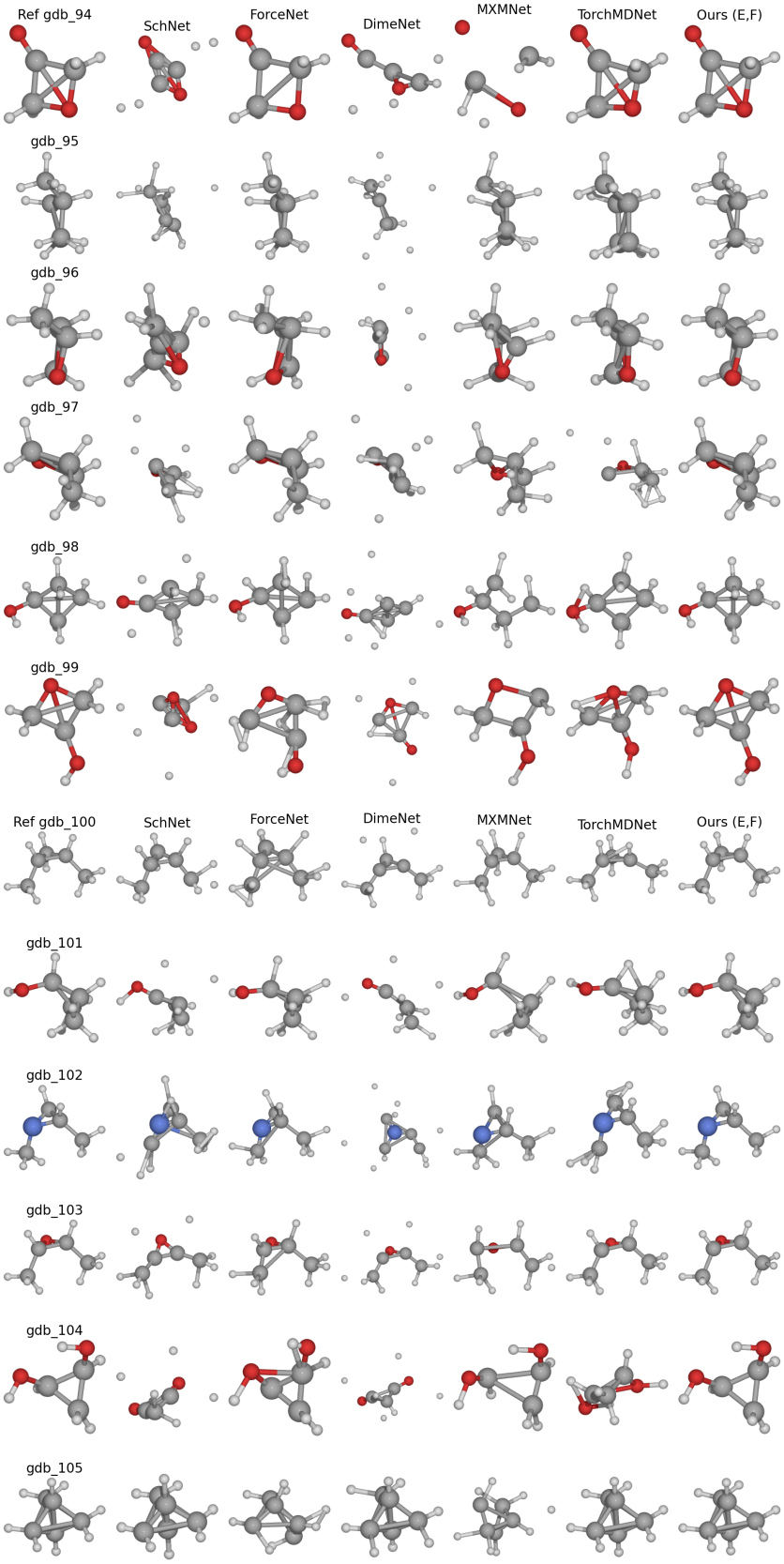

Figure XI: Additional structural optimization results (8/9)

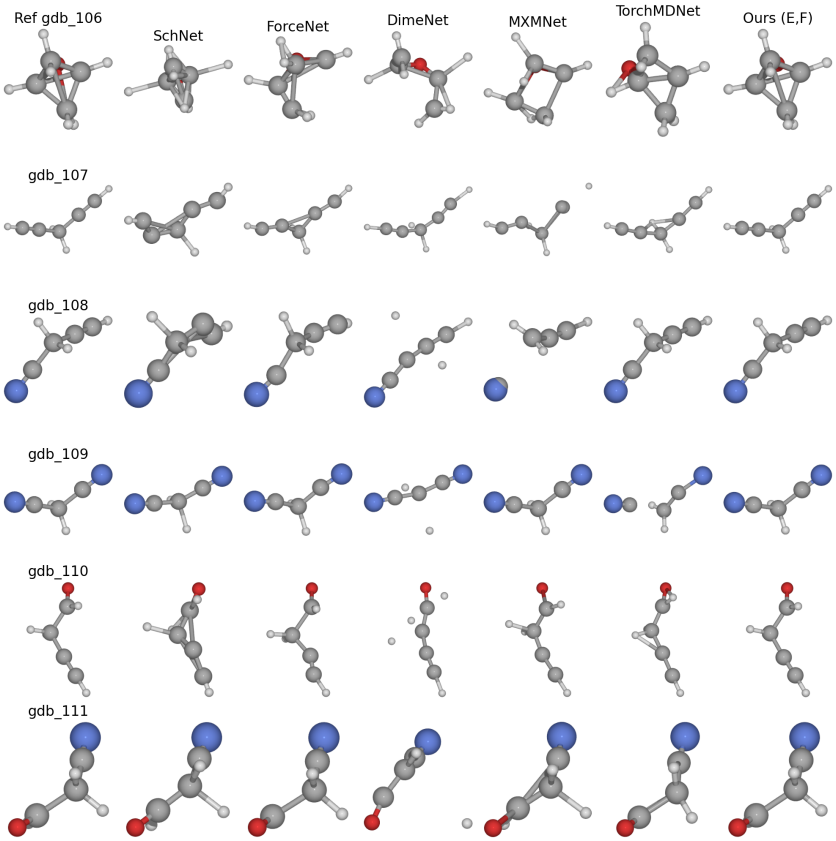

Figure XII: Additional structural optimization results (9/9)