Appendix — Enhancing Multimodal Protein Function Prediction through Dual-Branch Dynamic Selection with Reconstructive Pre-training

1 Extended Experiments of Main Modules

To further validate the rationality of the module design, we conducted an in-depth exploration of the two key components in DSRPGO: BInM and DSM.

1.1 BInM Exploration

To explore the feature interaction capability within BInM, we designed an experiment in which we separately removed the cross-attention module. DSRPGO-BInM-0 and DSRPGO-BInM-1 represent the removal of the bottom and top cross-attention modules from BInM.

Based on the results in Table 1, we find that removing either of the cross-attention modules leads to a decline in overall performance. This validates BInM's ability to capture interactions.

Table 1: Performance Comparison Under Different Interaction Settings. The comparison results of F_{max} score, micro-averaged AUPR (m-AUPR), macro-averaged AUPR (M-AUPR), F1-score (F1), and accuracy (ACC), for BPO, MFO, and CCO. The best results are highlighted in bold, and the sub-optimal results are underlined.

Method	$\mathbf{F_{max}}$			m-AUPR			M-AUPR			F1			ACC		
	BPO	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**
DSRPGO	**0.458**	**0.254**	**0.452**	**0.330**	**0.166**	**0.371**	**0.182**	**0.114**	**0.239**	**0.272**	**0.241**	**0.357**	**0.346**	**0.124**	**0.262**
DSRPGO-BInM-0	0.433	0.189	0.351	0.314	0.113	0.282	0.175	0.106	0.192	0.271	0.179	0.306	0.296	0.086	0.166
DSRPGO-BInM-1	0.434	0.193	0.331	0.314	0.111	0.249	0.170	0.112	0.173	0.268	0.178	0.287	0.289	0.090	0.163

1.2 DSM Exploration

To explore the impact of threshold t selection in the DSM module on model performance, we compare the model's performance under different threshold values. Since there are six experts, the average selection probability for each expert is $1/6$. Therefore, we set the threshold values around $1/6$. Notably, the DSRPGO model employs a learnable, adaptive threshold t.

Based on the experimental results in Table 2, we find that the model performs best when using a learnable, adaptive threshold t. This demonstrates the effectiveness of automatic threshold selection in the DSM module.

2 Extended Experiments of Pre-training Encoders

In this work, the corresponding PSSI encoder and PSeI encoder are trained according to the spatial structure feature and sequence feature in the pre-training stage. To verify the effectiveness of the PSSI encoder and the PSeI encoder, the ablation experiment of the key module of the encoder is carried out in this section.

Table 2: Performance Comparison of Different Threshold t in DSM Module. The comparison results of F_{max} score, micro-averaged AUPR (m-AUPR), macro-averaged AUPR (M-AUPR), F1-score (F1), and accuracy (ACC), for BPO, MFO, and CCO. The best results are highlighted in bold.

Threshold t	$\mathbf{F_{max}}$			m-AUPR			M-AUPR			F1			ACC		
	BPO	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**
auto (ours)	**0.458**	**0.254**	**0.452**	**0.330**	**0.166**	**0.371**	**0.182**	**0.114**	**0.239**	**0.272**	**0.241**	**0.357**	**0.346**	**0.124**	**0.262**
0.1	0.430	0.202	0.428	0.326	0.122	0.314	0.103	0.103	0.221	0.266	0.189	0.325	0.341	0.097	0.252
0.14	0.429	0.201	0.386	0.325	0.123	0.297	0.109	0.109	0.216	0.269	0.187	0.325	0.330	0.099	0.252
0.1616	0.433	0.204	0.427	0.327	0.123	0.286	0.110	0.110	0.212	0.264	0.189	0.333	0.335	0.097	0.261
0.18	0.438	0.205	0.416	0.328	0.123	0.279	0.101	0.101	0.225	0.266	0.190	0.300	0.335	0.102	0.227
0.2	0.429	0.204	0.380	0.326	0.123	0.261	0.107	0.107	0.215	0.269	0.189	0.317	0.346	0.102	0.227
0.3	0.426	0.206	0.411	0.324	0.122	0.271	0.105	0.105	0.224	0.272	0.189	0.333	0.330	0.100	0.227
0.4	0.430	0.205	0.405	0.320	0.121	0.307	0.113	0.113	0.214	0.272	0.193	0.329	0.308	0.100	0.252

Table 3: Ablation experiments of key components in the pre-trained encoder. The comparison results of F_{max} score, micro-averaged AUPR (m-AUPR), macro-averaged AUPR (M-AUPR), F1-score (F1), and accuracy (ACC), for BPO, MFO, and CCO. The Cost Time (ms) represents the time it takes for the model to test each protein sample. The best results are highlighted in bold.

Method	$\mathbf{F_{max}}$			m-AUPR			Cost Time (ms)
	BPO	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**	
Spatial-BiMamba Block	0.370	**0.240**	**0.419**	0.244	**0.156**	**0.371**	**0.539**
Spatial-Multihead Attention	**0.430**	0.223	0.350	**0.291**	0.154	0.313	5.661
Sequence-BiMamba Block	0.290	**0.237**	0.308	0.185	**0.157**	0.230	**0.280**
Sequence-Multihead Attention	**0.329**	0.219	**0.345**	**0.199**	0.148	**0.248**	0.450

In this experiment, the pre-training framework is the same as stage 1 of DSRPGO. Here, BiMamba Block in the pre-trained encoder can be replaced by Multihead Attention Block Wu et al. (2023), and vice versa. Then, we use the pre-trained encoder for feature extraction in the fine-tuning task. During fine-tuning, we only use an MLP for classification.

As shown in Table 3, the model names are formed by combining the feature and framework components with a hyphen. This indicates that the results are obtained by pre-training the feature with the framework component and then fine-tuning it.

We model the spatial structure features and sequence features using encoders with different components. Experimental results in Table 3 show that the encoder using BiMamba Block as a component for modeling spatial structure features shows significant advantages in MFO and CCO. When modeling sequence features, the encoder utilizing Multihead Attention as a component demonstrates considerable advantages in BPO and CCO. Additionally, we observe that BiMamba Block is significantly more efficient in processing spatial structure information, requiring much less time to predict a protein's function compared to Multihead Attention. Considering both inference efficiency and performance, we decide to use an encoder with BiMamba Block as the component for spatial structure features modeling. For sequence information, while BiMamba Block achieves better inference efficiency, we prioritize model performance and opt for Multihead Attention to model sequence features.

3 Impact of Sequence Similarity

To ensure the validity of our experimental design and avoid potential data leakage, we analyze the sequence similarity between the test set and the combined training and validation sets for BPO, MFO, and CCO. We calculate the similarity for each of these sets and categorize the results into different similarity ranges.

Figure 1 shows the number of proteins in the test set within each similarity range. We observe that the majority of proteins in our test set exhibit an average sequence similarity of less than 50% with the proteins in the combined training and validation sets. Only a few proteins have an average

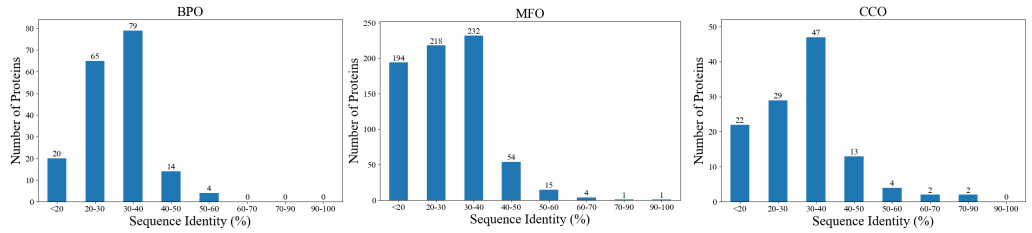

Figure 1: Distribution of sequence identity across proteins in the test dataset. The x-axis represents the sequence similarity ranges. The y-axis indicates the number of proteins within each range. Each bar denotes the number of proteins in the test set that fall within the corresponding similarity range.

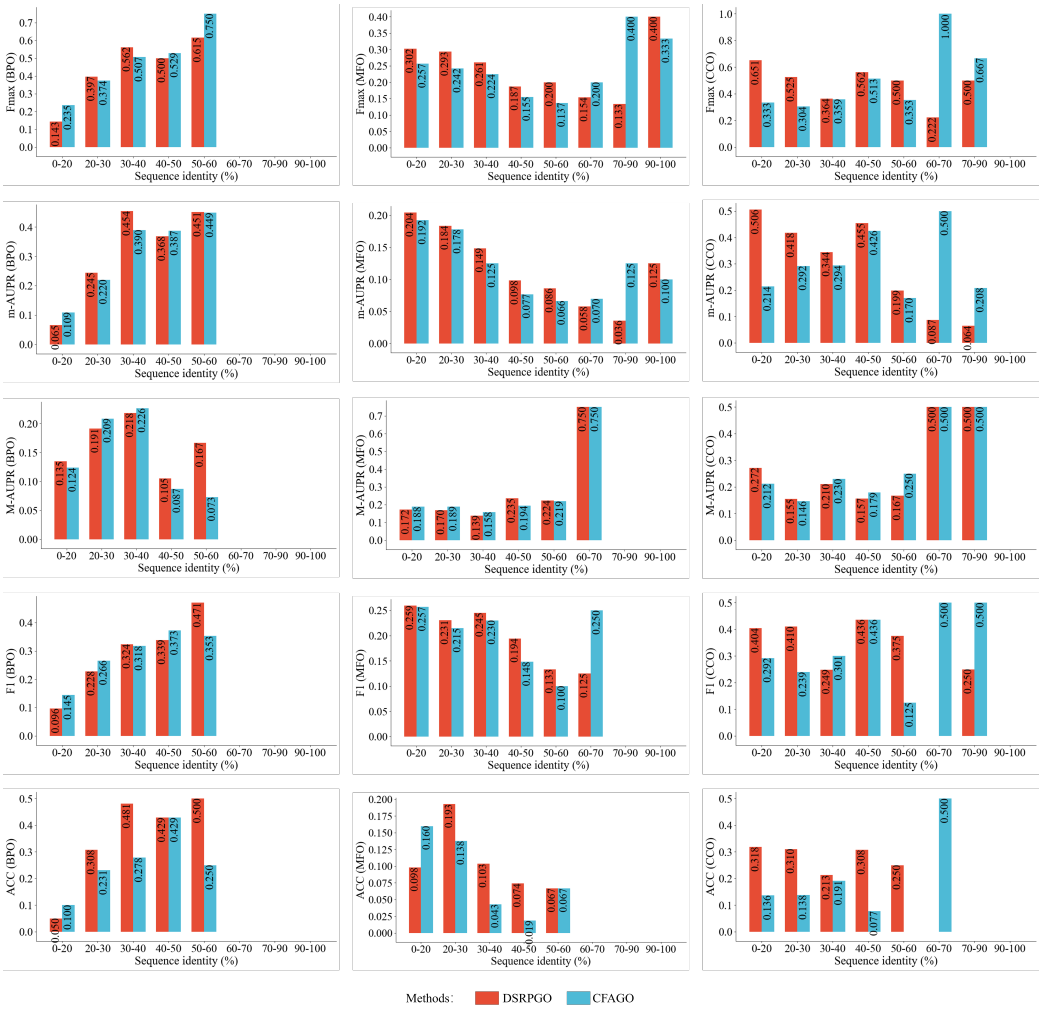

Figure 2: Performance of the different methods in Different Sequence Similarity Ranges. Figures show the results of F1-score (F1), accuracy (ACC), F-max score, macro-averaged AUPR (M-AUPR), and micro-averaged AUPR (m-AUPR) for BPO, MFO and CCO, respectively. The red pillar represents the results of the DSRPGO model and the blue pillar shows the results of the CFAGO model.

Table 4: Performance of the different methods in Different Sequence Similarity Ranges. The comparison results of F_{max} score, micro-averaged AUPR (m-AUPR), macro-averaged AUPR (M-AUPR), F1-score (F1), and accuracy (ACC), for BPO, MFO, and CCO. The Range represents the proteins that have average similarity within a specific range.

Method	Range	$\mathbf{F_{max}}$			m-AUPR			M-AUPR			F1			Acc		
		BPO	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**
CFAGO	0-100	0.439	0.236	0.366	0.328	0.159	0.337	0.188	0.138	0.210	0.283	0.234	0.314	0.338	0.100	0.210
	0-20	0.235	0.257	0.333	0.109	0.192	0.214	0.124	0.188	0.212	0.145	0.257	0.292	0.100	0.160	0.136
	20-30	0.374	0.242	0.304	0.220	0.178	0.292	0.209	0.189	0.146	0.266	0.215	0.239	0.231	0.138	0.138
	30-40	0.507	0.224	0.359	0.390	0.125	0.294	0.226	0.158	0.230	0.318	0.230	0.301	0.278	0.043	0.191
	40-50	0.529	0.155	0.513	0.387	0.077	0.426	0.087	0.194	0.179	0.373	0.148	0.436	0.429	0.019	0.077
	50-60	0.750	0.137	0.353	0.449	0.066	0.170	0.073	0.219	0.250	0.353	0.100	0.125	0.250	0.067	0.000
	60-70	-	0.200	1.000	-	0.070	0.500	-	0.750	0.500	-	0.250	0.500	-	0.000	0.500
	70-90	-	0.400	0.667	-	0.125	0.208	-	0.000	0.500	-	0.000	0.500	-	0.000	0.000
	90-100	-	0.333	-	-	0.100	-	-	0.000	-	-	0.000	-	-	0.000	-
DSRPGO (Ours)	0-100	0.458	0.330	0.452	0.330	0.166	0.371	0.182	0.114	0.239	0.272	0.241	0.357	0.346	0.124	0.262
	0-20	0.143	0.302	0.651	0.065	0.204	0.506	0.135	0.172	0.272	0.096	0.259	0.404	0.05	0.098	0.318
	20-30	0.397	0.293	0.525	0.245	0.184	0.418	0.191	0.17	0.155	0.228	0.231	0.41	0.308	0.193	0.31
	30-40	0.562	0.261	0.364	0.454	0.149	0.344	0.218	0.139	0.21	0.324	0.245	0.249	0.481	0.103	0.213
	40-50	0.5	0.187	0.562	0.368	0.098	0.455	0.105	0.235	0.157	0.339	0.194	0.436	0.429	0.074	0.308
	50-60	0.615	0.2	0.5	0.451	0.086	0.199	0.167	0.224	0.167	0.471	0.133	0.375	0.5	0.067	0.25
	60-70	-	0.154	0.222	-	0.058	0.087	-	0.75	0.5	-	0.125	0	-	0	0
	70-90	-	0.133	0.5	-	0.036	0.064	-	0	0.5	-	0	0.25	-	0	0
	90-100	-	0.4	-	-	0.125	-	-	0	-	-	0	-	-	0	-

Table 5: Comparison with Structure-based and PLM-based methods. The comparison results of F_{max} score, and micro-averaged AUPR (m-AUPR) for BPO, MFO, and CCO. The best results are highlighted in bold.

Method	Focus	$\mathbf{F_{max}}$			m-AUPR		
		BPO	**MFO**	**CCO**	**BPO**	**MFO**	**CCO**
DeepFRI	Structure based	0.362	**0.461**	0.385	0.308	**0.382**	0.360
PredGO	Structure + PLM based	0.108	0.455	0.252	0.058	0.254	0.183
DSRPGO (Ours)	Multi-modal based	**0.440**	0.282	**0.421**	**0.332**	0.172	**0.392**

similarity greater than 70%. Based on these results, we conclude that the time-based split used in the CAFA challenge is reasonable and does not introduce significant sequence similarity overlap between our training and test sets.

In the comparative experiments described in Section 3.2, BLAST, which primarily relies on sequence similarity, performs poorly. This further supports the notion that the sequence similarity between the test set and the combined training and validation sets is relatively low.

Additionally, we explore the model's performance in predicting protein functions within different similarity ranges. We divide the test set into 9 similarity intervals.

In Figure 2, we present the performance of DSRPGO and CFAGO across different protein similarity ranges. Specifically, range (n-m) represents the average similarity of proteins in the range of n-m. For both BPO and CCO, our method demonstrates significantly better performance than CFAGO, particularly for proteins with low similarity. And the detailed results are shown in Table 4. The symbol "-" indicates no proteins in that interval. Due to the scarcity of proteins with similarity above 50%, we do not provide further analysis for this group. For proteins with similarity below 50%, the model performs best in predicting proteins in the similarity range of 30 to 40 for BPO. For MFO and CCO, the model performs best for proteins in the similarity range of 0 to 30. These results show that our proposed DSRPGO model is stable across different sequence similarities and is not reliant on the sequence similarity of the data.

4 Comparison with Structure-based and PLM-based Methods

To ensure a fair comparison, we conduct an additional experiment, evaluating the performance of the structure-based method DeepFRI (Gligorijević et al., 2021), and PredGO (Zheng et al., 2023) method using a protein language model (PLM). The results shown in Table 5 demonstrate that our approach outperforms both methods in BPO and CCO aspects.

Table 6: Performance Comparison Under Different Learning Rate Settings. The comparison results of F_{max} score, micro-averaged AUPR (m-AUPR), macro-averaged AUPR (M-AUPR), F1-score (F1), and accuracy (ACC), for BPO, MFO, and CCO. The best results are highlighted in bold.

LR	F_{max}			m-AUPR			M-AUPR			F1			ACC		
	BPO	MFO	CCO	BPO	MFO	CCO	BPO	MFO	CCO	BPO	MFO	CCO	BPO	MFO	CCO
1.00E-03 (ours)	**0.458**	0.254	**0.452**	**0.330**	**0.166**	**0.371**	0.182	0.114	0.239	0.272	0.241	**0.357**	**0.346**	**0.124**	**0.262**
1.00E-04	0.352	**0.265**	0.387	0.254	0.160	0.291	**0.195**	0.114	0.242	0.261	**0.261**	0.333	0.220	0.071	0.261
1.00E-05	0.063	0.146	0.291	0.034	0.055	0.262	0.101	0.064	0.231	0.035	0.009	0.202	0.000	0.000	0.134
2.00E-03	0.451	0.242	0.423	0.310	0.145	0.361	0.169	0.113	0.225	**0.280**	0.246	0.325	0.302	0.078	0.185
2.00E-04	0.422	0.231	0.435	0.307	0.148	0.324	0.187	**0.123**	0.225	0.269	0.234	0.337	0.264	0.085	0.244
2.00E-05	0.181	0.166	0.311	0.082	0.073	0.291	0.125	0.097	**0.264**	0.132	0.089	0.239	0.011	0.001	0.151

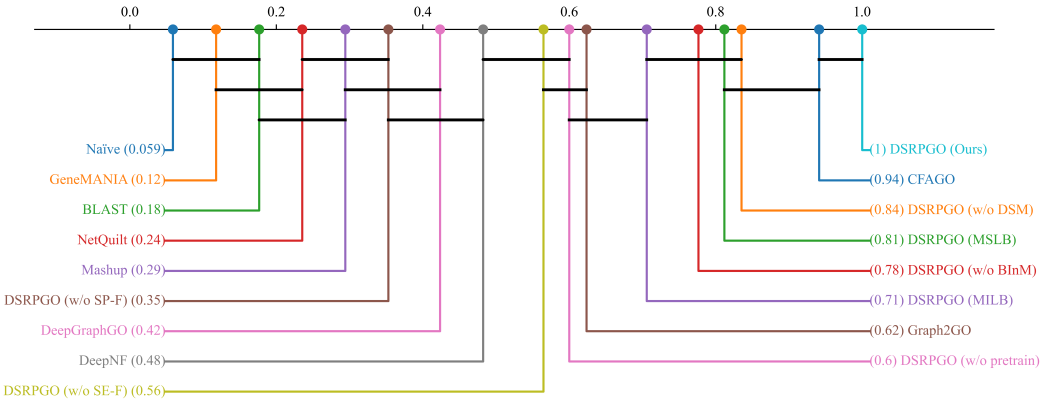

Figure 3: Critical Difference Diagrams.

5 Impact of Learning Rate

To investigate the impact of the learning rate (LR) on model performance, we conduct experiments with various learning rates ranging from 1e-5 to 2e-3. All other hyperparameters, including the optimizer, batch size, and training epochs, are kept constant to isolate the effect of the learning rate.

From Table 6, we can see that the model performs best when the learning rate is set to 1e-3. This may be because a lower learning rate leads to underfitting of the model.

6 Critical Difference Diagrams for Statistical Comparison

To assess the significance of the results and compare the performance of different approaches, we use critical difference diagrams (CDDs) (Benavoli et al., 2016). These diagrams, as described in scikit-posthocs documentation, are particularly useful in visualizing whether differences between approaches are statistically significant.

The CDD, as shown in Figure 3, presents the overall comparison of all models across various aspects and metrics. To provide a more detailed analysis, we include additional CDDs to compare the models' performance across different aspects, metrics, and experiments. The CDDs used for the comparative experiments are shown in Figures 4, 5 and 6.The CDDs used for the module ablation experiments are shown in Figures 7, 8 and 9. By examining the CDDs, we can observe that DSRPGO ranks first in almost all cases, indicating that our model demonstrates significant advantages in single-species protein function prediction.

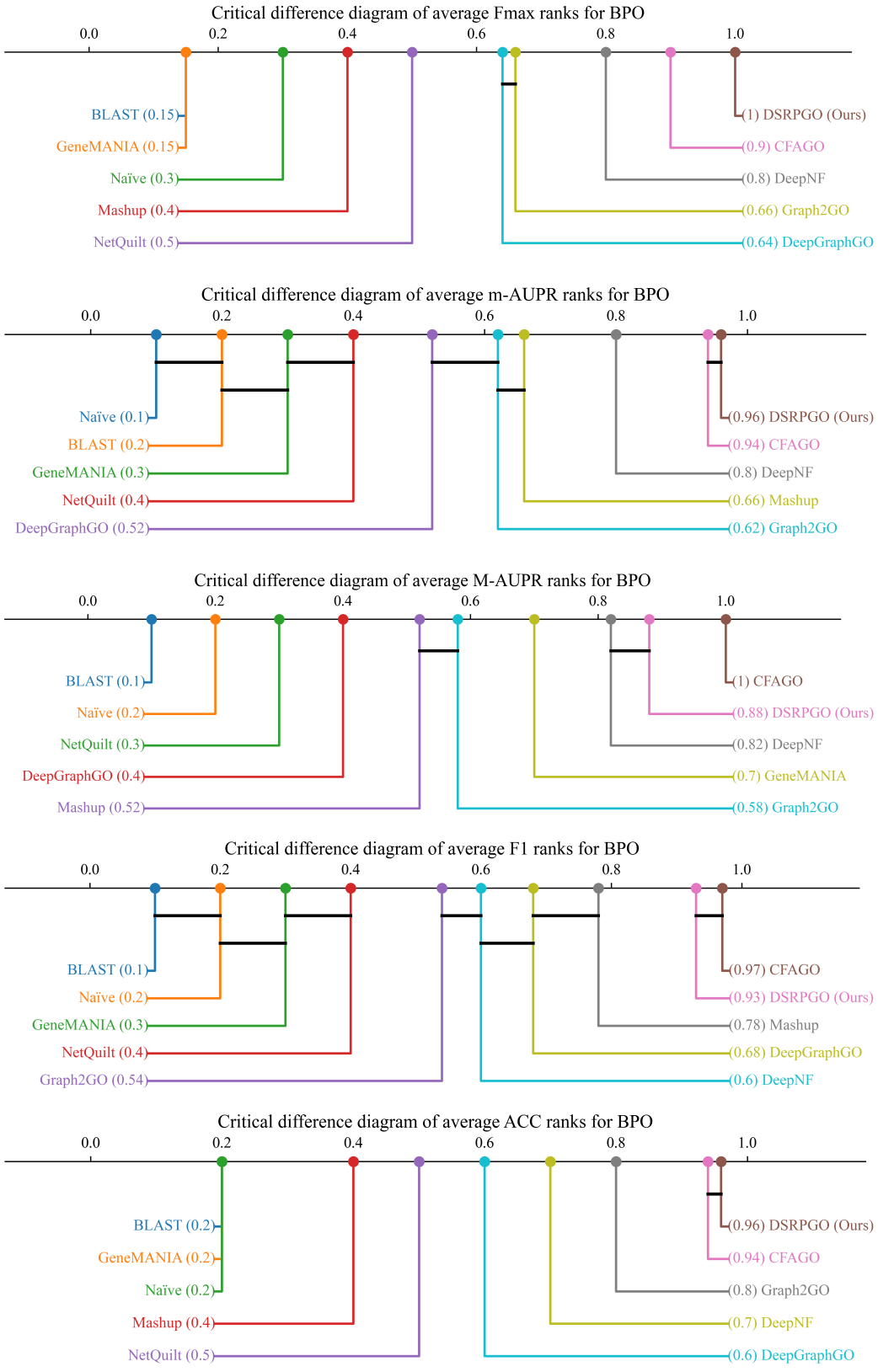

Figure 4: Critical Difference Diagram for Comparative Experiments on BPO.

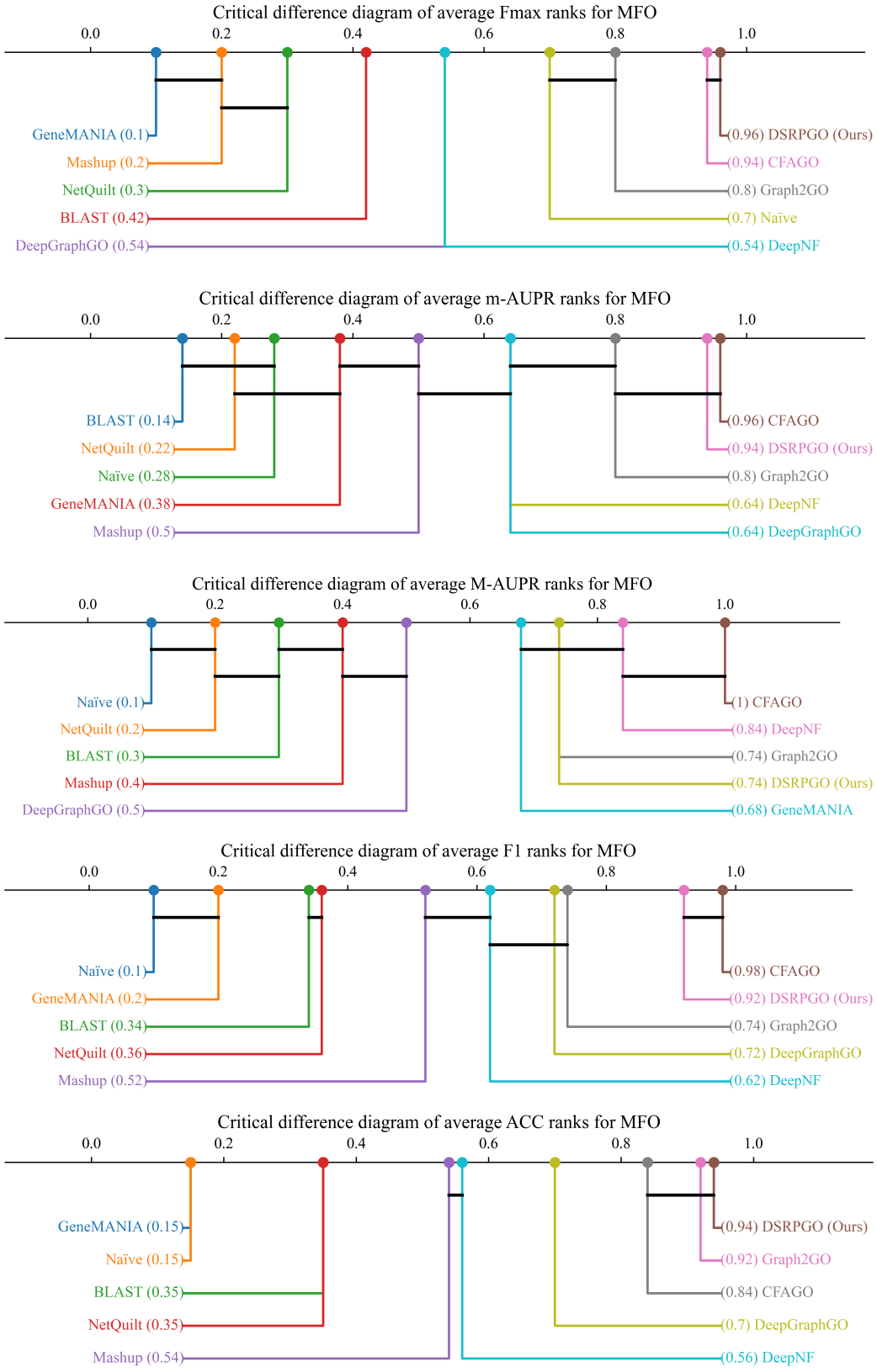

Figure 5: Critical Difference Diagram for Comparative Experiments on MFO.

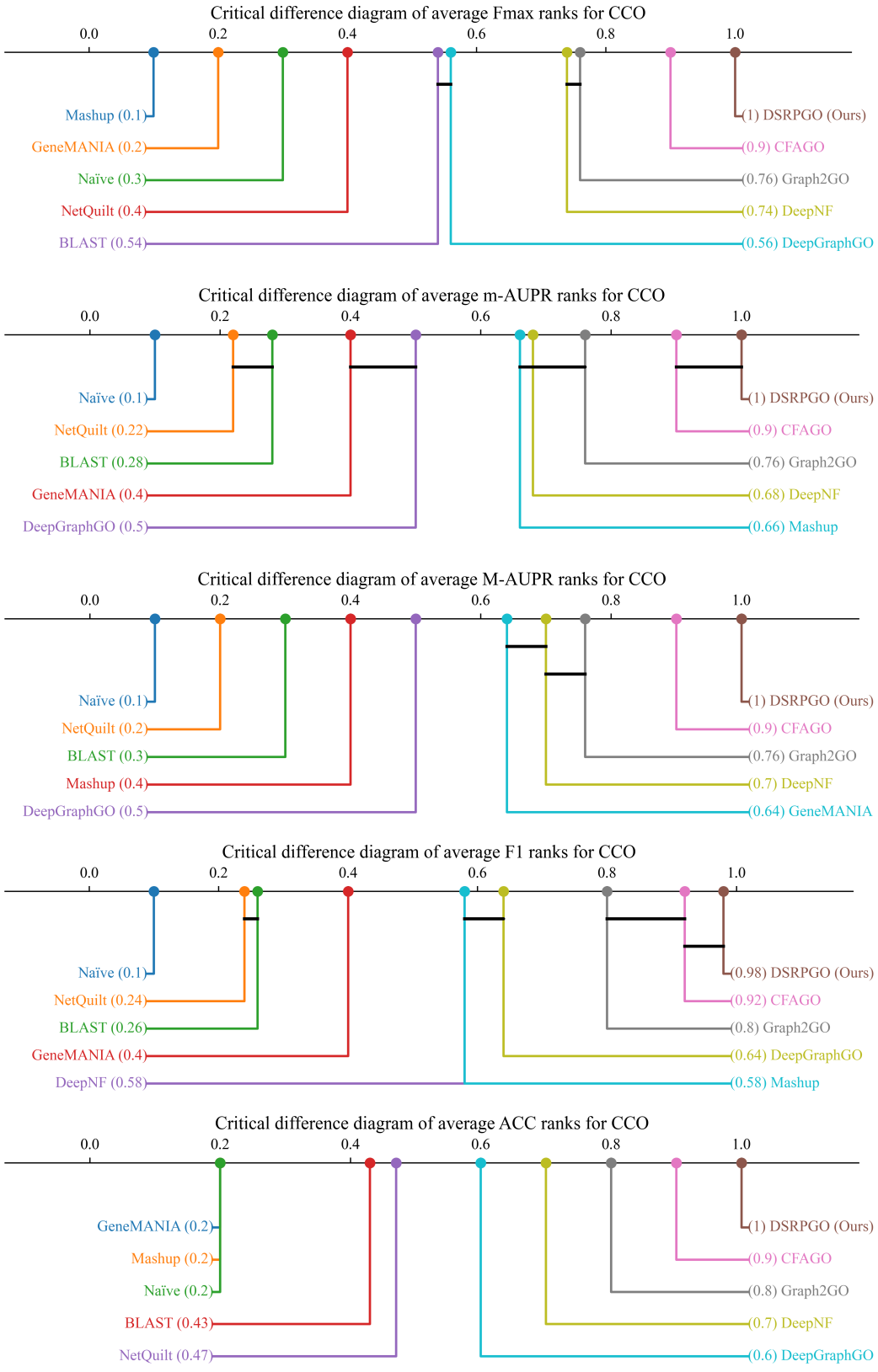

Figure 6: Critical Difference Diagram for Comparative Experiments on CCO.

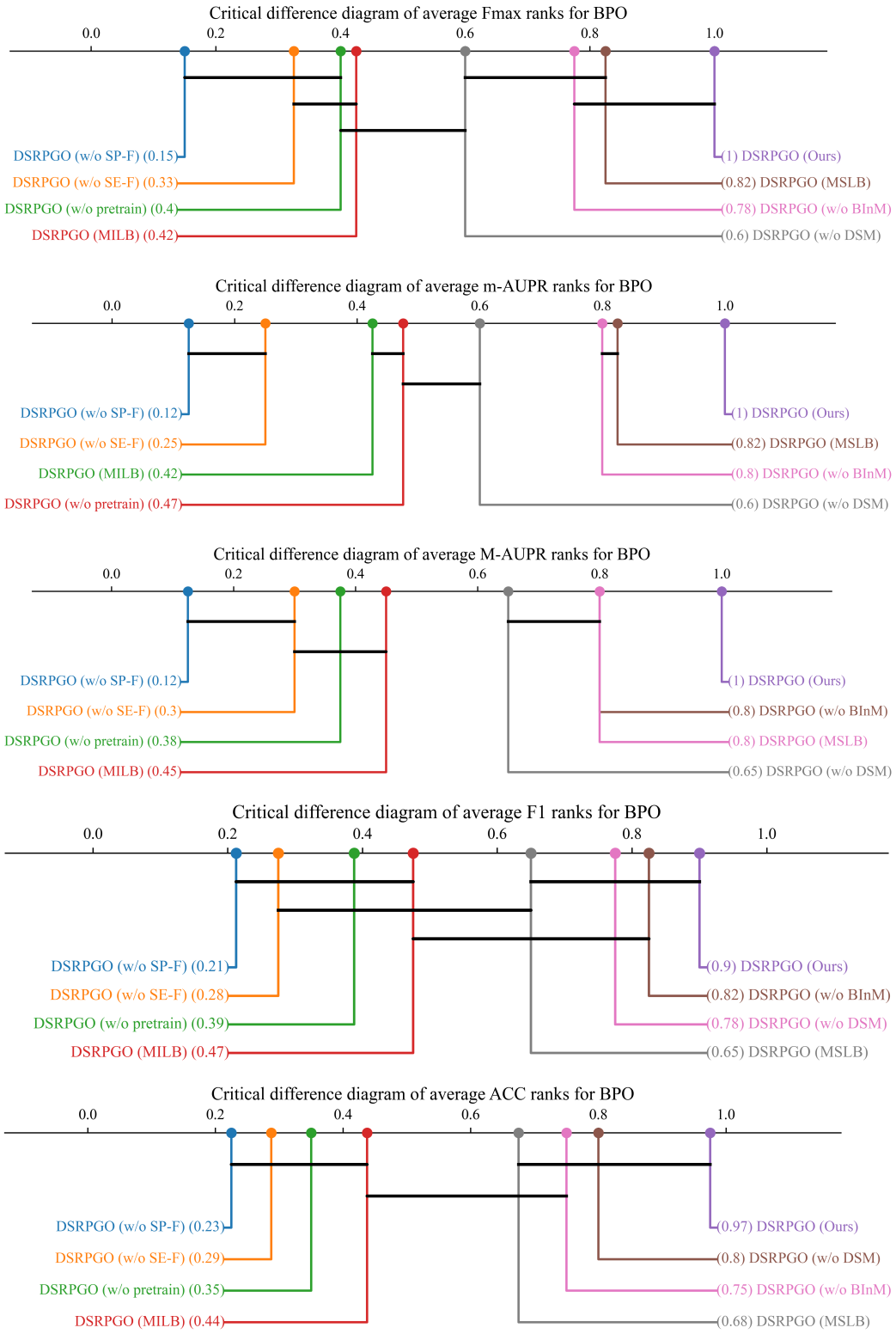

Figure 7: Critical Difference Diagram for Module Ablation on BPO.

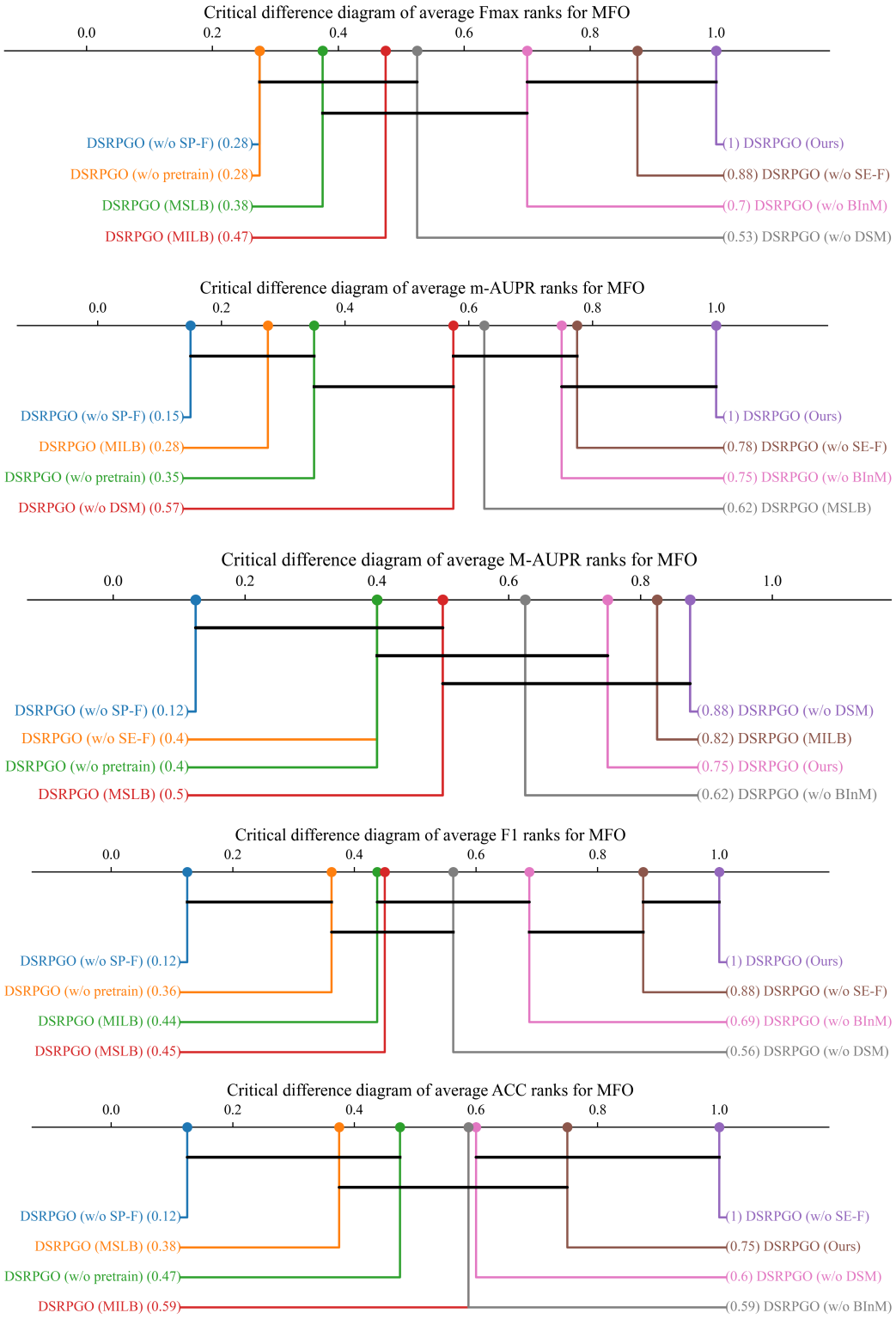

Figure 8: Critical Difference Diagram for Module Ablation on MFO.

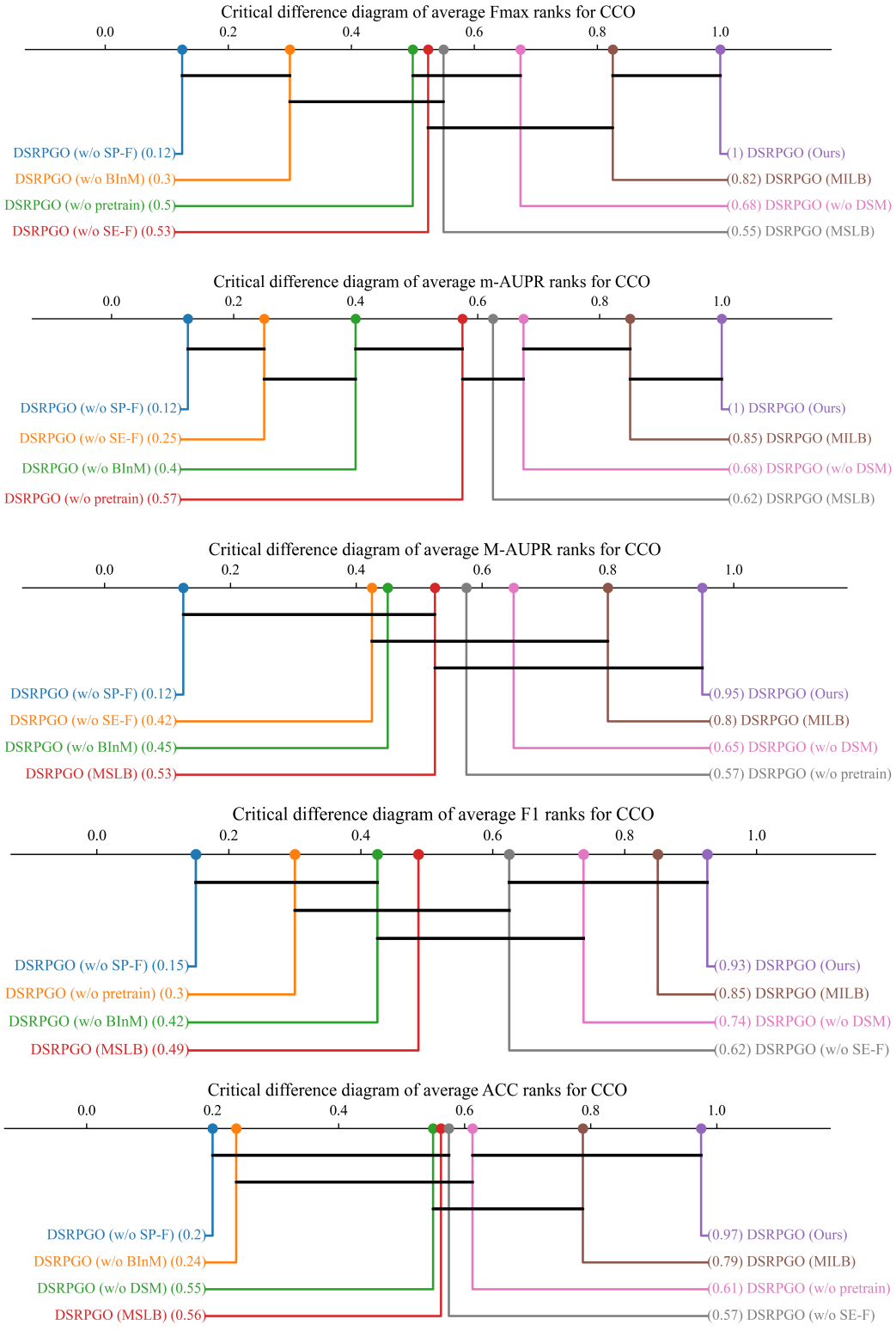

Figure 9: Critical Difference Diagram for Module Ablation on CCO.

References

Alessio Benavoli, Giorgio Corani, and Francesca Mangili. Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research*, 17(1):152–161, 2016.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

Zhourun Wu, Mingyue Guo, Xiaopeng Jin, Junjie Chen, and Bin Liu. Cfago: cross-fusion of network and attributes based on attention mechanism for protein function prediction. *Bioinformatics*, 39(3):btad123, 2023.

Rongtao Zheng, Zhijian Huang, and Lei Deng. Large-scale predicting protein functions through heterogeneous feature fusion. *Briefings in Bioinformatics*, 24(4):bbad243, 2023.