Supplemental Information

Appendix A. Frequency Histograms

A) BBBP: “Chiralities Distribution”

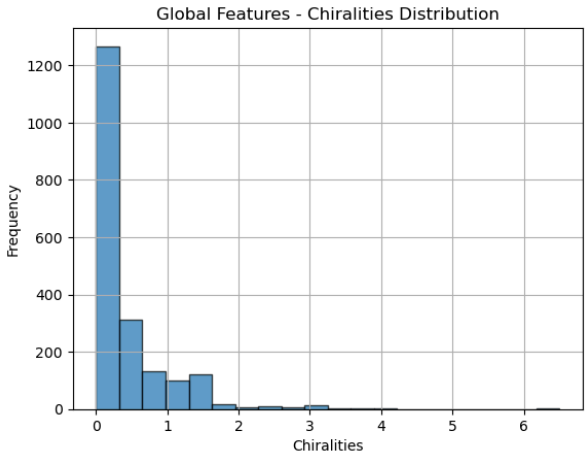


Figure 1. Chiralities distribution in the BBBP dataset. B) Atom degree distribution in the BBBP dataset.

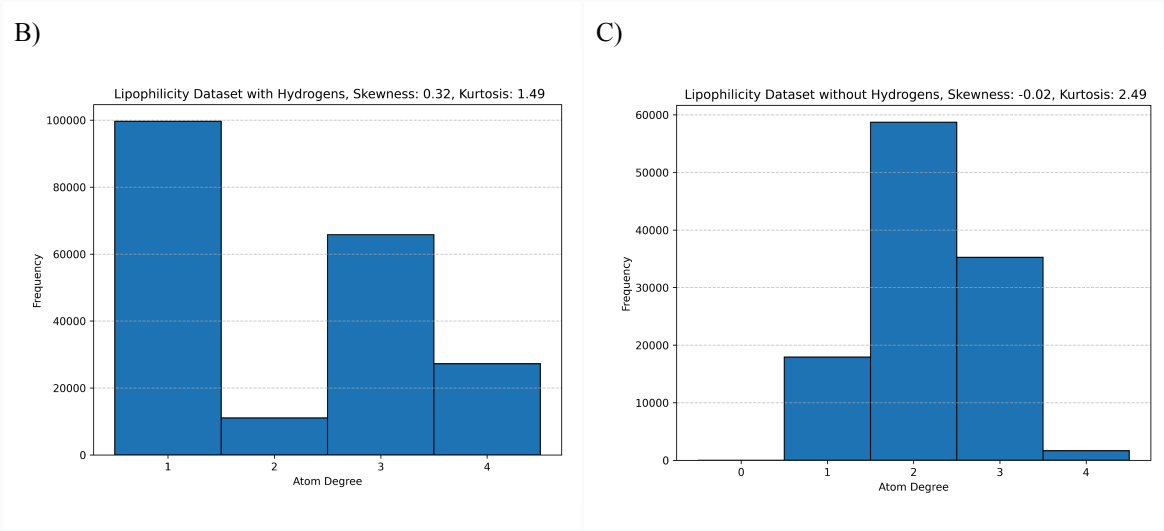
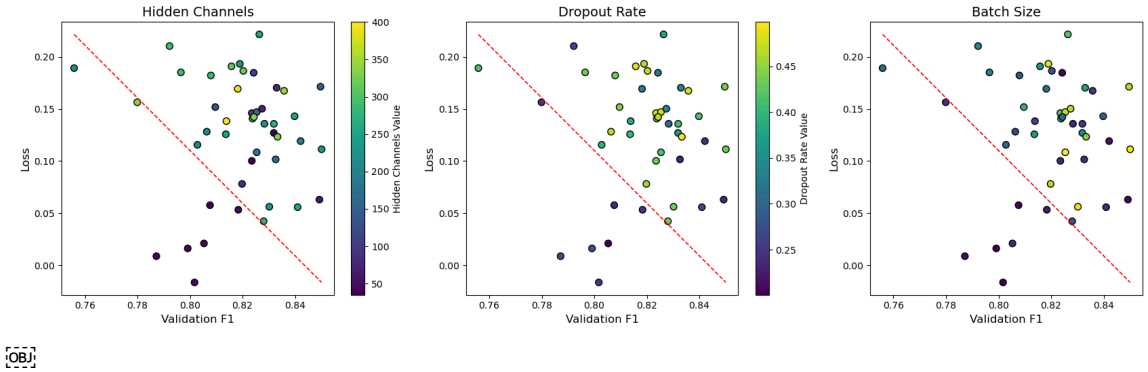


Figure 2. Distribution changes derived from removing hydrogen in the Lipophilicity dataset.

Appendix B.

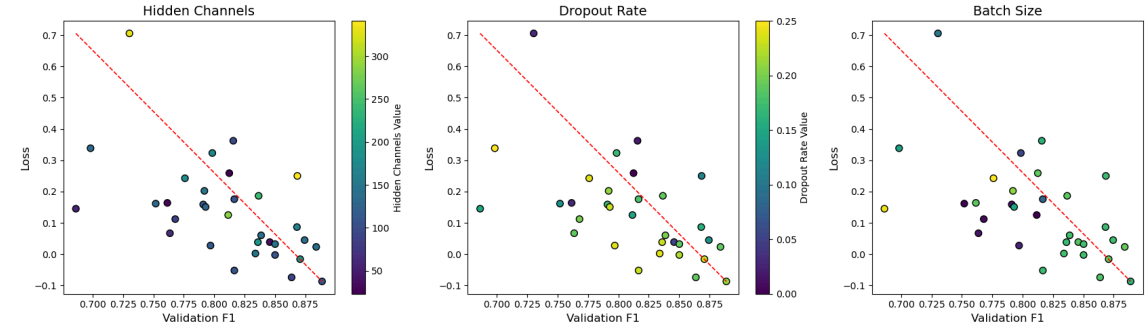
Pareto plots for the hyperparameters optimized using 5-fold-cross validation procedures with a dual direction objective: loss difference minimization and maximization of Validation F1.

A) BACE - UMP

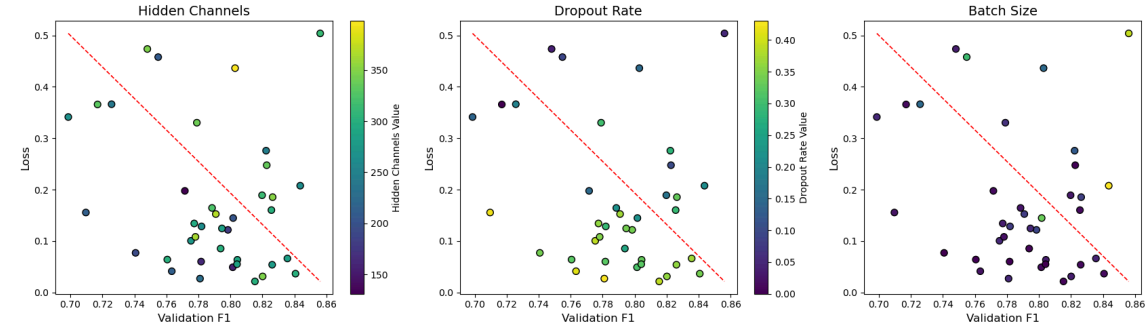


OBJ

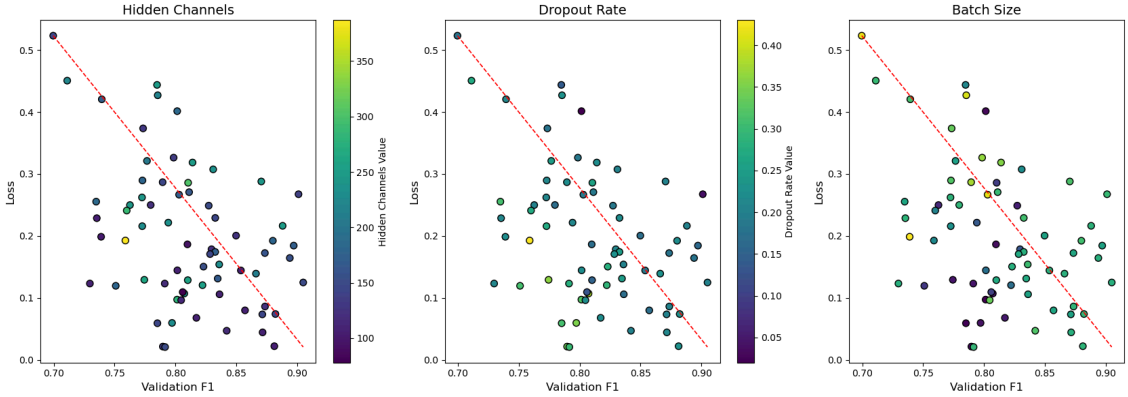
B) BACE - CBMP



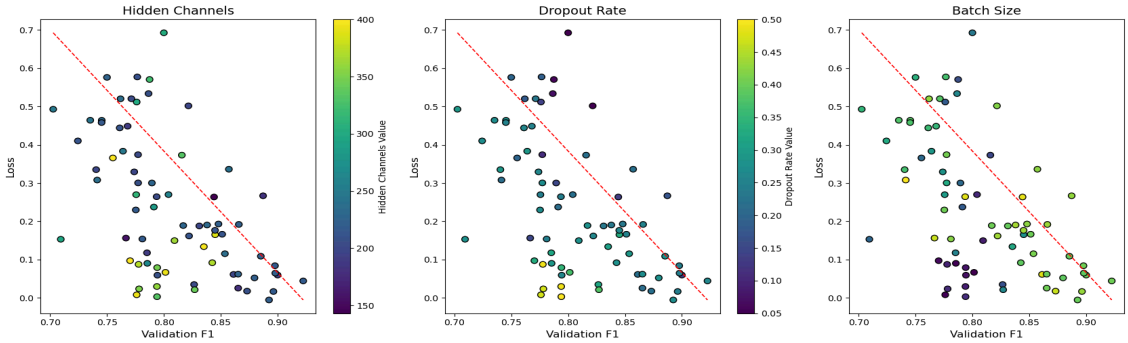
C) BACE - ABMP



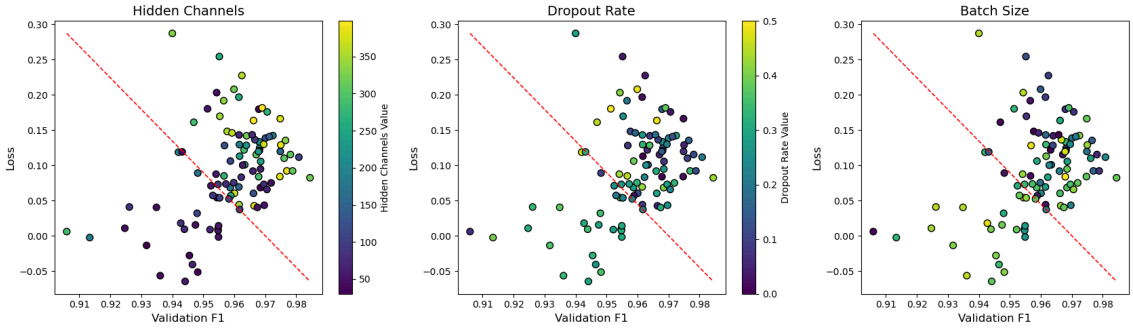
D) BACE - BMP + SN



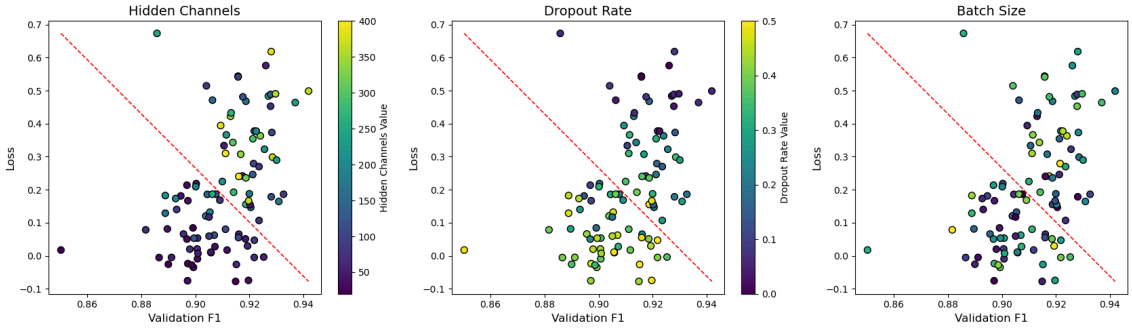
E) BACE - ABMP + SN



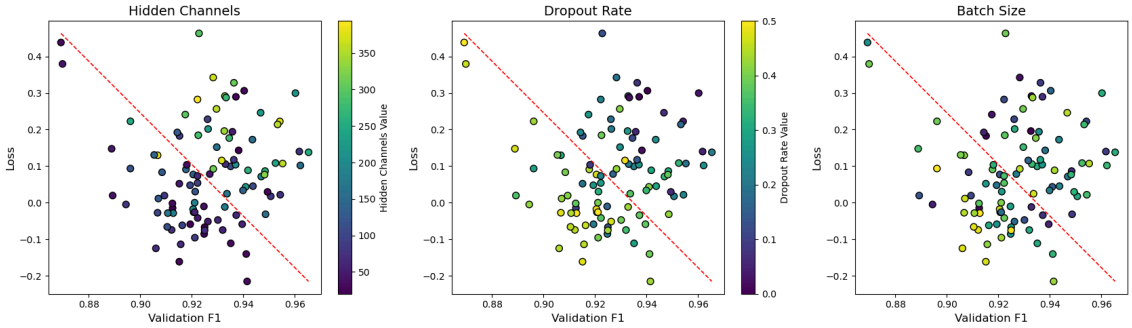
F) BBBP - UMP



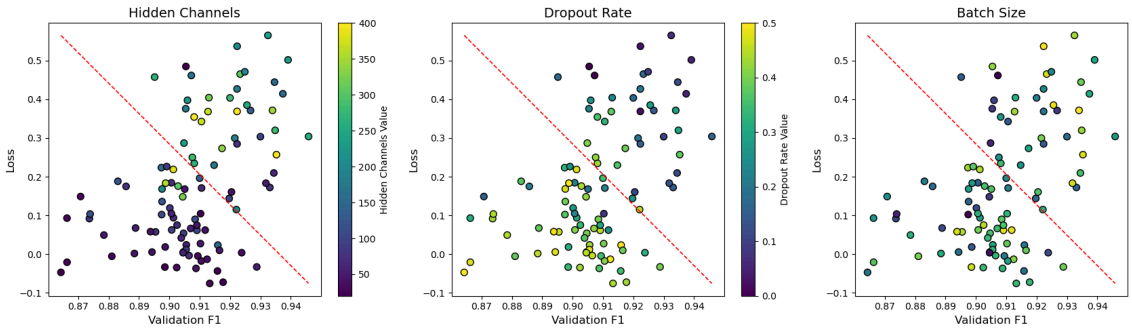
G) BBBP - CBMP



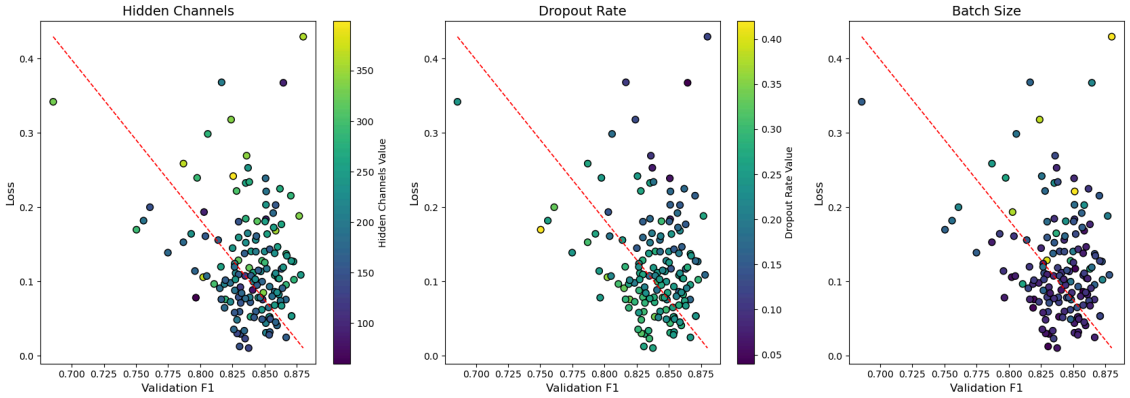
H) BBBP - BMP + SN



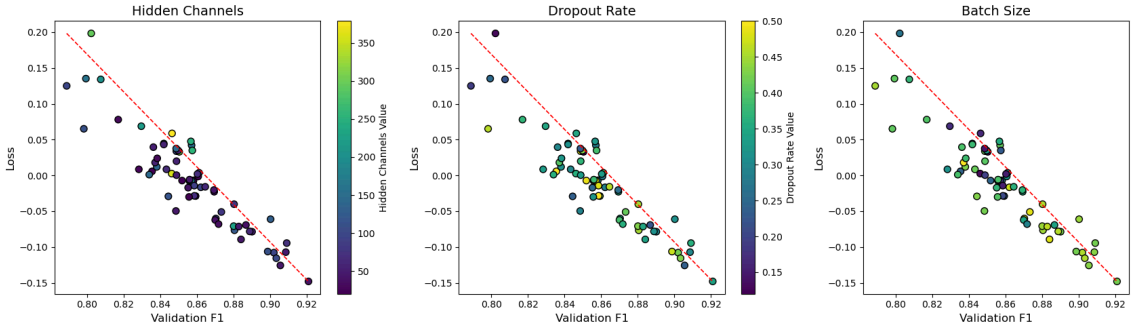
H) BBBP - ABMP + SN



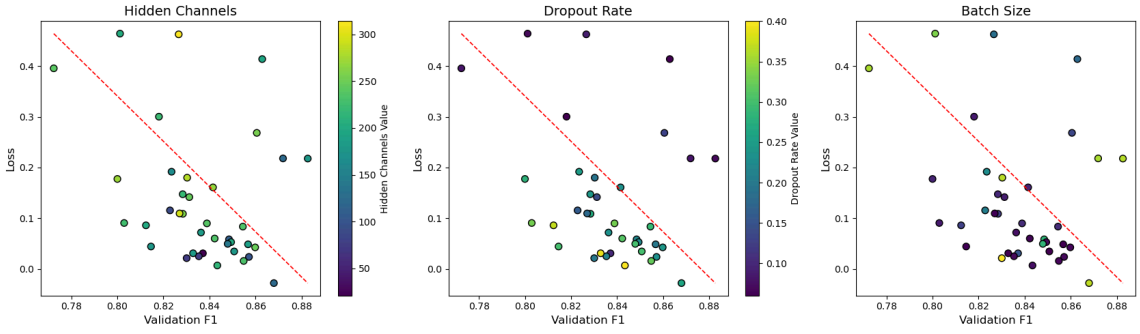
I) TRPA1 - BMP



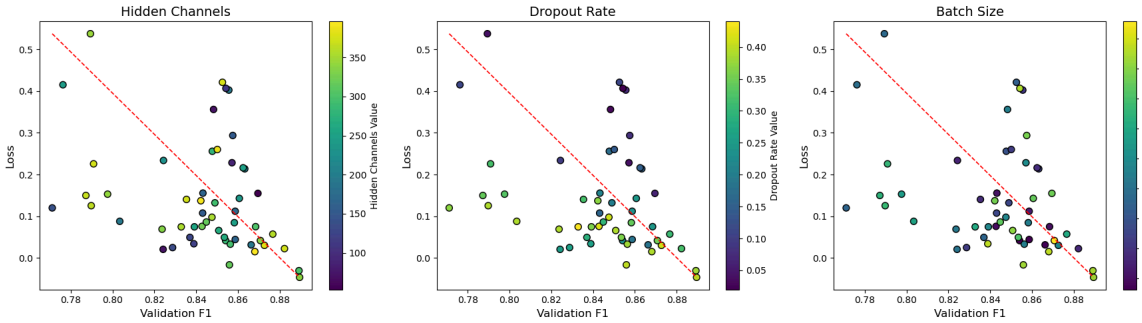
J) TRPA1 - UMP



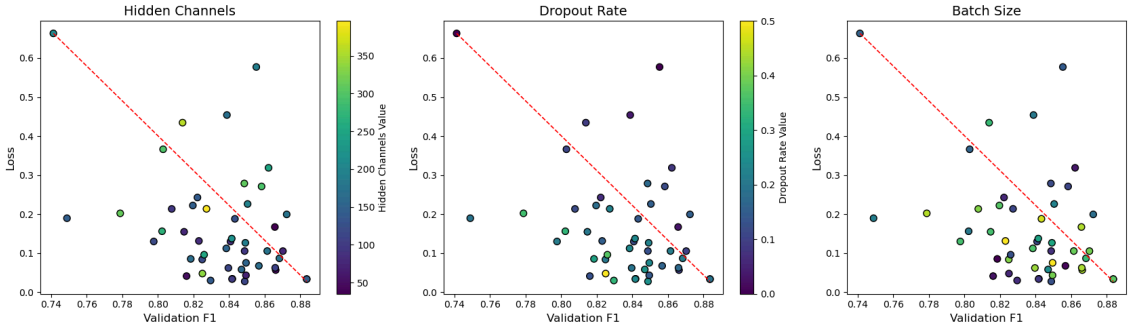
K) TRPA1 - CBMP



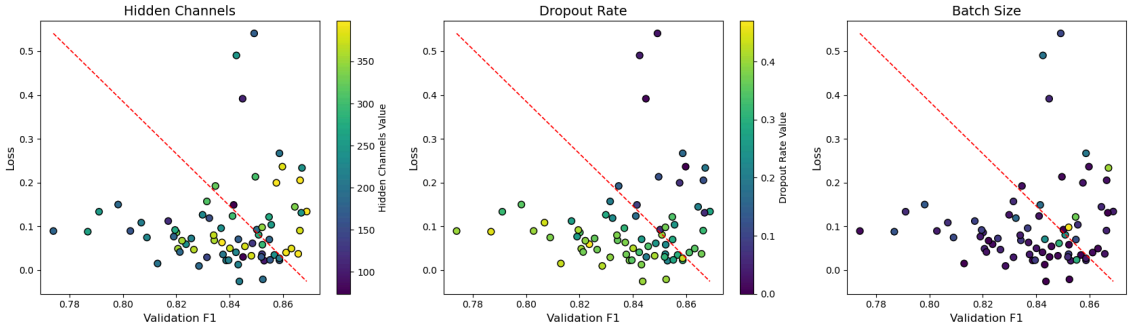
L) TRPA1 - ABMP



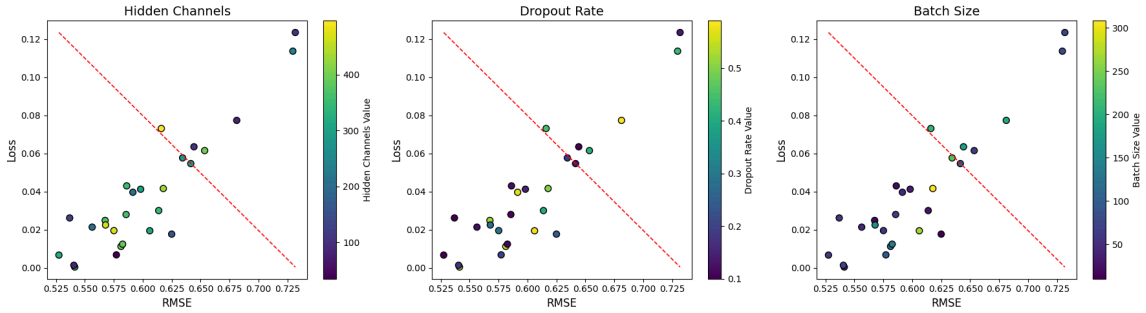
M) TRPA1- BMP + SN



N) TRPA1 - ABMP + SN



O) Lipophilicity -ABMP



Appendix C. Effects of Data Augmentation

Among the most common techniques to work with unbalanced datasets is data augmentation, which refers to the oversampling of the minority class[1]. Oversampling balances the number of class training instances by synthetically increasing the number of minority class instances.

We evaluated the impact of class augmentation by increasing the size of the less populated class by a factor of 1x across two datasets, testing this approach on two different models. Then we made a histogram to visualize the class proportions for each cluster.

According to a cluster analysis derived from a bottom-up hierarchical approach (Figure 3), the BBBP has fewer and larger distinct clusters visible at the cutoff distance, suggesting more homogeneous relationships among compounds than the BACE dataset. Taking a weighted average 0:1 class ratio per cluster, where the weight is calculated by dividing the cluster size by the dataset size, results in 2.5:1 for the BACE and 0.5:1 for the BBBP. This procedure results in increased ratios compared to the initial raw class division reported earlier (1.2:1 and 0.3:1 for BACE and BBBP, respectively)

For the BBBP dataset, augmentation improved the weighted average class ratio to 0.8:1. In contrast to the BACE dataset, where an apparent total class ratio of 1.1:1, continues to have the same weighted average ratio (2.5:1). The latter is explained by a disproportionate distribution of minority-class samples within clusters. On the other hand, class representation across clusters was uniform after augmentation in the BBBP dataset, as detailed in the histogram of cluster distributions with augmented data for both datasets.

Despite these adjustments, augmentation did not improve performance across the three key evaluation metrics for the two datasets (Figure 4). Furthermore, the true positive rates for the augmented class declined,

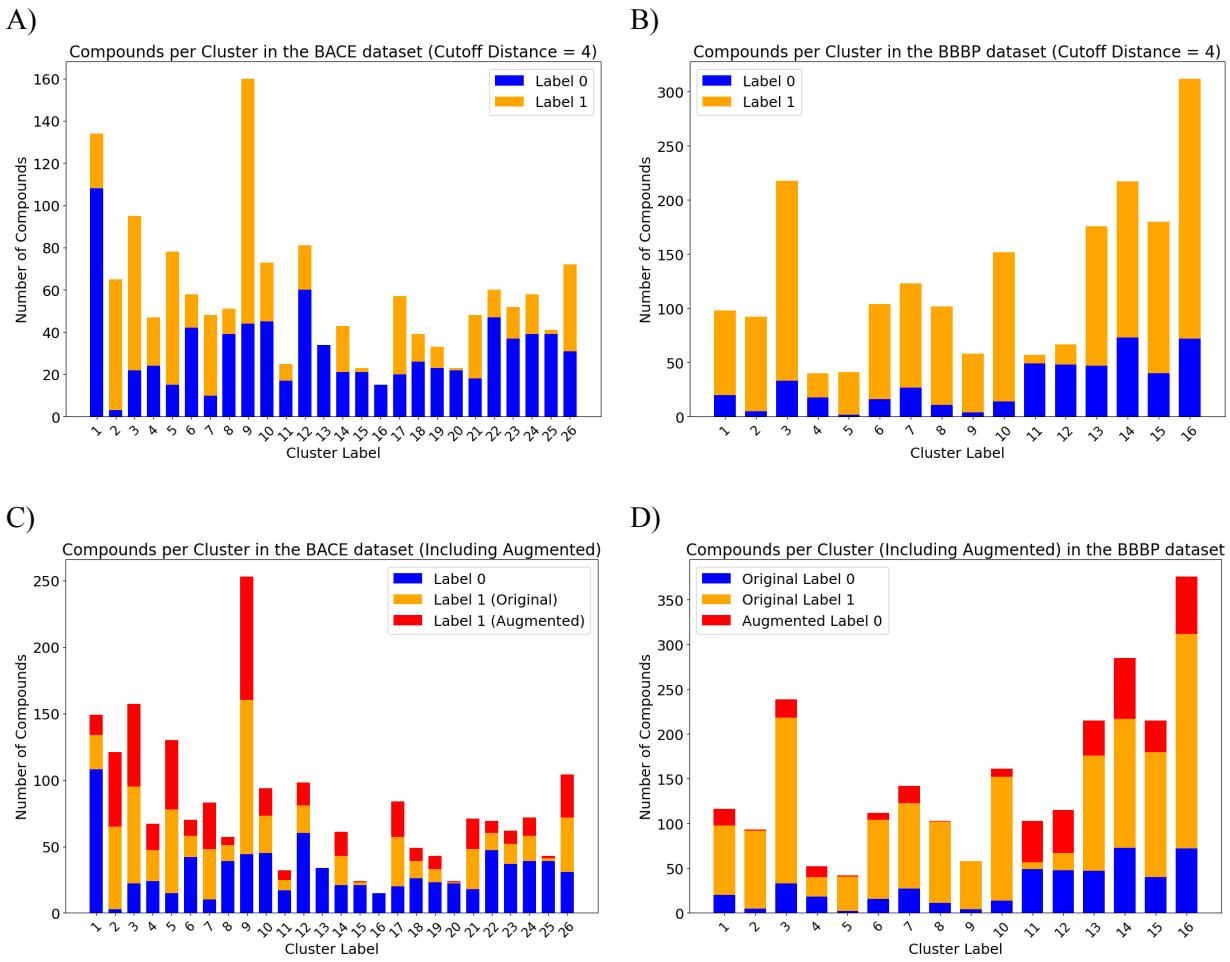


Figure 3. The structural diversity within benchmarking datasets can be visualized with hierarchical clustering and subsequent cluster analysis reveals the class imbalance at a cluster level. In A) and C) The hierarchical clustering for the BACE and BBBP datasets are shown respectively. The clusters were extracted using a distance cut-off = 4 and frequency counts on each of the two classes were made to visualize class balances. In B) and D) are the histograms for the cluster class distributions for the BACE and BBBP sets respectively.

suggesting that the approach may have negatively impacted the ability of the model to predict minority-class instances accurately. This indicates that duplicating samples within a class may introduce bias, reducing the true diversity of the class and leading to the misclassification of unseen data. Interestingly, the non-augmented class benefited, as the model may have focused on redundant features in the augmented class, ultimately improving performance for the original class data.

In conclusion, duplicating samples to address class imbalance is only effective when the primary goal is to improve the prediction of the more representative class, accepting a higher true-value rate for this one at the expense of diminished performance for the less represented.

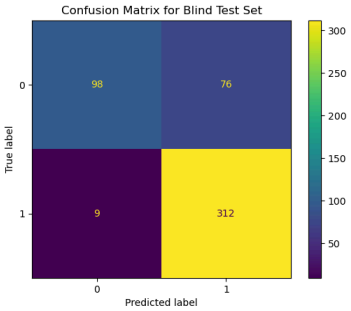
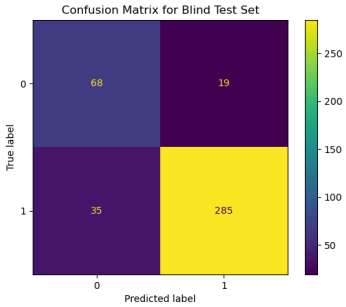
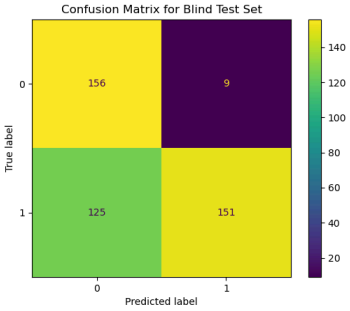
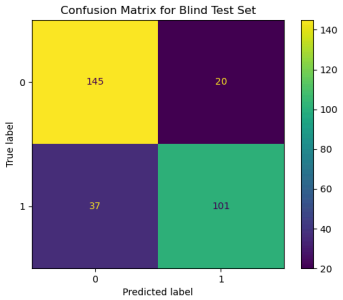
Model - Dataset-Class augmented 1x	Confusion Matrix with Augmentation	Confusion Matrix with No Augmentation
UMP-BBBP		
BMP+SN-BACE		

Figure 4. Effects of data augmentation by random re-structuration of the SMILES string as a strategy to fight class imbalance in the BACE dataset using model UMP and the BBBP dataset using model BMP+SN.

Appendix D. Final hyperparameters used for the reported performance metrics

Model-Dataset	Batch Size	Drop Out	Hidden Channels	Learning Rate	Epochs	Tested Random-Seeds
BMP-BACE	32	0.5	64	0.0032	50	[32, 42, 52, 62, 72]
BMP+SN-BACE	171	0.19	54	0.0032	50	[32, 42, 52, 62, 72]
UMP-BACE	168	0.26	224	0.0032	50	[92, 102, 112, 122, 132]
CBMP-BACE	154	0.21	111	0.0032	50	[92, 102, 112, 122, 132]
ABMP-BACE	184	0.28	205	0.0032	50	[22, 32, 42, 52, 62]
ABMP+SN-BACE	155	0.26	217	0.0032	50	[32, 42, 52, 62, 72]
BMP-BBBP	29	0.41	64	0.0032	50	[42, 52, 62, 72, 82]
BMP+SN-BBBP	157	0.21	85	0.003	50	[92, 102, 112, 122, 132]
UMP-BBBP	234	0.23	35	0.0032	50	[92, 102, 112, 122, 132]
CBMP-BBBP	58	0.37	58	0.0032	50	[92, 102, 112, 122, 132]
ABMP-BBBP	105	0.23	56	0.0032	50	[32, 42, 52, 62, 72]
ABMP+SN-BBBP	198	0.4	53	0.0032	50	[92, 102, 112, 122, 132]
BMP-TRPA1	208	0.19	171	0.0032	50	[92, 102, 112, 122, 132]
UMP-TRPA1	43	0.33	69	0.0032	50	[92, 102, 112, 122, 132]
CBMP-TRPA1	215	0.26	125	0.0032	50	[92, 102, 112, 122, 132]
ABMP-TRPA1	207	0.4	339	0.0032	50	[22, 32, 42, 52, 62]
BMP+SN-TRPA1	151	0.19	146	0.0032	50	[92, 102, 112, 122, 132]
ABMP+SN-TRPA1	212	0.19	180	0.0032	50	[32, 42, 52, 62, 72]
BMP - Lipoph.	203	0.45	362	0032	50	[92, 102, 112, 122, 132]
UMP - Lipoph.	19	0.38	179	0032	200	[92, 102, 112, 122, 132]
CBMP - Lipoph.	145	0.35	166	0032	200	[92, 102, 112, 122, 132]
ABMP - Lipoph.	65	0.10	322	0.0032	200	[92, 102, 112, 122, 132]
BMP+SN - Lipoph.	147	0.11	167	0032	200	[92, 102, 112, 122, 132]
ABMP+SN - Lipoph.	199	0.30	110	0032	200	[92, 102, 112, 122, 132]