Informed Augmentation Selection Improves Tabular Contrastive Learning

Arash Khoeini*, Shuman Peng*, Martin Ester
School of Computing Science
Simon Fraser University
akhoeini@sfu.ca shumanp@sfu.ca
ester@sfu.ca

Abstract

While contrastive learning (CL) has demonstrated success in image data, its application to tabular data remains relatively unexplored. The effectiveness of CL heavily depends on data augmentations, yet the suitability of tabular augmentation techniques for contrastive learning remains unclear. In this study, we assess the compatibility of various tabular augmentation techniques with CL by examining their impact on feature space characteristics (i.e., uniformity and alignment) which serve as proxies for downstream performance. Our investigation reveals that augmentations impact feature space quality, and that achieving a balance between uniformity and alignment is essential for good downstream performance. We then propose a novel framework for selecting augmentation combinations that strike this balance. Experimental results on 21 tabular datasets from the OpenML-CC18 benchmark and on the TCGA cancer genomics dataset consistently demonstrate the effectiveness of our proposed framework in enhancing downstream performance.

1 Introduction

Deep neural networks excel in classification and regression tasks with large labeled datasets, but labeled data is often scarce, while unlabeled data is abundant. To address this, self-supervised learning (SSL) pre-trains models on unlabeled data in image and text domains, allowing effective downstream task performance with minimal labeled data (Noroozi and Favaro, 2016; Gidaris et al., 2018; Chen et al., 2020a; Grill et al., 2020; Chen et al., 2020b; Chen and He, 2020; Devlin et al., 2018). While SSL has mainly been applied to images and text, it has recently expanded to tabular data. Two key SSL approaches for tabular data are pretext training and contrastive learning (CL). While most tabular SSL methods rely on pretext training, wherein the model learns to solve auxiliary tasks on unlabeled data, CL offers the advantage of not requiring the careful design of pretext tasks. However, CL for tabular data has been relatively underexplored compared to pretext-based approaches.

CL-based SSL methods are widely used in the image domain, and they aim to bring similar (positive) examples closer and separate dissimilar (negative) examples in the representation space He et al. (2020); Chen et al. (2020b). CL's effectiveness relies heavily on data augmentations (Chen and Li, 2020; Caron et al., 2020; Jaiswal et al., 2020; Tian et al., 2020; Wang and Qi, 2021), which are employed to generate positive examples without the need for additional annotation. While CL-pretrained models are ultimately assessed on downstream task performance, real-world applications often lack ground truth labels during CL-based pre-training, necessitating proxies for downstream performance. Recent studies show that CL induces feature spaces with properties like uniformity and alignment, which correlate with downstream task performance Wang and Isola (2020). Superior downstream performance is associated with a balanced feature space exhibiting high degrees of

*Equal Contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

uniformity and alignment, observed in image and text data Wang and Isola (2020). However, the impact of feature uniformity and alignment on downstream performance remains unexplored within the tabular data domain, and there is a lack of investigation into the appropriateness of tabular augmentations for achieving these desirable properties.

In this work, we aim to understand which tabular data augmentation techniques are suitable for tabular CL by investigating the feature space characteristics generated by each augmentation. We employ a widely used form of the InfoNCE contrastive loss (van den Oord et al., 2018; Chen et al., 2020b), following existing tabular CL works (Bahri et al., 2022), and we investigate six prevalent tabular data augmentation techniques found in the literature in terms of feature space characteristics.

We observe that the augmentation techniques have a strong impact on the quality of feature space characteristics – some augmentation techniques favor feature alignment while others favor the uniformity of feature spaces. Based on our observations and also findings from Wang and Isola (2020) showcasing that a good balance between uniformity and alignment is crucial for good downstream performance, we devise a novel and practical framework for selecting a combination of augmentation techniques that pairs an augmentation that achieves the best uniformity with one that achieves the best alignment, striking a good balance of uniformity and alignment. To the best of our knowledge, this is the first study to systematically guide the selection and combination of different types of augmentations for tabular contrastive learning. We evaluate our framework on 21 tabular datasets from OpenML-CC18, and our experiments demonstrate that our framework for selecting a suitable pair of augmentations for CL pre-training leads to better classification accuracy on downstream tasks compared to performing CL pre-training using single augmentation techniques. Moreover, models pre-trained with our selected augmentation combination consistently achieve better downstream performance across various corruption rates, an important hyper-parameter for augmenting tabular data. We further extend and evaluate our framework on The Cancer Genome Atlas (TCGA) dataset that consists of cancer patient gene expression data. We show that performing CL pre-training using our proposed framework consistently achieves better downstream performance across a number of corruption rates than pre-training using single augmentations.

2 Empirical Analysis of Feature Quality Metrics

In contrastive-learned feature spaces, features are more evenly distributed compared to those in supervised-learned spaces, with similar samples positioned closer together on the unit hypersphere. These properties, termed "feature uniformity" and "feature alignment" by Wang and Isola (2020), are key to the quality of learned representations. Feature uniformity ensures a balanced distribution that preserves information, while feature alignment measures how close similar samples are, therefore assessing the semantic structure. Wang and Isola (2020) shows that spaces with higher uniformity and alignment perform better in downstream tasks in image and text domains. Unlike Wang and Isola (2020) that optimize these characteristics, we analyze the uniformity and alignment of contrastive-learned spaces using negated version of their loss terms, where higher values are preferred.

$$\text{Alignment} \triangleq - \mathop{\mathbb{E}}_{(x,x')\sim p_{pos}} [\|f(x) - f(x')\|_2^{\alpha}], \quad \alpha > 0^2 \tag{1}$$

$$\text{Uniformity} \triangleq - \log \mathop{\mathbb{E}}_{(x,x')\sim p} [e^{-t\|f(x)-f(x')\|_2^2}], \quad t > 0 \tag{2}$$

In the rest of this section, we undertake empirical analysis aimed at addressing two fundamental research questions: (1) How do data augmentation techniques influence the feature uniformity and alignment of contrastive pre-trained models for tabular data (Section 2.3)? and (2) Are feature uniformity and alignment reliable indicators of downstream performance in the tabular domain (Section 2.4)?

2.1 Augmentations

In this paper, we explore six tabular data augmentation techniques: joint replacement (**JR**), marginal replacement (**MR**), joint Mixup (**JM**), marginal Mixup (**MM**), Gaussian noise (**GN**), and feature

[2]In our empirical analyses, $\alpha = 2$ is used to evaluate feature alignment.

dropout (**DO**) (Bahri et al., 2022; Verma et al., 2020). To perform augmentations, a binary feature mask is first defined using corruption rate $c \in (0, 1)$, where each augmentation is applied to $c \times 100\%$ of the input features.

In **MR** each masked feature is replaced with a random value from its marginal distribution (Bahri et al., 2022). In **JR** masked features are replaced with values from a single random sample (Yoon et al., 2020). In **MM** and **JM**, the original masked features are interpolated with features of another sample using Mixup rate $\lambda \in [0, 1]$ Zhang et al. (2017); Verma et al. (2020); Lee et al. (2020). In **GN**, noise $\mathcal{N}(0, 0.5^2)$ is added to features (Bahri et al., 2022). In DO, selected features are set to zero.

2.2 Datasets, model training and evaluation details

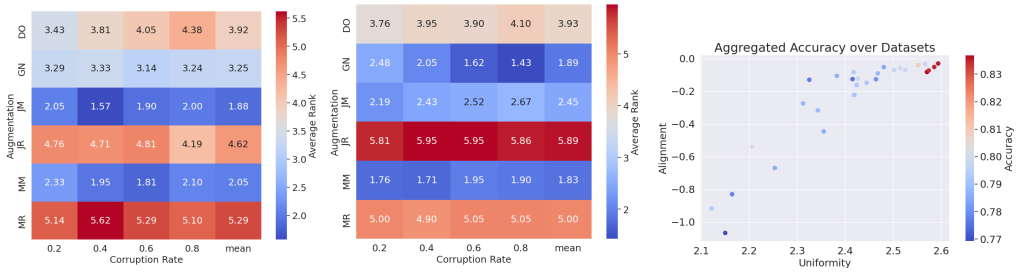
We use 21 of the real-world tabular datasets for classification tasks from the OpenML-CC18 benchmark. We follow the authors of (Bahri et al., 2022) for splitting each tabular dataset into unlabelled pre-training and labelled downstream fine-tuning and testing sets, with more details provided in Appendix D.3. We follow Bahri et al. (2022) and Chen et al. (2020b) for model design, using a neural network with a feature extractor f, and either a pre-training head g or a classification head h. During pre-training, the model consists of f and g, while fine-tuning uses f and h. All components are fully connected feed-forward networks with ReLU activations. Models were trained on each dataset with four corruption rates ($c = 0.2, 0.4, 0.6, 0.8$) and six augmentation techniques, yielding 24 models per dataset. Each model was trained ten times with different splits, and classification accuracy was used to evaluate downstream performance. Rank-based heat maps were generated to compare augmentation techniques across corruption rates, considering metrics such as uniformity, alignment, and accuracy.

2.3 How does data augmentation impact feature spaces?

Our investigation into augmentation techniques and corruption rates for contrastive learning on tabular data reveals that data augmentations have a strong impact on feature spaces (illustrated in Figure 1). Specifically, some augmentations (e.g., JM) consistently improve feature uniformity while others (e.g., MM) yield better feature alignment. In addition, most augmentations are robust to different corruption rates for feature uniformity and alignment, with the exception of GN for alignment and DO for uniformity.

2.4 Are feature uniformity and alignment indicative of downstream performance?

We observe that models with feature spaces that exhibit greater uniformity and alignment consistently achieve higher classification accuracy with a linear classifier. In comparison, models with less uniform and aligned feature spaces achieve lower classification accuracy. In particular, lower downstream performance is observed for models with higher alignment than uniformity and for models with higher uniformity than alignment. This aligns with the findings of (Wang and Isola, 2020) from the image and text domain, which show that improving only one of uniformity and alignment degrades downstream performance. Multiple regression analysis supports these results, showing a positive correlation between uniformity (coefficient: 0.0511) and alignment (coefficient: 0.0466) with downstream accuracy.



(a) Uniformity (b) Alignment (c) Alignment

Figure 1: Comparing augmentation techniques based on mean ranking across 21 OpenML-CC18 datasets for (a) feature uniformity and (b) alignment at different corruption rates. The last column in each heatmap shows the mean ranking across all corruption rates. Lower ranks indicate better performance. (c) Aggregated accuracy over all 21 OpenML-CC18 datasets.

3 Our Framework for Tabular CL

In the preceding sections, we made two key observations: (i) data augmentations significantly affect feature space properties (Section 2.3), and (ii) balancing high degrees of feature uniformity and alignment is crucial for optimal downstream performance (Section 2.4). Based on these insights, we propose a novel framework for tabular contrastive learning that strikes a balance between uniformity and alignment, while also attaining high levels of both, through data augmentations.

Our approach selects augmentations in a dataset-specific manner by identifying (1) the augmentation that maximizes uniformity (A_1) and (2) the one that maximizes alignment (A_2), based on mean rankings across corruption rates (Figure 1). During contrastive pre-training with InfoNCE loss, a positive sample is generated by randomly applying either A_1 or A_2, ensuring a balance between uniformity and alignment, leading to improved downstream performance.

4 Experimental Results

In this section, we validate the effectiveness of our framework for tabular contrastive learning through empirical experiments on real-world tabular datasets. We conduct experiments on 21 OpenML-CC18 tabular datasets, with aggregated results presented in Figure 2. Our framework selects a different pair of augmentation for each dataset, which we denote as Ours in Figure 2. Additionally, we evaluate our framework on the more complex The Cancer Genome Atlas (TCGA) dataset, which includes multiple distinct downstream tasks (Table 1). Our experiments show that our framework effectively selects augmentation pairs that enhance the feature space characteristics of contrastive pre-trained models, consequently improving their downstream performance on tabular datasets. Details of experiments and full experimental results are available in Section D of the Appendix.

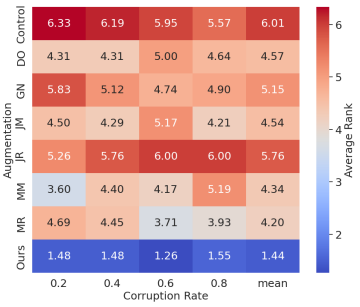


Figure 2: Ranking of models' downstream accuracy across 21 OpenML-CC18 tabular datasets. Each column shows mean rankings for augmentation techniques and the supervised baseline (Control) with 10 independent runs per model. The "mean" column aggregates ranks across all corruption rates. Lower ranks (more blue) indicate better performance.

5 Conclusion

In this study, we conducted a comprehensive analysis of the impact of data augmentations on contrastive-learned feature spaces in tabular datasets, evaluating six widely-used augmentation techniques. Our empirical findings confirm a positive correlation between feature uniformity/alignment and downstream task accuracy in the tabular domain. Moreover, we demonstrate that models striking a balance between high degrees of feature uniformity and alignment exhibit superior downstream task accuracy, and that some augmentations promote higher uniformity while others enhancing alignment. Building upon these insights, we devised a novel framework for selecting augmentation combinations that facilitate the attainment of a balanced feature space in models. Our experiments on 21 tabular datasets from OpenML-CC18 and on the TCGA cancer genomics show that the augmentation combination identified by our framework improves the balance between uniformity and alignment, boosting downstream performance over single augmentations.

Augmentation	Task1	Task2	Task3	Task4	Task5
Control	0.750 ± 0.016	0.710 ± 0.060	0.689 ± 0.052	0.709 ± 0.074	0.712 ± 0.030
DO	0.701 ± 0.064	0.713 ± 0.090	0.661 ± 0.090	0.697 ± 0.061	0.746 ± 0.076
GN	0.706 ± 0.070	0.693 ± 0.090	0.680 ± 0.092	0.685 ± 0.077	0.736 ± 0.084
JM	0.758 ± 0.052	0.795 ± 0.074	0.710 ± 0.063	0.756 ± 0.056	**0.807 ± 0.057**
JR	0.744 ± 0.062	0.786 ± 0.072	0.706 ± 0.083	0.753 ± 0.037	0.788 ± 0.057
MM	0.701 ± 0.072	0.715 ± 0.085	0.676 ± 0.072	0.723 ± 0.053	0.730 ± 0.093
MR	0.685 ± 0.042	0.724 ± 0.065	0.611 ± 0.109	0.624 ± 0.090	0.723 ± 0.090
JM+GN (ours)	**0.761 ± 0.044**	**0.807 ± 0.040**	**0.721 ± 0.075**	**0.771 ± 0.064**	**0.802 ± 0.066**

Table 1: Mean AUROC and standard deviation for each augmentation on TCGA downstream tasks, averaged across corruption rates and 5 runs per model.

Acknowledgment

This research was supported by the NSERC Discovery Grant. We would like to thank Shichong Peng for providing thorough feedback throughout the course of this research.

References

Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. 2022. Scarf: Self-Supervised Contrastive Learning using Random Feature Corruption. In *International Conference on Learning Representations*. https://openreview.net/forum?id=CuV_qYkmKb3

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *ArXiv* abs/2006.09882 (2020).

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020b. A Simple Framework for Contrastive Learning of Visual Representations. *ArXiv* abs/2002.05709 (2020).

Ting Chen and Lala Li. 2020. Intriguing Properties of Contrastive Losses. In *Neural Information Processing Systems*.

Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020a. Improved Baselines with Momentum Contrastive Learning. *ArXiv* abs/2003.04297 (2020).

Xinlei Chen and Kaiming He. 2020. Exploring Simple Siamese Representation Learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 15745–15753.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* (2018).

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems* 34 (2021), 18932–18943.

Jean-Bastien Grill, Florian Strub, Florent Altch'e, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *ArXiv* abs/2006.07733 (2020).

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A Survey on Contrastive Self-supervised Learning. *ArXiv* abs/2011.00362 (2020).

Kyungeun Lee, Ye Seul Sim, Hyeseung Cho, Suhee Yoon, Sanghyu Yoon, and Woohyung Lim. 2023. Binning as a Pretext Task: Improving Self-Supervised Learning in Tabular Domains. In *NeurIPS 2023 Second Table Representation Learning Workshop*. https://openreview.net/forum?id=btK3lk5puP

Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. 2020. i-mix: A domain-agnostic strategy for contrastive representation learning. *arXiv preprint arXiv:2010.08887* (2020).

Kushal Alpesh Majmundar, Sachin Goyal, Praneeth Netrapalli, and Prateek Jain. 2022. MET: Masked Encoding for Tabular Data. In *NeurIPS 2022 First Table Representation Workshop*. https://openreview.net/forum?id=vMHs3HR7r0A

Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*. Springer, 69–84.

Hossein Sharifi-Noghabi, Parsa Alamzadeh Harjandi, Olga Zolotareva, Colin C Collins, and Martin Ester. 2021. Out-of-distribution generalization from labelled and unlabelled gene expression data for drug response prediction. *Nature Machine Intelligence* 3, 11 (2021), 962–972.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning. *ArXiv* abs/2005.10243 (2020).

Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. 2021. SubTab: Subsetting Features of Tabular Data for Self-Supervised Representation Learning. In *Neural Information Processing Systems*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *ArXiv* abs/1807.03748 (2018).

Vikas Verma, Minh-Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc V. Le. 2020. Towards Domain-Agnostic Contrastive Learning. In *International Conference on Machine Learning*.

Feng Wang and Huaping Liu. 2020. Understanding the Behaviour of Contrastive Loss. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 2495–2504.

Tongzhou Wang and Phillip Isola. 2020. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *International Conference on Machine Learning*.

Xiao Wang and Guo-Jun Qi. 2021. Contrastive Learning With Stronger Augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2021), 5549–5560.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics* 45, 10 (2013), 1113–1120.

Jing Wu, Suiyao Chen, Qi Zhao, Renat Sergazinov, Chen Li, Shengjie Liu, Chongchao Zhao, Tianpei Xie, Hanqing Guo, Cheng Ji, et al. 2024. SwitchTab: Switched Autoencoders Are Effective Tabular Learners. *arXiv preprint arXiv:2401.02013* (2024).

You Wu, Omid Bazgir, Yongju Lee, Tommaso Biancalani, James Lu, and Ehsan Hajiramezanali. 2023. Multitask-Guided Self-Supervised Tabular Learning for Patient-Specific Survival Prediction. In *NeurIPS 2023 Second Table Representation Learning Workshop*. `https://openreview.net/forum?id=3gBqMkELhZ`

Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. 2020. VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain. In *Neural Information Processing Systems*.

Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. 2017. mixup: Beyond Empirical Risk Minimization. *ArXiv* abs/1710.09412 (2017).

A Appendix

B Preliminaries

Problem setup Let $D^u = \{X^u\}$ denote an unlabeled tabular dataset of N^u samples with d input features (covariates), and let $D^l = \{X^l, Y\}$ denote a labeled tabular dataset of N^l samples with the same d features and with class labels $Y \in \{1, ..., C\}$, where C is the number of classes. The unlabeled dataset D^u is assumed to be far larger than the labeled dataset D_l: $N^u \gg N^l$. The goal is to use a contrastive learning method to pre-train a feature extractor f using D^u and then apply the pre-trained f to a downstream classification task based on D^l. D^l is used to train a linear classifier on top of feature extractor f to perform downstream classification. Optionally, D^l is also used to fine-tune the feature extractor. In our experiments, we fine-tune the feature extractor to align with standard practices in contrastive learning Bahri et al. (2022); Chen et al. (2020b).

B.1 Contrastive Learning

Contrastive learning (CL) is a self-supervised representation learning technique that learns a feature space from *unlabeled data* such that semantically similar (positive) samples are close in proximity and semantically dissimilar (negative) samples are far apart in the feature space. CL's effectiveness relies heavily on data augmentation techniques (Chen and Li, 2020; Caron et al., 2020; Jaiswal et al., 2020; Tian et al., 2020; Wang and Qi, 2021). These techniques enable the generation of positive examples without additional annotation, thereby producing multiple distinct views of a single example. We discuss augmentation techniques in Section 2.1.

In this work, we use a widely adopted variant of contrastive loss Chen et al. (2020b); Wang and Isola (2020); Bahri et al. (2022) that is derived from the InfoNCE loss (van den Oord et al., 2018). The contrastive loss for a single positive pair (x_i, x_i') is given by:

$$\mathcal{L}_{\text{CL}} = -\log \frac{\exp(\text{sim}(v_i, v_i')/\tau)}{\exp(\text{sim}(v_i, v_i')/\tau) + \sum_{k=1}^{K} \exp(\text{sim}(v_i, v_k)/\tau)} \tag{3}$$

where v_i is the representation of an original sample x_i, and v_i' is the representation of a positive counterpart x_i' (an augmented version of x_i). Negative counterparts x_k are represented as v_k in the representation space, which encompass all samples $x_k, k \in [1, K]$ that neither correspond to x_i or its augmentations. Here, $sim(v_i, v_j)$ denotes a similarity measure between the representations (commonly the cosine similarity), and τ is a temperature scaling parameter that adjusts the sensitivity of the loss function to differences in similarity. The denominator sums over one positive and K negative pairs, effectively normalizing the similarity of the positive pair against the sum of similarities across all pairs.

C Related Works

Tabular self-supervised representation learning Self-supervised learning (SSL) has garnered significant attention for learning expressive representations from unlabelled data in image and text domains. Recently, there has been a surge in efforts to extend SSL to tabular data, with methods falling into three main categories. The first category leverages pretext tasks, often involving the reconstruction of original samples from corrupted versions or predicting the applied corruption Yoon et al. (2020); Lee et al. (2023); Majmundar et al. (2022); Wu et al. (2024). The self-supervised VIME method by Yoon et al. (2020) introduces two tabular-compatible pretext tasks: recovering the binary mask vector applied to the original sample and reconstructing the original input. Additionally, Lee et al. (2023) employs binning as a pretext task, replacing continuous tabular feature values with corresponding bin indices. The work by Majmundar et al. (2022) introduces a masked input reconstruction pretext task for tabular data.

The second category employs contrastive learning (CL), aiming to map semantically similar examples close in the latent representation space and dissimilar examples far apart. The SCARF method, proposed by Bahri et al. (2022), utilizes the marginal replacement augmentation technique, randomly sampling values from the empirical marginal distribution of features to replace a portion $c \in (0, 1]$ of the original input features. Based on the SimCLR method Chen et al. (2020b), SCARF optimizes the InfoNCE contrastive loss (van den Oord et al., 2018). Methods introduced in Verma et al. (2020); Lee et al. (2020) employ variants of Mixup noise for tabular data augmentation.

The third category integrates both pretext training and CL, resulting in hybrid methods with multiple components in the loss functions. Examples include the SubTab method Ucar et al. (2021) and its follow-up work Wu et al. (2023).

Differing from existing tabular CL works, we aim to gain insights into the effectiveness of different tabular augmentation techniques for contrastive pre-training by using feature uniformity and alignment as proxies for downstream performance. Building upon these insights, we adopt a systematic approach to selecting combinations of augmentation techniques, diverging from the conventional practice of employing a single technique (Yoon et al., 2020; Bahri et al., 2022).

Factors influencing contrastive feature spaces Recent work by Wang and Liu (2020) has demonstrated that the properties of feature space, such as uniformity and tolerance, can be adjusted by tuning the temperature τ hyper-parameter of contrastive losses. Feature tolerance, akin to feature

Algorithm 1 Pseudo-code for pre-training function

Require: X ▷ X is pre-training dataset
  function PRE_TRAIN(X)
    for each iteration **do**
      $p \leftarrow$ RANDOM$(0, 1)$ ▷ $p \in [0, 1]$
      if $p > 0.5$ **then**
        $X\_pos \leftarrow$ A_1(X)
      else
        $X\_pos \leftarrow$ A_2(X)
      end if
      $Z \leftarrow$ MODEL(X)
      $Z\_pos \leftarrow$ MODEL(X_pos)
      $loss \leftarrow$ CONTRASTIVE_LOSS(Z, Z_pos)
      OPTIMIZER$(loss)$
    end for
  end function

alignment as discussed in Wang and Isola (2020), requires labels for measurement, unlike alignment, which does not necessitate label access. Smaller temperature τ values lead to more uniform feature spaces, but with less tolerance to samples within the same category, meaning that similar samples have low degrees of similarity in the feature space. Conversely, larger τ values lead to less uniform feature spaces but with greater tolerance to samples from the same category Wang and Liu (2020). In this work, we explore the impact of data augmentations on the uniformity and alignment of feature spaces in tabular data, and investigate how these factors influence downstream performance. Different from (Wang and Liu, 2020; Wang and Isola, 2020), which only studied the behavior of feature space properties on image and text modalities, we study their behavior on tabular data.

D Complete Experimental results

D.1 Training Details

Closely following SCARF Bahri et al. (2022), we use a feed-forward neural network with three hidden layers, each with 256 units, and an output layer of the same size, followed by a linear classifier. We include a pre-training head g, with a hidden layer and a output layer of size 256, which is removed after pre-training and replaced with a classifier head. Using this pre-training head is introduced in SimCLR Chen et al. (2020b), and became the common approach in contrastive learning.

For FT-Transformer architecture we used the official implementations of its authors with all the default hyper-parameters [3]. We use the Adam optimizer with a learning rate of 0.001 for pre-training and fine-tuning, with a batch size of 128. Consistent with Bahri et al. (2022), we employ the contrastive pre-training loss for early stopping. However, in contrast, we employ the classification loss rather than classification accuracy to guide early stopping during fine-tuning for downstream tasks. We set a maximum of 1000 epochs for pre-training and 200 epochs for fine-tuning, implementing early stopping with a patience of 5. We monitor pre-training and fine-tuning loss on a validation set held out for this purpose. The same validation set is used for both stages, but for pre-training, we create a static validation set by generating 10 augmentations for each sample, resulting in 10 positive pairs per sample. This static set is used to evaluate our method after each epoch during pre-training.

All reported results use the feed-forward network mentioned above by default. Results using an FT-Transformer backbone will be explicitly noted.

D.2 Pseudo-code of our Framework

D.3 Experimental Results on the OpenML-CC18 Datasets

Datasets We use 21 of the real-world tabular datasets for classification tasks from the OpenML-CC18[4] benchmark. See Table 3 of the Supplementary for details. Following Bahri et al. (2022), we

[3] https://github.com/yandex-research/rtdl-revisiting-models

[4] https://docs.openml.org/benchmark/

remove the MNIST, Fashion-MNIST, and CIFAR10 datasets from OpenML-CC18 to focus on true tabular datasets. The datasets for this study were chosen to ensure relevance and applicability of the analysis, based on two primary criteria: (1) datasets containing only numerical features with at least 50 features, or (2) datasets including at least one categorical feature with a total feature count exceeding 15. This selection was driven by the need for datasets that support specific augmentation techniques applicable exclusively to numerical features (e.g., Gaussian noise) and to ensure the datasets were sufficiently high-dimensional for robust analysis. Following the authors of Bahri et al. (2022), for each tabular dataset 70% of the entire dataset is used for unlabelled pre-training,[5] 25% of the 70% pre-training data is used to perform supervised fine-tuning of the model for downstream classification, 10% of the entire dataset is used for validation of the fine-tuned models, and the remaining 20% of the dataset is used for testing the model's downstream classification performance.

D.3.1 Baselines

In addition to the six augmentations mentioned in Section 2.1, we also train a model called Control for comparison. Control uses the same architecture and hyper-parameters as the other methods. However, it differs in that it is trained using only the labeled data (in the fine-tuning/training set of a downstream task) through supervised learning. In other words, the main difference between Control and the other baseline models is that Control starts fine-tuning with a randomly initialized feature extractor, while the others begin with a feature extractor that has been pre-trained using contrastive learning and specific augmentation techniques. Therefore Control is not affected by different corruption rates. The purpose of including Control is to assess the benefits of pre-training the feature extractor with the unlabeled data.

Following our framework, we select a pair of augmentation techniques for each individual dataset (shown in Table 2). This pair comprises augmentations with the highest uniformity and alignment for the dataset. We denote the selected pair as "Ours" and observe that it consistently achieves better downstream accuracy than any single augmentation technique. This is shown by the ranking heat map in Figure 2. Further, we plot the mean uniformity and alignment ranks for each augmentation technique and our combination to check which augmentations enable a CL-trained model to achieve a good balance between uniformity and alignment.

We observe that our combination leads to the most uniform and aligned feature spaces, shown by the red dot closest to the origin in Figure 6 (the corresponding rank-based heat maps are included in the Supplementary Figure 3). This observation corroborates our findings in Section 2 and those from image datasets Wang and Isola (2020), where CL-pretrained models with a balance between uniformity and alignment lead to better downstream classification performance. We conducted the same experiment, but this time using the FT-Transformer feature extractor f, a transformer architecture specifically designed for tabular datasets Gorishniy et al. (2021) (see Section D.1 for details). The results were similar, with the pairs selected by our framework achieving the highest average downstream accuracy across the datasets. The results for the FT-Transformer feature extractor are shown in Figure 8.

D.4 Further Experiments on Cancer Patient Genomics Datasets

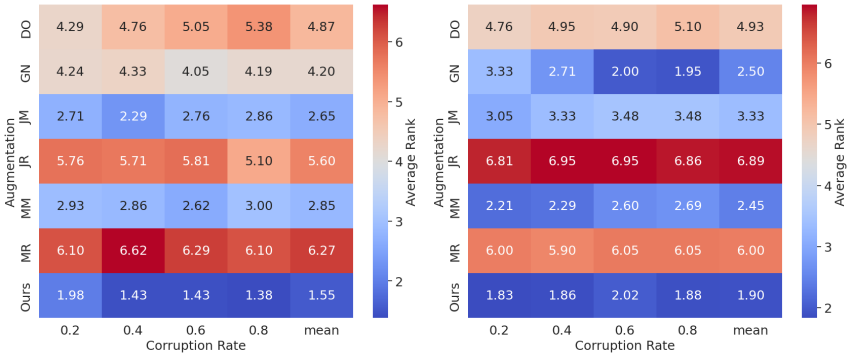
D.4.1 Experimental Setup

Dataset In addition to the 21 tabular benchmark datasets employed in Section 2, we use The Cancer Genome Atlas (TCGA) dataset Weinstein et al. (2013) to evaluate our framework in biomedical settings, and perform a series of experiments across 5 distinct downstream tasks. More specifically, we formulate one binary classification downstream task for each cancer type, resulting in five separate cancer-type classification tasks. The TCGA dataset serves as an expansive resource aimed at enhancing our grasp of cancer's molecular dynamics via genome sequencing analysis. In the TCGA dataset, rows represent patients, columns (attributes) represent the genes, and the continuous attribute values represent the level of expression of a particular gene in a patient. Our pre-training data combines five specific cancer datasets: Lung Adenocarcinoma, Breast Invasive Carcinoma, Pancreatic Adenocarcinoma, Prostate Adenocarcinoma, and Kidney Carcinoma. The data is preprocessed following Sharifi-Noghabi et al. (2021), except we do not reduce the number of genes as they did. The combined dataset contains 2721 samples (patients) and 18,312 shared features (genes), and is

[5]These datasets are labelled, but we withhold their labels for pre-training.

Table 2: Datasets and Selected Pair of Augmentations

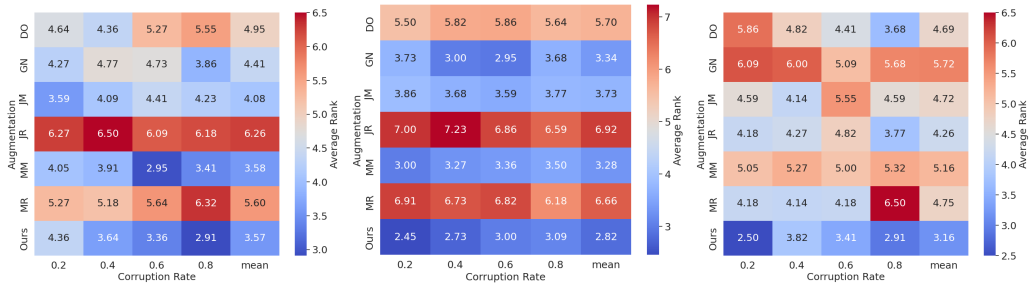
Dataset	Selected Pair of Augmentations
DNA	JM+MM
Bioresponse	MM+GN
madelon	JM+MM
mfeat-pixel	MM+GN
cnae-9	JM+MM
first-order-theorem-proving	JM+MM
isolet	MM+GN
mfeat-factors	MM+GN
mfeat-fourier	JM+GN
mfeat-karhunen	JM+MM
optdigits	JM+MM
ozone-level-8hr	JM+GN
semeion	JM+GN
spambase	JM+MM
nomao	MM+GN
har	JM+GN
sick	MM+MM
eucalyptus	MM+JM
credit-g	MM+GN
cylinder-bands	MM+JM
bank-marketing	MM+JM



(a) Uniformity (b) Alignment

Figure 3: Uniformity and alignment rank-based heat maps comparing our combination of augmentations (denoted by "Ours") with single augmentations on the OpenML-CC18 benchmarking datasets.

standardized using Z-score normalization. Statistics of five cancer datasets are shown in Table 4. For each task, we create balanced training, validation, and test sets, consisting of 100, 50, and 100 samples, respectively. To ensure the integrity of the model's learning process, we exclude these 250 samples from the pre-training dataset, guaranteeing that all examples in the downstream tasks are previously unseen by the model. We assess the binary classification performance for each downstream task using two metrics: the Area Under the ROC curve (AUROC) and classification accuracy.



(a) Uniformity (b) Alignment (c) Accuracy

Figure 4: Uniformity, alignment and accuracy rank-based heat maps comparing our combination of augmentations with single augmentations on the OpenML-CC18 benchmarking datasets. Here, a FT-Transformer model was used as the encoder.

Implementation and Training Details In this experiment, we adopt a model architecture akin to the one outlined in Bahri et al. (2022), detailed in Section D.1. Details on the model and training procedure are in Section D.1 of the Supplementary. Our primary objective is to assess the comparative efficacy of various augmentation methods rather than striving for the highest classification accuracy; therefore, we refrain from extensive hyper-parameter tuning.

Table 3: The 16 tabular datasets from OpenML-CC18 with numerical features that were utilized in our empirical analyses.

Dataset	# of Samples	# of Features	# of Categorical Features	# of Classes
mfeat-pixel	2000	241	0	10
bioresponse	3751	1777	0	2
first-order-theorem-proving	6118	52	0	6
madelon	2600	501	0	2
spambase	4601	58	0	2
mfeat-fourier	2000	77	0	10
mfeat-factors	2000	217	0	10
ozone-level-8hr	2534	73	0	2
mfeat-karhunen	2000	65	0	10
dna	3186	181	0	3
isolet	7797	618	0	26
optdigits	5620	65	0	10
semeion	1593	257	0	10
har	10299	562	0	6
cnae-9	1080	857	0	9
nomao	34465	119	0	2
sick	3772	30	23	2
eucalyptus	736	19	6	5
bank-marketing	45211	17	9	2
credit-g	1000	21	14	2
cylinder-bands	540	40	20	2

Table 4: Five TCGA datasets that were utilized in our experiments.

Dataset	Number of Samples	Number of Features
TCGA_LUAD	507	18312
TCGA_BRCA	1051	18312
TCGA_PAAD	131	18312
TCGA_PRAD	498	18312
TCGA_KIRC	534	18312

D.4.2 Compelete Experimental Results

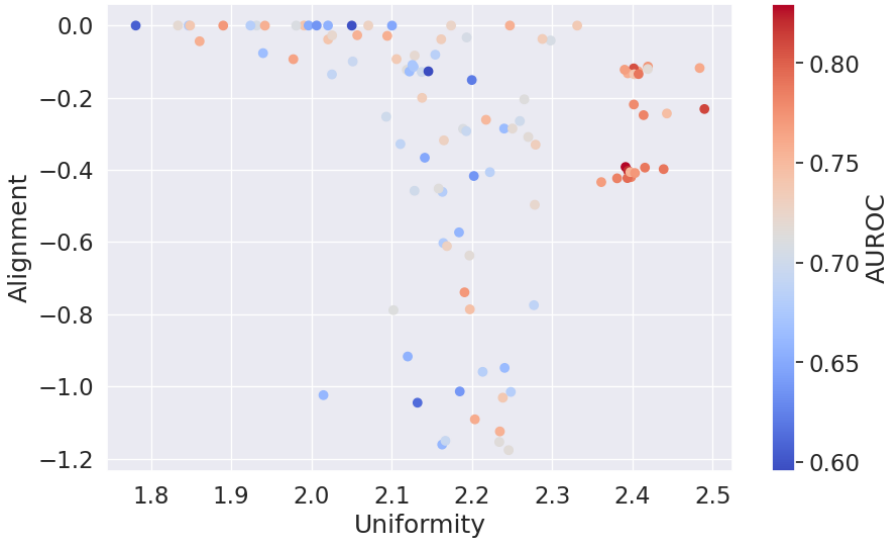


Figure 5: Higher uniformity and alignment correlate with improved downstream performance on TCGA datasets. Downstream AUROC values represent the average across 5 tasks.

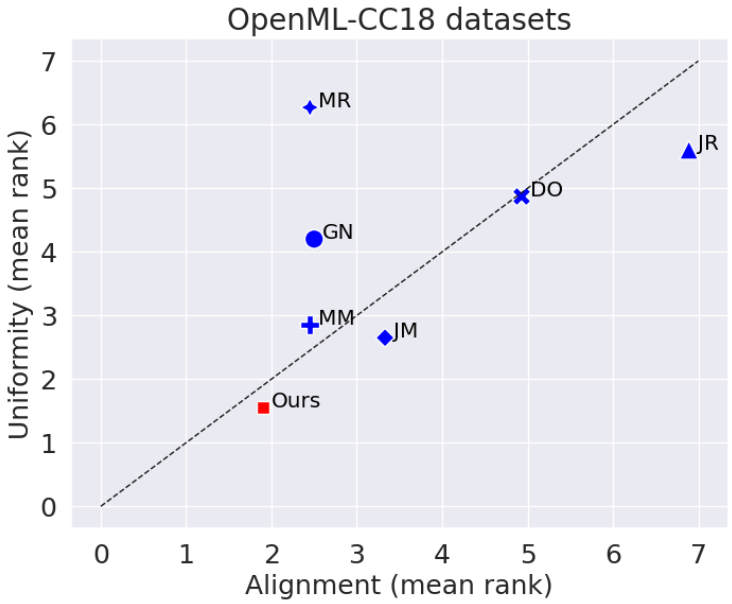
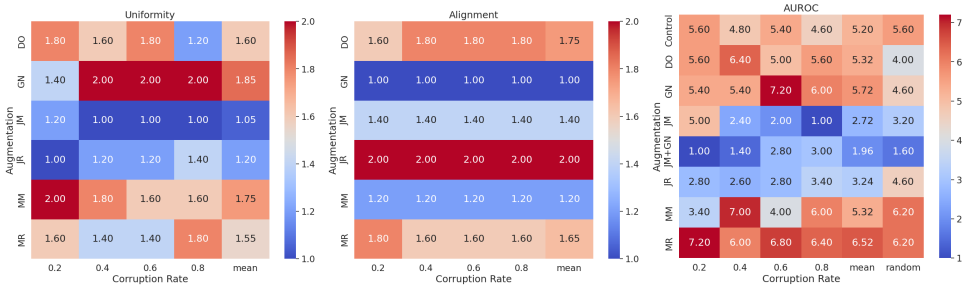


Figure 6: The augmentation pair selected by our framework leads to the most uniform and aligned feature spaces (bottom left).



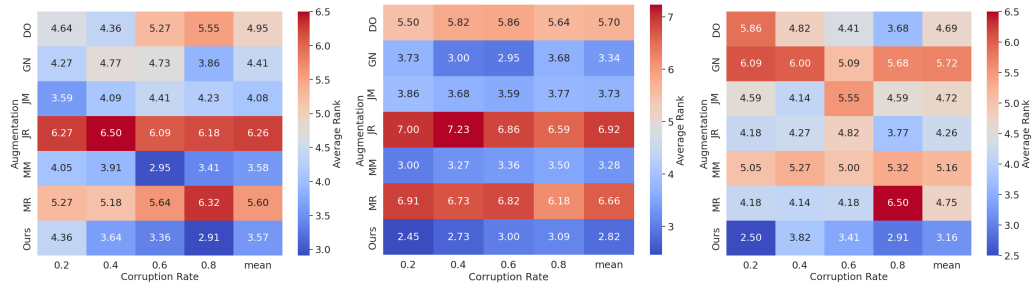
(a) Uniformity (b) Alignment (c) AUROC

Figure 7: Rank-based heat maps showing (a) uniformity, (b) alignment, and (c) AUROC on the TCGA datasets. The "random" column in (c) indicates randomly sampled corruption rates.

We first verify that balanced feature uniformity and alignment are positively correlated with downstream classification accuracy on the TCGA datasets. Figure 5 shows that this correlation also holds on the TCGA datasets, where the upper right quadrant of the scatter plot (i.e., the region showing higher alignment and uniformity) hosts the highest density of red colored points that indicate higher downstream AUROC. Each point in this plot shows the mean AUROC across 5 downstream tasks.

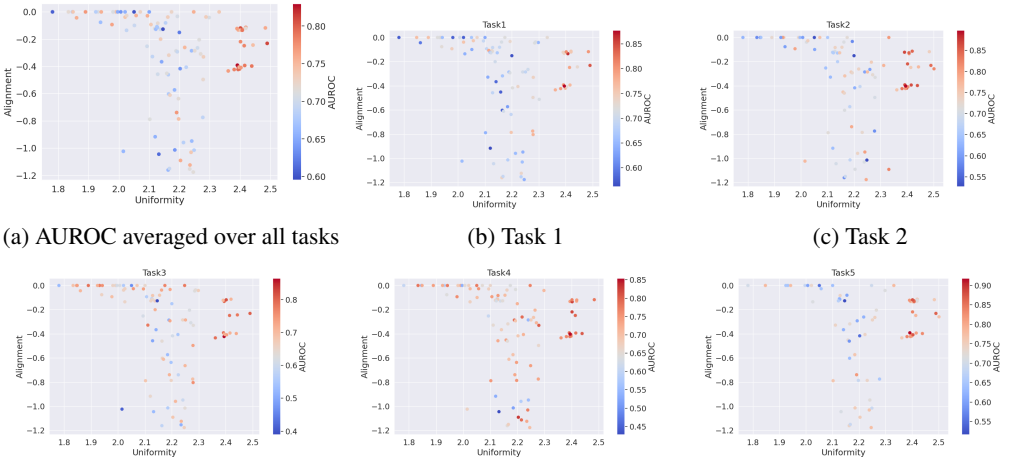
The multiple regression analysis again reveals a positive correlation between both alignment and uniformity with downstream performance, with the coefficients for uniformity and alignment being 0.091 and 0.059, respectively. We observe the same correlation for each single downstream task and include the results Figure 9.

Following our framework presented in Section 3, we begin by analyzing the uniformity and alignment of feature spaces generated by each of the augmentation techniques on the TCGA dataset. Figures 7a and 7b illustrate the mean uniformity and alignment ranking of each augmentation over all five model training runs for each corruption rate (shown in each column). We observe that the models pre-trained with joint Mixup (JM) exhibit the highest uniformity, while those utilizing Gaussian Noise (GN) demonstrate superior alignment. Consequently, we select these two augmentations as our combination and conduct pre-training from scratch using this combination (referred to as JM+GN). Figure 7c demonstrates that JM+GN outperforms all models trained using a single augmentation (except for



(a) Uniformity (b) Alignment (c) Accuracy

Figure 8: Uniformity, alignment and accuracy rank-based heat maps comparing our combination of augmentations with single augmentations on the OpenML-CC18 benchmarking datasets. Here, a FT-Transformer model was used as the encoder.



(a) AUROC averaged over all tasks (b) Task 1 (c) Task 2

(d) Task 3 (e) Task 4 (f) Task 5

Figure 9: Scatter plots of correlation with AUROC for each TCGA downstream task

joint Mixup) in terms of downstream AUROC across all corruption rates. Models pre-trained using JM+GN also significantly outperform models pre-trained using joint Mixup for lower corruption rates.

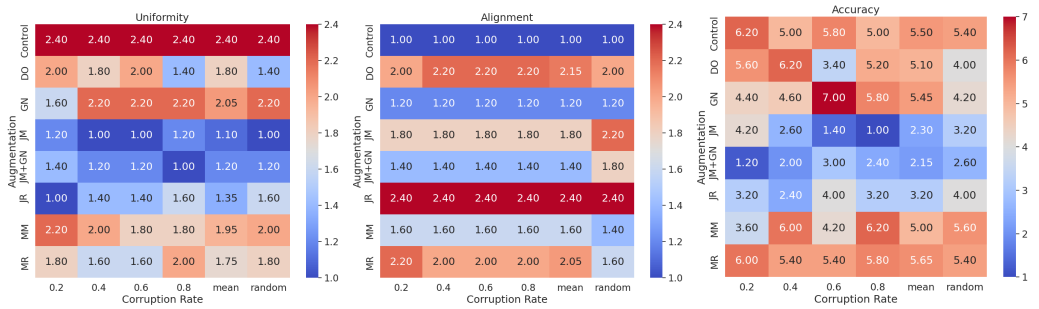
Moreover, our analysis reveals that JM+GN models achieve the highest downstream AUROC when averaged across corruption rates, as evidenced by the lowest mean rank value of 1.96 in the penultimate column (labeled "mean") of Figure 7c. Furthermore, Table 1 presents the mean AUROC for each downstream task, averaged over all corruption rates, for models pre-trained using each single augmentation technique as well as our combination. JM+GN models notably outperform other augmentation techniques, achieving the highest downstream AUROC scores on four out of the five tasks.

D.4.3 Experimental Results with Random Corruption Rates

Noticing that our combination consistently obtains similar levels of uniformity and alignment across various corruption rates, we proceed to investigate the impact of randomly sampling the corruption rate. The corruption rate serves as a crucial hyper-parameter for data augmentation, often tuned based on the model's performance on downstream task validation sets Bahri et al. (2022). Given the potential unavailability of labeled downstream data during model pre-training, randomly sampling the corruption rate emerges as a viable approach for eliminating this hyper-parameter.

Our experiments show (Figure 10) that our augmentation combination consistently outperforms single augmentation techniques when using randomly sampled corruption rates $c \in [0.2, 0.4, 0.6, 0.8]$. This superior performance is evident in the "random" column of Figure 7c. Table 5 complements

this observation by presenting the AUROC scores for each downstream task obtained with models pre-trained using single augmentations and our combination under randomly sampled corruption rates. Notably, our augmentation combination yields the highest downstream AUROC scores on three of the tasks and ranks second-best on the remaining two tasks.



(a) Uniformity (b) Alignment (c) Accuracy

Figure 10: Rank-based heat maps comparing (a) uniformity, (b) alignment, and (c) accuracy of our combination of augmentations with single augmentations on the TCGA datasets.

Augmentation	Task1	Task2	Task3	Task4	Task5
DO	**0.762 ± 0.048**	0.744 ± 0.021	0.601 ± 0.017	0.730 ± 0.064	0.791 ± 0.019
GN	0.708 ± 0.075	0.716 ± 0.060	0.676 ± 0.059	0.741 ± 0.079	0.756 ± 0.072
JM	0.759 ± 0.052	**0.778 ± 0.021**	0.657 ± 0.095	0.736 ± 0.033	0.727 ± 0.130
JR	0.743 ± 0.059	0.747 ± 0.082	0.648 ± 0.096	0.710 ± 0.033	0.774 ± 0.046
MM	0.750 ± 0.020	0.722 ± 0.050	0.611 ± 0.134	0.664 ± 0.063	0.699 ± 0.083
MR	0.705 ± 0.065	0.713 ± 0.080	0.632 ± 0.077	0.663 ± 0.126	0.794 ± 0.045
JM+GN (ours)	0.759 ± 0.037	**0.770 ± 0.080**	**0.697 ± 0.044**	**0.760 ± 0.043**	**0.810 ± 0.070**

Table 5: Mean and standard deviation of AUROC for each augmentation on TCGA downstream classification tasks with random corruption rates (averaged over 5 runs).

Imbalanced Downstream Task Dataset We conducted further experiments on an imbalanced version of the TCGA dataset, where the downstream data class ratio was 25:75. The results, presented in Figure 11, show that our approach also outperforms all other augmentation techniques in this setting.

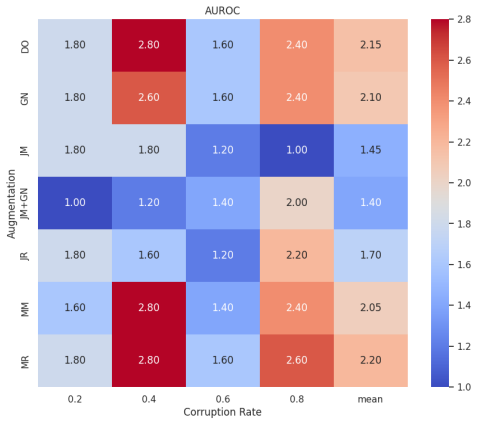


Figure 11: AUROC rank-based heat maps compare our combination of augmentations with single augmentations on the TCGA datasets for downstream tasks involving imbalanced classes.

Ablation Study Our ablation study compares our framework's selected pair with all other pairs. The chosen pair outperformed the others on average. Detailed results are in Figure 12.

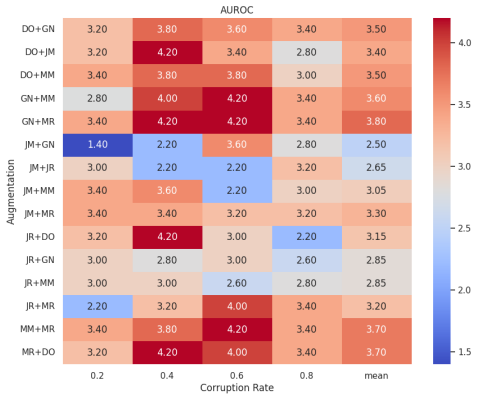


Figure 12: AUROC rank-based heat maps comparing all possible augmentation pairs on TCGA dataset. JM + GN, picked by our framework, outperforms all other pairs on average.

D.5 Limitations and Future Work

In this study, we focused on combining a pair of augmentations alternately (i.e., applying one at a time). This approach is more principled and aligns better with our framework, as it allows us to directly improve uniformity and alignment. An interesting avenue for future research would be to apply both augmentations simultaneously and to study their complex compound effects on feature space qualities. Additionally, extending our framework to systematically include more than two augmentations offers another promising direction for exploration. While we explored the interaction between augmentations and contrastive-learned feature space qualities in this work, we recognize that this is not the only factor affecting feature space quality. Future research could investigate the complex interplay between other factors, such as model architecture and methods, and how they influence feature space quality.