MUTEMBED: SELF-SUPERVISED LEARNING OF BIOLOGICAL LATENT EMBEDDINGS FROM CANCER MUTATIONAL PROFILES

Aakansha Narain[1], **Wu Jialun Andy**[1,2], **Hannan Wong**[1], **Vedant Sandhu**[1], **Jason J. Pitt**[1,2,3,4]
[1] Cancer Science Institute of Singapore, National University of Singapore
[2] Yong Loo Lin School of Medicine, National University of Singapore
[3] NUS Centre for Cancer Research, Yong Loo Lin School of Medicine, National University of Singapore
[4] NUS Artificial Intelligence Institute, National University of Singapore
{aakansha,jason.j.pitt}@nus.edu.sg

ABSTRACT

Cancer genomes possess diverse mutational patterns across multiple profiles, including single base substitutions (SBS), small insertions and deletions (ID), copy number variations (CN), and structural variants (SV). These profiles provide distinct, yet complementary perspectives to understanding a tumor's genomic landscape, which is essential for optimal patient care. Learning unified representations across this complex mutational landscape can reveal deeper insights into cancer biology, therapeutic interventions, and patient stratification. We present MutEmbed, a self-supervised framework that uses attention mechanisms to weigh and integrate information across mutational profiles, capturing their latent biological interdependencies. We use SBS, ID, CN, and SV calls for samples from the Pan-cancer Analysis of Whole Genomes (PCAWG) dataset (n = 2748). Using MutEmbed, we derive embeddings for each sample and demonstrate their biological relevance by analyzing cancer-type specific clustering patterns and enrichment patterns with DNA damage and repair pathway activities.

1 INTRODUCTION

Genomic instability (GI) is a well-established hallmark of cancer driven by the accumulation of genetic alterations, including single base substitutions (SBS), small insertions and deletions (ID), copy number variations (CN), and structural variants (SV). Patterns in these mutational profiles can inform underlying mechanisms of GI and reveal key insights into tumor evolution, DNA repair deficiencies, and potential therapeutic vulnerabilities. When analyzed collectively, the relationships within and between these profiles could reveal synergistic patterns that may be overlooked in single profile analyses (Everall et al., 2023). This approach provides a more comprehensive understanding of tumor heterogeneity, which can inform treatment strategies and ultimately improve patient outcomes. Computational methods, including machine learning and deep learning, are increasingly being studied to extract meaningful features from complex biological data and uncover insights into tumor biology and precision oncology. Anaya et al. (2023) showed an attention-based model that analyzed somatic mutations considering their local context using weakly supervised learning, and achieved superior performance in downstream tasks such as classification of tumor type and prediction of microsatellite status using the derived features. However, they only consider single domain contexts (for example, they show using SBS features) rather than integrating multiple sources of information.

We build on this idea by extending attention mechanisms across multiple mutational profiles with MutEmbed, a self-supervised framework that learns unified cancer sample representations. Our approach projects each mutation profile, or modality (SBS96, ID83, CN48, and SV32 - corresponding to the number of features for each mutation profile), into a shared embedding space where cross-modal attention enables information sharing between different mutation types. Through a reconstruction-based training objective, the model learns to preserve and integrate information across all profiles into a compressed latent representation, allowing biological relationships between

mutation types to naturally emerge without requiring explicit context cues or labels. These learned representations can capture meaningful patterns about the samples' tumor biology, enabling downstream applications such as cancer type prediction and pathway analysis.

Table 1: Comparison of classification performance (F1 score) across primary cancer types (top) and subtypes (bottom). Only cancer types with more than 50 samples were included.

	Bone	Breast	CNS	Colorectal	Esophagus	Kidney	Liver	Lung	Lymph	Myeloid	Ovary	Pancreas	Prostate	Skin	Stomach
MutEmbed	**0.47**	**0.64**	**0.79**	**0.74**	0.63	**0.87**	**0.93**	0.73	**0.92**	**0.70**	**0.71**	**0.67**	**0.80**	**0.97**	**0.30**
SBS96	0.27	0.24	0.58	0.57	**0.64**	0.67	0.91	0.73	0.78	0.69	0.59	0.39	0.50	0.87	0.25
ID83	0.14	0.34	0.62	0.47	0.42	0.85	0.84	**0.74**	0.54	0.34	0.44	0.40	0.61	0.75	0.19
CN48	0.02	0.14	0.29	0.17	0.29	0.42	0.43	0.18	0.46	0.39	0.37	0.39	0.54	0.08	0.04
SV32	0.08	0.12	0.31	0.11	0.28	0.24	0.21	0.11	0.26	0.35	0.32	0.13	0.48	0.20	0.09

	Adenocarcinoma	BNHL	CLL	Endocrine	HCC	MPN	Medullo	Melanoma	Pilocytic astrocytoma	RCC	SCC
MutEmbed	**0.84**	**0.82**	**0.82**	**0.67**	**0.93**	0.72	**0.76**	**0.95**	**0.64**	**0.91**	**0.62**
SBS96	0.63	0.74	0.55	0.44	0.92	**0.83**	0.36	0.88	0.39	0.73	0.50
ID83	0.56	0.40	0.50	0.34	0.80	0.34	0.55	0.72	0.37	0.88	0.43
CN48	0.31	0.25	0.52	0.50	0.45	0.54	0.37	0.17	0.55	0.26	0.23
SV32	0.35	0.18	0.23	0.15	0.33	0.33	0.12	0.22	0.59	0.29	0.26

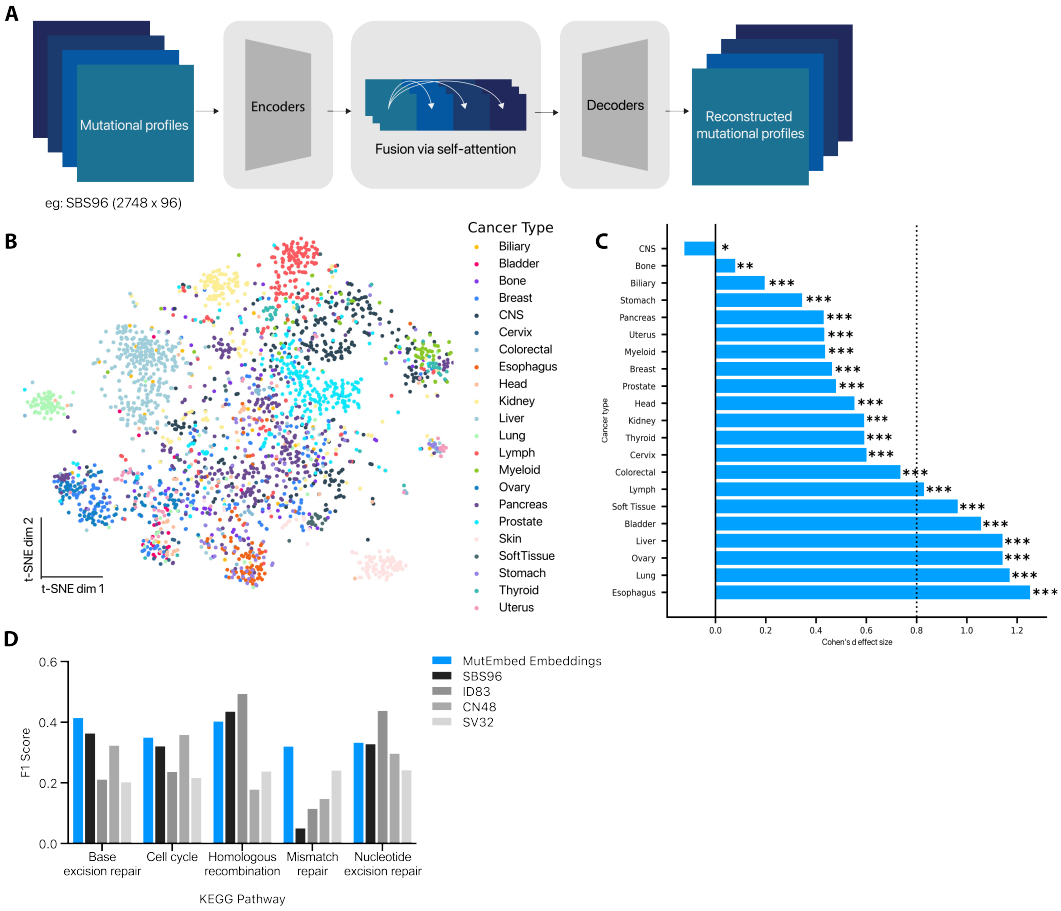

Figure 1: **A** MutEmbed architecture. **B** t-SNE visualisation of MutEmbed embeddings, labelled by cancer type. **C** Cohen's d effect size measuring similarity of samples within their cancer types in comparison to samples outside their cancer types. Statistical significance was assessed using an unpaired two-tailed t-test comparing within-cancer similarity scores versus between-cancer similarity scores. Significance levels are based on Bonferroni-adjusted p-values (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$), with an adjusted significance threshold of 0.002 to account for the 22 cancer types tested. **D** Classification performance on canonical DNA damage response and repair pathways.

2 METHODS

We define the MutEmbed network to integrate heterogeneous feature matrices $M_i \in \mathbb{R}^{n \times d_i}$, where n is the number of samples and d_i represents the dimensionality of the i-th feature space. The model processes k different feature matrices through a projection layer $P : \mathbb{R}^{d_i} \rightarrow \mathbb{R}^e$ that maps each input space to a common embedding dimension e. The projected features are stacked column-wise to form a sequence of feature representations for each sample $X = [x_1; x_2; ...; x_k] \in \mathbb{R}^{n \times k \times e}$.

We omit positional encodings, as the order of feature matrices is not relevant across samples and represents distinct mutational profiles rather than sequential data. The stacked features are processed through l layers of attention mechanisms. The final representation is compressed into a latent space $z \in \mathbb{R}^{64}$ through mean pooling across the feature dimension followed by linear projection. Decoders reconstruct the original feature spaces through separate linear transformations. The model is trained to minimize the mean squared reconstruction error $\mathcal{L} = \sum \|M_i - M_i'\|^2$ across all feature spaces.

3 RESULTS

Latent Space Organization of Mutational Profiles After training MutEmbed, we analyzed the 64-dimensional latent representations using t-SNE visualization (Figure 1 A), which revealed distinct clustering patterns by cancer type. We validated clustering by using Cohen's d to measure the effect size on intra- and inter-cancer cosine similarities and running t-tests with Bonferroni correction to assess significance (Figure 1 B). Strong clustering (Cohen d's > 0.8) – especially in skin, lung, liver, kidney, and esophageal cancers that have distinct mutational signatures (for example, known exogenous exposures like UV light for skin cancer and tobacco smoking for lung cancer map to specific and distinct signatures (Alexandrov et al., 2013)) – suggests our embeddings capture cancer-type-specific mutation patterns. Conversely, weaker clustering in cancer types such as CNS, bone, and stomach cancers may reflect similarities in tumor heterogeneity, aberrant mutational processes, or active biological pathways across cancer types.

Cancer Classification Performance We used the embeddings for multivariate prediction of primary cancer type and subtype, which is particularly useful when determining the tumor of origin for metastatic cancers is difficult (Pavlidis & Pentheroudakis, 2012). The test F1 scores (Table 1) for strongly clustered cancer types – skin (0.97), liver (0.93), and lymph (0.92) – corroborate these groupings. Similarly, high F1 scores (> 0.8) for several subtypes indicate that the embeddings capture not only tissue of origin but also underlying molecular mechanisms. For instance, the strong performance in distinguishing Chronic lymphocytic leukemia (CLL) and B-cell non-Hodgkin lymphoma (BNHL), both lymphoid cancers, suggests that despite their shared lineage, they have distinct molecular evolution paths, which the embeddings successfully differentiate. Notably, our approach significantly outperformed classifiers trained on individual mutation profiles alone, highlighting the advantage of integrating cross-modality data. Furthermore, while the results in Table 1 are after using weighted loss to correct class imbalance, MutEmbed shows equally superior performance without any correction (Appendix A.2), indicating that the embeddings alone are sufficiently distinct to separate classes, unlike other features.

Biological Pathway Analysis To assess whether the learned embeddings encode information related to pathway-level biology, we attempted to predict dominant pathway activity states. As a proxy for pathway activity, we applied Gene Set Variation Analysis (GSVA) to RNA-seq expression profiles for a subset of the PCAWG samples to compute enrichment scores for a selected set of canonical DNA damage response and repair pathways from the KEGG database on a per-sample basis. GSVA generates normalized, sample-specific enrichment scores that reflect the relative expression of pathway genes. For this preliminary analysis, we simplified the pathway labelling by selecting, for each sample, the pathway with the largest absolute GSVA enrichment score. This captures the pathway showing the strongest deviation in expression (either up or down regulation) without distinguishing directionality. Figure 1 D shows the performance from the individual mutation profiles and the learnt embeddings. Unlike with previous downstream tasks, the MutEmbed embeddings do not perform as well compared to the individual profiles, only having higher predictive power for Base Excision Repair and Mismatch repair. This reflects the complexity of such an analysis from a technical perspective, as the relationship between mutation patterns and pathway activity is not biologically direct. Additionally, this approach would likely benefit from incorporat-

ing the directionality of pathway regulation, as well as more sophisticated labelling strategies, given that multiple pathways may be relevant for a single sample. Moreover, because the embeddings are trained to learn a general representation of mutation profiles, they may not explicitly prioritize pathway-specific features, but instead capture broader, context-dependent patterns - as was also observed in the cancer type classification task.

4 CONCLUSION

We showed that MutEmbed effectively integrates diverse mutational profiles into a shared representation and captures meaningful cancer-type-specific patterns without explicit biological labels. These embeddings may be used for relevant downstream prediction tasks, and could also be analyzed further from a biological perspective to understand tumor heterogeneity better. In future work, we aim to refine our embeddings by integrating additional genomic and transcriptomic modalities, incorporating pathway-aware training objectives, and applying more sophisticated phenotype-labeling strategies. We also plan to validate the utility of these embeddings across a wider range of pathway analyses, cancer phenotypes, and clinical outcomes.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2025 Tiny Papers Track.

REFERENCES

Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

Jordan Anaya, John-William Sidhom, Faisal Mahmood, and Alexander S. Baras. Multiple-instance learning of somatic mutations for the classification of tumour type and the prediction of microsatellite status. *Nature Biomedical Engineering*, 8:57–67, 2023.

Erik N. Bergstrom, Mi Ni Huang, Ludmil B. Alexandrov, et al. Sigprofilermatrixgenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics*, 20(1), 2019.

Peter J. Campbell, Gad Getz, and The PCAWG Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, 2020.

Andrew Everall, Avraam Tapinos, David C. Wedge, et al. Comprehensive repertoire of the chromosomal alteration and mutational signatures across 16 cancer types from 10,983 cancer patients. *medRxiv*, 2023.

Mary J. Goldman, Brian Craft, David Haussler, et al. Visualizing and interpreting cancer genomics data via the xena platform. *Nature Biotechnology*, 38(6):675–678, 2020.

Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC Bioinformatics*, 14(1):7, January 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-7. URL https://doi.org/10.1186/1471-2105-14-7.

Arthur Liberzon, Aravind Subramanian, Jill P. Mesirov, et al. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

Nicholas Pavlidis and George Pentheroudakis. Cancer of unknown primary site. *The Lancet*, 379: 1428–1435, 2012.

A APPENDIX

A.1 METHODS - ADDITIONAL DETAILS

A.1.1 DATA ACQUISITION AND PRE-PROCESSING

We used 2748 samples from the Pan-cancer Analysis of Whole Genomes (PCAWG) dataset (Campbell et al., 2020), which is available via the Legacy International Cancer Genome Consortium (ICGC) 25K server. For each whole genome sequencing (WGS) sample, we downloaded SBS96 (from single base substitutions) and ID83 (from small insertions and deletions) matrices from the ICGC portal, and used SigProfilerMatrixGenerator's *CNVMatrixGenerator* and *SVMatrixGenerator* scripts to generate the CN48 and SV32 mutational profiles (Bergstrom et al., 2019). All four mutational profiles for each WGS sample were independently frequency standardized and normalized prior to training MutEmbed.

A.1.2 MUTEMBED MODEL ARCHITECTURE AND IMPLEMENTATION

MutEmbed was trained on the complete set of PCAWG samples, optimizing embeddings by minimizing the reconstruction loss for each mutational profile. Within the model, for each WGS sample, we calculate cross-profile attention to dynamically share information across modalities and learn correlation patterns.

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V \tag{1}$$

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{E}}\right) \tag{2}$$

where Z is the stacked projected profiles, $W_Q, W_K, W_V \in \mathbb{R}^{E\times E}$ are learnable projection matrices, and A represents the attention scores.

After L attention layers, the latent representation is computed as:

$$H = \text{ReLU}(W_L \cdot \text{mean}(Z', \text{dim} = 1)) \tag{3}$$

where $W_L \in \mathbb{R}^{E\times 64}$ projects to a lower-dimensional latent space.

We used Adam optimizer with a learning rate of 0.0006, batch size of 32 and 500 epochs. For this experiment, we set the hidden dimension to be 265, latent dimension to be 64, and used only one layer of attention.

A.1.3 Classification Tasks Implementation

For the multi-class classification tasks, we trained a simple MLP consisting of 2 linear layers (hidden dimension = 128) and a ReLU layer. Only cancer types with more than 50 samples across the dataset were used. Weighted cross entropy loss was used to correct for class imbalance. We used a 60/20/20 train/val/test split and ran 10 trials for each experiment (MutEmbed embeddings, SBS96, ID83, CN48 and SV32) across the same initially randomized seeds over 50 epochs with early stopping. We used a batch size of 32, learning rate of 0.001 with Adam optimizer.

A.1.4 Gene Set Variation Analysis Details

We obtained normalized RNAseq gene expression data (35608 genes) for a subset (n = 1214) of the PCAWG samples from the University of California Santa Cruz XENA portal (Goldman et al., 2020). We filtered genes based on their status in the Consensus CDS project, retaining only those that were public or updated under active review. This resulted in a final set of 17640 genes. We applied Gene Set Variation Analysis (GSVA) (Hänzelmann et al., 2013) to our gene expression matrix to derive sample-level pathway enrichment scores for selected KEGG pathways. Using the GSVA R package with default parameters, gene expression measurements (rows: genes, columns: samples) were non-parametrically transformed to calculate an enrichment score for each pathway in each sample. These GSVA scores were then used for downstream clustering and comparative analyses. To identify relevant pathways, we queried the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011) using the following terms: KEGG AND ((DNA AND Damage) OR (DNA AND Repair) OR (Homologous AND Recombination)). From the results, we selected five key pathways: cell cycle, homologous recombination, mismatch repair, base excision repair, and nucleotide excision repair. Associations obtained for each dimension were filtered based on a q-value threshold (< 0.05) for the normalized enrichment score (NES).

A.2 Results - Additional Details

Table 2: Comparison of classification performance (F1 score) without weighted loss across primary cancer types (top) and subtypes (bottom).

	Bone	Breast	CNS	Colorectal	Esophagus	Kidney	Liver	Lung	Lymph	Myeloid	Ovary	Pancreas	Prostate	Skin	Stomach
MutEmbed	**0.41**	**0.70**	**0.80**	**0.78**	**0.64**	**0.89**	**0.94**	**0.77**	**0.93**	**0.70**	**0.71**	**0.75**	**0.80**	**0.96**	**0.32**
SBS96	0.00	0.46	0.67	0.51	0.66	0.60	0.90	0.73	0.80	0.55	0.31	0.45	0.52	0.86	0.10
ID83	0.01	0.40	0.63	0.52	0.29	0.83	0.83	0.69	0.60	0.24	0.41	0.47	0.61	0.76	0.03
CN48	0.00	0.27	0.51	0.00	0.12	0.29	0.45	0.02	0.46	0.42	0.33	0.43	0.58	0.00	0.00
SV32	0.00	0.20	0.36	0.00	0.17	0.17	0.36	0.00	0.22	0.24	0.32	0.26	0.48	0.08	0.00

	Adenocarcinoma	BNHL	CLL	Endocrine	HCC	MPN	Medullo	Melanoma	Pilocytic astrocytoma	RCC	SCC
MutEmbed	**0.90**	**0.89**	**0.84**	**0.74**	**0.92**	0.70	**0.81**	**0.99**	**0.58**	**0.93**	**0.68**
SBS96	0.80	0.61	0.22	0.10	0.87	**0.84**	0.17	0.90	0.02	0.72	0.25
ID83	0.79	0.03	0.43	0.13	0.78	0.15	0.68	0.75	0.37	0.91	0.21
CN48	0.71	0.00	0.21	0.64	0.36	0.19	0.26	0.00	0.46	0.12	0.00
SV32	0.69	0.00	0.25	0.11	0.20	0.09	0.00	0.61	0.19	0.00	0.00