HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2020 January 01.

Published in final edited form as:

Pac Symp Biocomput. 2020 ; 25: 379–390.

Learning a Latent Space of Highly Multidimensional Cancer Data

Benjamin Kompa[†],

Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

Beau Coker

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

Abstract

We introduce a Unified Disentanglement Network (UFDN) trained on The Cancer Genome Atlas (TCGA), which we refer to as UFDN-TCGA. We demonstrate that UFDN-TCGA learns a biologically relevant, low-dimensional latent space of high-dimensional gene expression data by applying our network to two classification tasks of cancer status and cancer type. UFDN-TCGA performs comparably to random forest methods. The UFDN allows for continuous, partial interpolation between distinct cancer types. Furthermore, we perform an analysis of differentially expressed genes between skin cutaneous melanoma (SKCM) samples and the same samples interpolated into glioblastoma (GBM). We demonstrate that our interpolations consist of relevant metagenes that recapitulate known glioblastoma mechanisms.

Keywords

machine learning; RNAseq; image translation; disentangled latent spaces; multidimensional data

1. Introduction

Deep learning is being applied to many difficult problems in genomics and medicine such as understanding cancer prognosis. Chaudhary et al. were able to robustly predict survival in liver cancer.[1] Cruz-Roa et al. leveraged deep learning to quantify the extent of breast cancer tumors in imaging data.[2] Other groups have trained networks to identify metastatic breast cancer and lymph node metastasis.[3]

There are significant questions remaining in oncology about the relationships between different cancer types. For instance, while there is an association between melanoma, a type of skin cancer, and glioblastoma, a type of brain cancer, little is known about the molecular underpinnings of this relationship.[4,5] Nevertheless, there is little work in machine learning being done on what changes are occurring at a gene expression level during metastasis.

Recently, deep generative models such as variational auto encoders (VAEs) and generative adversarial networks (GANs) have made large advances in image, audio, and text

[†] kompa@fas.harvard.edu.

generation.[6–8] VAEs and GANs learn generative distributions on lower-dimensional encodings of input data.[9] VAEs have found genomic applications. Rampasek et al. applied VAEs to learn drug responses based on gene expression data.[10] Way et al. trained a VAE called Tybalt to encode The Cancer Genome Atlas (TCGA).[9] Huang et al. have developed a theory of cancer development as a progression along a low dimensional space, justifying exploration of cancer metastasis using machine learning algorithms that learn low dimensional representations.[11]

A new VAE-GAN hybrid architecture known as the Unified Feature Disentanglement Network (UFDN) learns fundamental features that distinguish input domains.[12] For multiple input data types, such as photographs, sketches, and watercolor paintings, the UFDN learns an VAE encoding of the data domains and trains a discriminator in the latent space to discriminate between domain types. Then, the UFDN can subsequently encode data from one domain and decode the data into a different domain.[12] An additional GAN distinguishes between real/fake images in the pixel space to promote high quality decodings.[12]

The primary goal of this work is to utilize the UFDN architecture to learn a disentangled latent space of cancer gene expression data, which allows for interpolation between cancer types.

2. Overview of UFDN-TCGA

In this work, we apply this new UFDN architecture to TCGA RNA-Seq data and learn a latent space embedding that allows us to convert between different cancer types given gene expression data. Given a sample's gene expression levels in one type of cancer, we can predict gene expression levels as if that cancer sample were of another type. This represents a generative, personalized model of metastasis. We can sample points in our latent space encoding and decode them into any new cancer domain.

Additionally, we can partially interpolate between cancer domains. UFDN decoding is not strictly binary—input data can be decoded into a mix of output domains. We investigate *partial interpolations* of one cancer type into another, mimicking the progressive nature of metastasis.

We analyze the performance of our TCGA-trained UFDN on two tasks: predicting whether a sample is from cancerous or normal tissue and predicting which cancer sub-type a sample consists of. Additionally, we investigate partial interpolations from skin cutaneous melanoma (SKCM) TCGA samples to glioblastoma (GBM) by looking at differential expression of genes. We compute metagenes that summarize gene expression changes using integrative non-negative matrix factorization. Finally, we analyze Gene Ontology (GO) term enrichment in highly activated metagenes for each interpolated dataset.

2.1. UFDN Architecture

Liu et al. develop a UFDN as a combination of an encoder E, a generator G, and two discriminators: D_V in the latent space and D_X in the pixel space.[12] In our application, pixel space is replaced by "gene expression space." E takes input data and encodes it in a latent

space. In our UFDN, we encode gene expression using fully connected networks. D_v learns to discriminate between domains, or cancer types. Then, generator G uses a latent space encoding z and a domain vector d_v to produce gene expression data in domain v.[12] Our UFDN uses $d_v \in \mathbb{R}^{33}$ since there are 33 cancer types in TCGA.

We define a *partial interpolation* with parameter $p \in [0,1]$ of an input of domain c to domain $\hat{c}$ to be the decoding of the input into into a composition of domains c and $\hat{c}$, with weight p given to domain $\hat{c}$. That is, the domain vector of the partial interpolation has components $d_{v_{\hat{c}}} = p$, $d_{v_c} = 1 - p$, and remaining components zero. For instance, a 0.25-GBM interpolation means an input has been decoded with $d_{vGBM} = 0.25$ and original domain entry is 0.75.

In the input space, D_x learns to distinguish between samples that have been decoded to their original domain c or a new domain $\hat{c}$.[12] The network is trained by iterative stochastic gradient updates to E, D_v, and D_x. For a more detailed exposition of the architecture of and gradient updates for training the UFDN, please see Section 3 of Liu et al. 2018.[12]

The encoder E and generator G are single layer networks, each with 500 hidden units, that learn a 100 dimensional latent space. The feature space discriminator D_v is a single layer network with 64 hidden units and the pixel space discriminator D_x is a two layer network with 500 and 100 hidden units. All networks are fully connected with leaky ReLU activation functions. We use 50,000 iterations of Adam updates with a learning rate of 10^{-4}.

3. Methods

3.1. Data Preprocessing

The data consisted of 10,433 samples of RNA-Seq gene expression levels across 33 cancer types for 20,501 genes from TCGA obtained via the R Package curatedTCGA.[13,14] For the purpose of this work, we only considered the RSEM[15] normalized expression levels. We divided the data 70%, 20%, and 10% to train, test, and holdout datasets, respectively.

Way et al. demonstrated that preprocessing gene expression levels by scaling gene-wise expression levels (across all samples) to between 0 and 1 yields a trainable latent space.[9] We adapted this procedure by first clipping expression levels to fall within 3 standard deviations from the mean of gene-wise expression levels followed by the same min-max normalization of Way et al..[9]

3.2. Classification Tasks

We assessed two classification tasks using the UFDN. The first *Cancer Status* task was classifying a sample as tumor or normal. The second *Domain* task was predicting cancer domain, one of 33 types of cancer in the TCGA.

We compared three different ways of using UFDN-TCGA on these tasks:

- *UFDN-MSE*: Classify a sample's type by encoding the sample and decoding it into all 33 domains, predicting the type of the domain with lowest reconstruction error as defined by mean square error (MSE).
- *Unsupervised UFDN*: Inspired by the unsupervised domain adaptation experiments from Liu et al.,[12] this algorithm predicts cancer status by encoding a sample into the latent space, then decoding it into the mesothelioma domain, regardless of input domain. We trained a random forest classifier to predict cancer status on mesothelioma training data, then use the prediction of this classifier to predict cancer status in the original input domain. The motivation for this approach is that the classifier trained on mesothelioma data is strong but the test data of interest is of a different cancer type.
- *Semi-supervised UFDN*: A hybrid of the two above algorithms used to predict cancer status and type. First, predict cancer type using *UFDN-MSE*. Then, predict cancer status using a random forest classifier trained on that specific type's status data.

3.3. Interpolation Analysis

We encoded 95 samples of SKCM (skin cutaneous melanoma) from our test set partition of the TCGA into our latent space using our trained UFDN. Then, we interpolated the samples into glioblastoma (GBM) at four different fractions of interpolation: 25%, 50%, and 75%, and 100%. The 100% interpolation represents a prediction of gene expression levels of the SKCM samples as GBM samples.

In order to analyze how gene expression changed between SKCM samples and these samples as GBM, we performed a differential expression analysis using **edgeR**.[16,17] This is an R package that uses a negative binomial distribution model to analyze significant gene expression changes between two groups.[16,17] Although normally **edgeR** works with raw read counts, more recently the package creator has stated that RSEM normalized reads are also suitable for use with **edgeR**.[18]

We applied the inverse transformation of our min-max normalization to our four interpolated datasets since our UFDN decodes gene expression levels to within the range of [0,1]. Then we used **edgeR** to find differentially expressed genes between SKCM samples and 100% GBM interpolated samples. A p-value threshold for differential expression was set at $p = .05/20501 = 2.438 * 10^{-6}$ to control for false discovery.

Analyzing every single gene that significantly changed between SKCM and GBM would be a computational challenge, so we used integrative non-negative matrix factorization (IntNMF) to learn metagenes that summarized gene expression changes.[19] IntNMF learns a reduced dimensionality representation across multiple datasets.[19] IntNMF learns a shared basis matrix $W \in \mathbb{R}^{p \times k}$ and where p is the number of features (here, the differentially expressed genes) and k is the number of metagenes, $k << p$. Each dataset D_j is described by a learned matrix $H_j \in \mathbb{R}^{k \times n}$ where n is the number of samples in the dataset.[19] Each row of H_j represents the linear combination of metagenes of W that combine to reconstruct the

original sample in D_j.[19] We chose $k = 60$ based on an analysis of the reconstruction error $\sum_j \|D_j - WH_j\|_F$, where F is the Frobenius norm. We learned W and H_j for each dataset using the R package IntNMF.[19]

Every element g of column $W^{(i)}$ is non-negative and represents the contribution of gene g to the i-th metagene.[19] Each element s of the n-th row of H_j represents the contribution of metagene s to the n-th sample of the j-th dataset. We can analyze how these metagenes change over the different interpolation datasets in order to understand how gene expression is changing.[19]

Finally, to understand the broad composition of the metagenes discovered by IntNMF, we used Gene Ontology (GO) enrichment analysis. GO terms are an ontology of three categories: biological processes, molecular function, and cellular component. They link together information about the functions and relationships of genes and proteins. topGO is an R package that analyzes if GO terms, which have been mapped to genes, show up more often than expected in a set of genes and associated scores for each gene.[20]

We used test similar to the Kolmogorov-Smirnov test known as Gene Score Enrichment Analysis that calculates p-values of enrichment based on a score for each gene.[20] We tested each metagene derived from IntNMF with the score for gene g as $W_g^{(i)}$. [20] By looking at the top scoring GO terms for each metagene, we understand what sort of genes are changing as we interpolate between cancer types.[20]

4. Results

4.1. UFDN Training and Performance

First, we validated that our UFDN learned a disentangled latent space representation of TCGA RNA-Seq data. Liu et al. define a latent space as *disentangled* if domain information is uncoupled from representation in the latent space.[12] Figure 3 shows the TCGA data and latent space encodings projected into UMAP space.[21] UMAP learns a Riemann manifold representation of the data.[21] We observed distinct clusters by cancer types for both the original data, but less distinct clusters for the encodings. This represents a disentangling of domain information and latent space representation and allows for interpolation between domains.

Next, we estimated the ability of our UFDN to take data from a source domain (original cancer type) and interpolate these data into a target domain (new cancer type). We considered the fraction of the k nearest neighbors, in the training data, of the interpolated samples that were in the target domain as a measure of success. These decoding rates are shown in Figure 4. There were certain cancers that the UFDN was able to more robustly interpolate into. These included glioblastoma, acute myloid leukemia, mesothelioma, and prostate adenocarcinoma, among others. Difficult cancers to interpolate into were sarcomas, which are a heterogeneous subcategory of soft tissue cancers, and cervical squamous cell carcinoma.

Finally, we analyzed our UFDN's performance on two classification tasks: *Cancer Status* and *Domain* prediction. Table 1 reports the performances of our three UFDN classification algorithms as compared to a random forest baseline. The random forests had a maximum depth of 15 and were composed of 100 trees. The *Semi-supervised UFDN* algorithm was able to match the performance of random forests on the cancer status task and was comparable on the cancer type task. Other UFDN algorithms were less successful compared to the baseline.

4.2. Gene Expression Changes

After interpolating 95 samples of SKCM from the test set into GBM, we analyzed which genes had significant changes in expression between the SKCM and 1.0-GBM samples. Using **edgeR**, we looked for genes that had differential expression that exceeded a significance threshold of $p = 2.43 * 10^{-6}$, which accounts for the Bonferroni correction. There were 10,557 genes that exceeded this threshold.

For the 10,557 differential expressed genes, we learned a shared basis W using IntNMF. By varying the rank of that basis, we were able to decrease the reconstruction error across datasets SKCM, 0.25-GBM, 0.5-GBM, 0.75-GBM, and 1.0-GBM. We chose $k = 60$ for subsequent analysis based on the inflection point of this reconstruction curve (see Supplementary Materials). Hutchins et al. suggest that this is an optimal way to select k for NMF.[22]

Finally, we visualized the rows of H_j for each dataset in {SKCM, 0.25-GBM, 0.50-GBM, 0.75-GBM, 1.00-GBM}. The columns of each heatmap in Figure 5 represent the relative activation of the respective metagene. As interpolation towards GBM increases, distinct metagenes increase their responsibility for reconstructing H_j. In SKCM, metagene 36 has the most representation in the data. For 0.25-GBM, 0.50-GBM, and 0.75-GBM, metagenes 15, 32, and 1 had the most representation in the data, respectively.

In the 1.00-GBM heatmap (Figure 5 E), we saw the increased activation of metagene 23. When we took 33 samples of TCGA GBM data from the test set and learned the matrix *HGBM* that minimized reconstruction error $\|D_{GBM} - W|H_{GBM}\|_F$ for the same, fixed, W learned previously by IntNMF, we observed the same metagene 23 dominating (Figure 5 F).

We proceeded to analyze the dominant metagene for every dataset H_j for GO term enrichment. In the interest of space, we only report the top 15 most enriched GO terms for metagene 23 based on p-value. Table 2 reports the GO term as well as p-value for each term.

Additional analysis was performed after controlling for false positive in **edgeR** results using the Wilcoxon signed-rank test. See the Supplementary Materials for this analysis.

5. Discussion

Our UFDN was able to learn a biologically relevant latent space encoding of TCGA data. Classification task results in Table 1 indicate that our UFDN was able to compete with random forests that were trained on all 20,501 gene expression features. This indicates our

algorithm was able to learn an efficient, useful embedding of gene expression data. Some UFDN classification methods likely performed worse than random forest methods due to a reduction in dimensionality. *UFDN-MSE*, semi-supervised, and unsupervised classification methods all encode gene expression from the 20,501 TCGA space into a 100 dimensional latent space. This encoding decreases the amount of information available to downstream classifiers (even after decoding), resulting in a decrease in performance. The goal of this analysis was not to learn a state-of-the-art classifier for cancer status/domain, but rather validate that our UFDN retains information about cancer status/domain.

Figure 3 demonstrates that we learned an encoding that disentangled domain information from latent space representation. Additionally, our UFDN could robustly interpolate into many cancer domains. Figure 4 demonstrates that interpolated gene expression levels are comparable to real gene expression levels. Since interpolated gene expression levels are consistently near real training samples of the target domain according to mean square error, we are accurately recapitulating gene expression levels.

We observed 10,557 differentially expressed genes between SKCM and 1.0-GBM interpolated samples. **edgeR** was mainly employed to reduce the number of genes analyzed with IntNMF. This reduction in dimensionality allowed us to make IntNMF computationally tractable. A further reduction in dimensionality was done by filtering with the Wilcoxon ranked-sign test for differentially express genes. 8,878 genes remained after Wilcoxon filtering. Alternative gene filtering methods could be considered in future works. The lower number of genes considered in IntNMF, the faster the learning of the shared basis W and dataset specific H_j. Analysis of the reconstruction error from IntNMF informed our choice of 60 metagenes (see Supplementary Materials). In Figure 5, we investigated how the relative weighting of each metagene change for each partial interpolation. We observed unique metagenes increasing in importance for each partial interpolation. This is an approximation of how gene expression profiles change during metastasis.

When we learned H_{GBM}, the representation of TCGA GBM samples with respect to the basis W, something remarkable happened. Note that W was not informed by the TCGA dataset GBM at all. W was simply the shared basis trained by IntNMF on interpolation datasets SKCM (equivalently, 0.00-GBM), 0.25-GBM, 0.5-GBM, 0.75-GBM, and 1.0-GBM. Yet when $H_{1.0-GBM}$ and H_{GBM} were compared side by side in Figure 5 E&F, their metagene activation profiles were dominated by the same metagene 23. Therefore, our interpolation from SKCM to GBM successfully recapitulated observed gene expression activity.

One advantage of the UFDN interpolations as compared to standard differential expression techniques is that we can look at which metagenes are activated for these partial interpolations. Metagene 23 would likely be recovered if you learn a new basis on just differentially expressed genes between TCGA-SKCM and TCGA-GBM. However, the UFDN interpolations allow us to examine what metagenes are activating as cells are transformed from one cell type to another *in silico*. Clearly, having gene expression data from cells undergoing metastasis would be ideal to understand the transition from SKCM to

GBM. The UFDN interpolations allow us to make hypotheses about which groups of genes are activating during metastasis.

Furthermore, when we explored several of the GO terms identified by a GO term enrichment analysis, metagene 23 was enriched for terms related to glioblastoma. GO:0008376 represents a glycoprotein with a known association to glioblastoma.[23,24] GO:0004126 refers to cytidine deaminase activity. Cytidine deaminase gene therapy has been identified as a potential treatment for glioblastoma.[25,26] GO:0048020 and GO:0008009 are associated with chemokines, which are implicated in glioblastoma development.[27,28] Our metagenes learned glioblastoma-specific genes and our UFDN interpolated skin cancer samples to glioblastoma. Further analysis of the metagenes activated during interpolations 0.25-GBM, 0.50-GBM, and 0.75-GBM could provide starting points for the investigation of the metastasis pathway from SKCM to GBM. This could help explain the association between melanoma and glioblastoma that is currently not understood.[4,5]

One factor that remains unexplored in this work is tumor purity. It would be interesting to see how different levels of tumor purity cluster in the UFDN latent space. Would all samples from one domain cluster together regardless of purity? How would they stratify within said cluster? These questions could be answered by using copy number information available in the TCGA and running FACETS to quantify purity.[29] We could also consider making synthetic datasets and training a new UFDN.

Ultimately, a significant limitation of this method is analyzing out of domain samples. This UFDN has been trained on specific cancer types and gene sets. When adding additional data sources, it is necessary to retrain the network. Additionally, the UFDN model currently requires a uniform number of input features across all samples. If some samples have incomplete feature sets, they likely cannot be used for training or evaluation.

6. Conclusion

Our UFDN learned a biologically relevant latent space that facilitated meaningful interpolations between cancer domains. Our latent space can be used to generate more examples of transitions between cancers types. Our interpolations from SKCM to GBM have feasible biological interpretations and suggest possible gene expression changes during the transition from melanoma to glioblastoma.

6.1. Code and Supplementary Materials

All of our code and Supplementary Materials is available at https://github.com/bkompa/UFDN-TCGA.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

BK was supported by NIH T32HG002295.

References

1. Chaudhary K, Poirion OB, Lu L and Garmire LX, Deep Learning–Based Multi-Omics integration robustly predicts survival in liver cancer, Clin. Cancer Res. 24, 1248 (3 2018). [PubMed: 28982688]
2. Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih NNC, Tomaszewski J, González FA and Madabhushi A, Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent, Sci. Rep. 7, p. 46450 (4 2017). [PubMed: 28418027]
3. Wang D, Khosla A, Gargeya R, Irshad H and Beck AH, Deep learning for identifying metastatic breast cancer (6 2016).
4. Desai AS and Grossman SA, Association of melanoma with glioblastoma multiforme, J. Clin. Orthod. 26, 2082 (5 2008).
5. Scarbrough PM, Akushevich I, Wrensch M and Il'yasova D, Exploring the association between melanoma and glioma risks, Ann. Epidemiol. 24, 469 (6 2014). [PubMed: 24703682]
6. Hsu C-C, Hwang H-T, Wu Y-C, Tsao Y and Wang H-M, Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks (4 2017).
7. Larsen ABL, Sønderby SK, Larochelle H and Winther O, Autoencoding beyond pixelsusing a learned similarity metric (12 2015).
8. Pu Y, Gan Z, Henao R, Yuan X, Li C, Stevens A and Carin L, Variational autoencoder for deep learning of images, labels and captions, in Advances in Neural Information Processing Systems 29, eds. Lee DD, Sugiyama M, Luxburg UV, Guyon I and Garnett R (Curran Associates, Inc., 2016) pp. 2352–2360.
9. Way GP and Greene CS, Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders, Pac. Symp. Biocomput. 23, 80 (2018). [PubMed: 29218871]
10. Rampasek L, Hidru D, Smirnov P, Haibe-Kains B and Goldenberg A, Dr.VAE: Drug response variational autoencoder (6 2017).
11. Huang S, Ernberg I and Kauffman S, Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective, Semin. Cell Dev. Biol. 20, 869 (9 2009). [PubMed: 19595782]
12. Liu AH, Liu Y-C, Yeh Y-Y and Wang Y-CF, A unified feature disentangler for Multi-Domain image translation and manipulation, in Advances in Neural Information Processing Systems 31, eds. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N and Garnett R (Curran Associates, Inc., 2018) pp. 2595–2604.
13. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C and Stuart JM, The cancer genome atlas Pan-Cancer analysis project, Nat. Genet. 45, 1113 (10 2013). [PubMed: 24071849]
14. Ramos M, curatedTCGAData: Curated data from the cancer genome atlas (TCGA) as MultiAssayExperiment objects (2018).
15. Li B and Dewey CN, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics 12, p. 323 (8 2011). [PubMed: 21816040]
16. Robinson MD, McCarthy DJ and Smyth GK, edger: a bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26, 139 (1 2010). [PubMed: 19910308]
17. McCarthy DJ, Chen Y and Smyth GK, Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation, Nucleic Acids Res. 40, 4288 (5 2012). [PubMed: 22287627]
18. Smyth G, EdgeR bioconductor support https://support.bioconductor.org/p/65890/#65910 (April, 2015), Accessed: 2018-12-11.
19. Chalise P and Fridley BL, Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm, PLoS One 12, p. e0176278 (2017).
20. Alexa A and Rahnenfuhrer J, topGO: enrichment analysis for gene ontology, R package version 2 (2010).

21. McInnes L, Healy J and Melville J, UMAP: Uniform manifold approximation and projection for dimension reduction (2 2018).
22. Hutchins LN, Murphy SM, Singh P and Graber JH, Position-dependent motif characterization using non-negative matrix factorization, Bioinformatics 24, 2684 (12 2008). [PubMed: 18852176]
23. Zhang Y, Iwasaki H, Wang H, Kudo T, Kalka TB, Hennet T, Kubota T, Cheng L, Inaba N, Gotoh M and Others, Cloning and characterization of a new human UDP-N-Acetylα-d-galactosamine: PolypeptideN-Acetylgalactosaminyltransferase, designated pp-GalNAc-T13, that is specifically expressed in neurons and synthesizes GalNAc α-Serine/Threonine antigen, J. Biol. Chem. 278, 573 (2003). [PubMed: 12407114]
24. Kroes RA, Dawson G and Moskal JR, Focused microarray analysis of glyco-gene expressionin human glioblastomas, J. Neurochem. 103, 14 (2007). [PubMed: 17986135]
25. Fischer U, Steffens S, Frank S, Rainov NG, Schulze-Osthoff K and Kramm CM, Mechanisms of thymidine kinase/ganciclovir and cytosine deaminase/ 5-fluorocytosine suicide gene therapy-induced cell death in glioma cells, Oncogene 24, 1231 (2 2005). [PubMed: 15592511]
26. Miller CR, Williams CR, Buchsbaum DJ and Gillespie GY, Intratumoral 5-fluorouracilproduced by cytosine deaminase/5-fluorocytosine gene therapy is effective for experimental human glioblastomas, Cancer Res. 62, 773 (2 2002). [PubMed: 11830532]
27. Zhou Y, Larsen PH, Hao C and Yong VW, CXCR4 is a major chemokine receptor on gliomacells and mediates their survival, J. Biol. Chem. 277, 49481 (12 2002). [PubMed: 12388552]
28. Rempel SA, Dudas S, Ge S and Gutiérrez JA, Identification and localization of the cytokineSDF1 and its receptor, CXC chemokine receptor 4, to regions of necrosis and angiogenesis in human glioblastoma, Clin. Cancer Res. 6, 102 (1 2000). [PubMed: 10656438]
29. Shen R and Seshan VE, FACETS: allele-specific copy number and clonal heterogeneity analysistool for high-throughput DNA sequencing, Nucleic Acids Res. 44, p. e131 (9 2016).

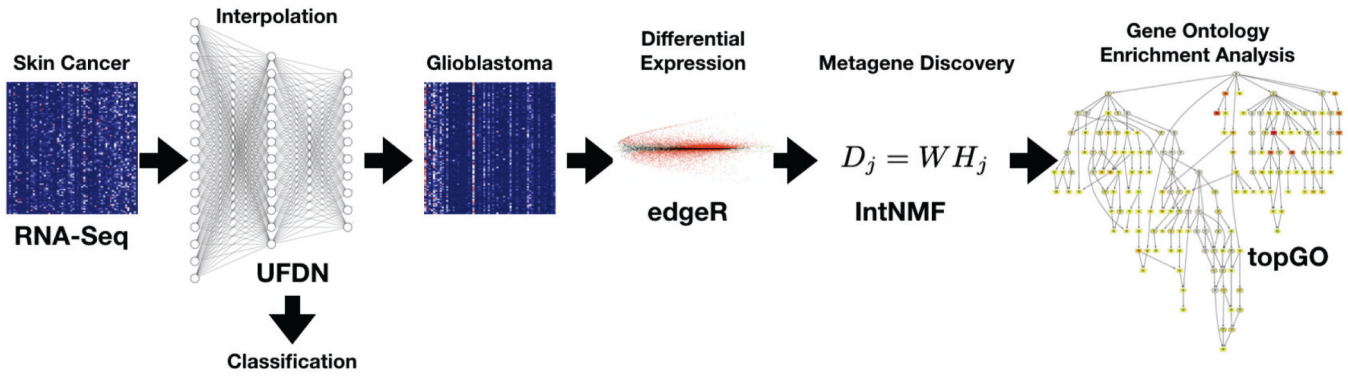

Fig. 1:

We encoded RNA-Seq samples from skin cutaneous melanoma and decoded them into glioblastoma using UFDN-TCGA, then analyzed which sets of genes were changing between cancer types.

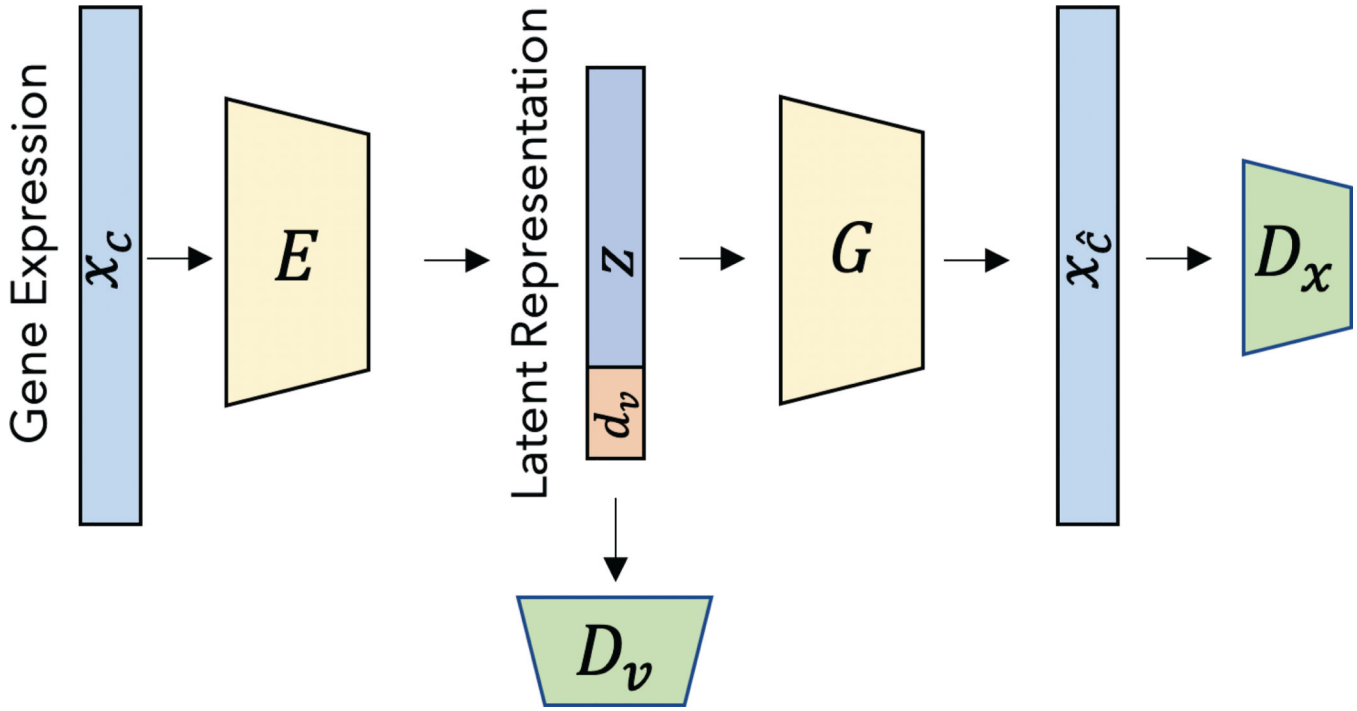

Fig. 2:
The neural network architecture of UFDN-TCGA.

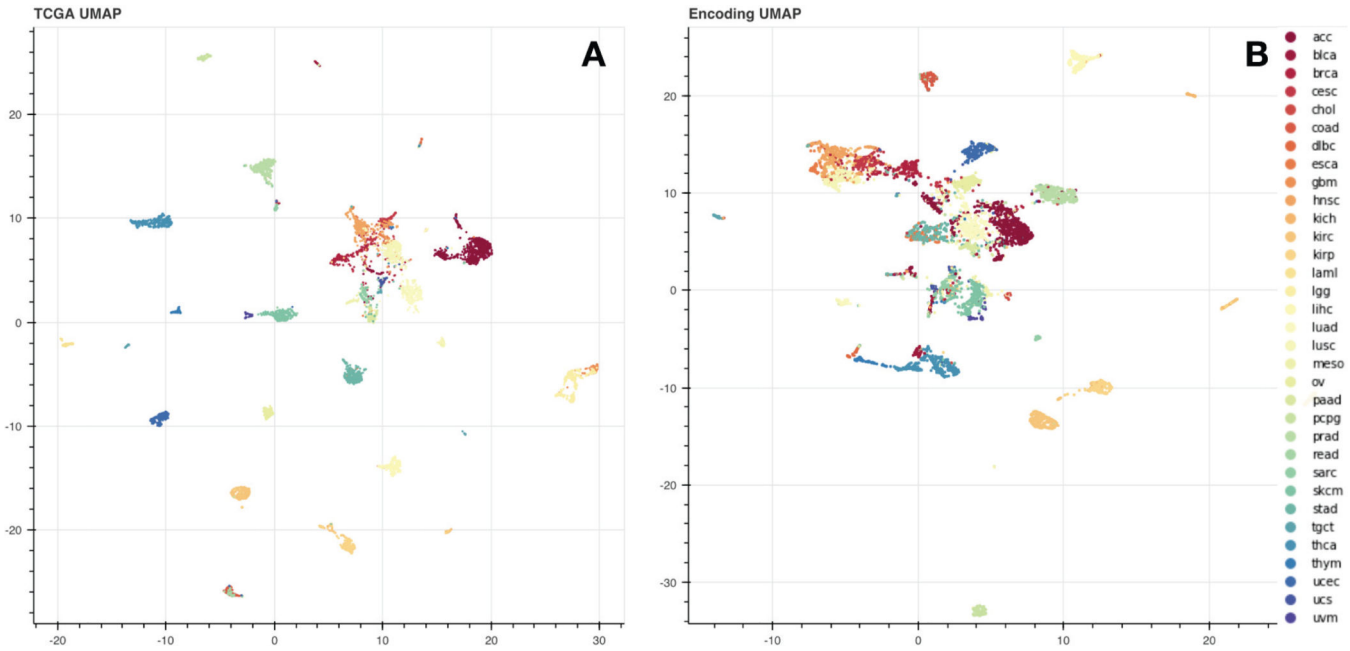

Fig. 3:

UMAP projections of the RNA-Seq TCGA data (Figure 3A) and UFDN latent space encodings of said data (Figure 3B). The full 20,501 dimensional representation of gene expression levels have more cancer specific clusters, while the 100 dimensional latent space encodings have uncoupled from domain information, to some extent.

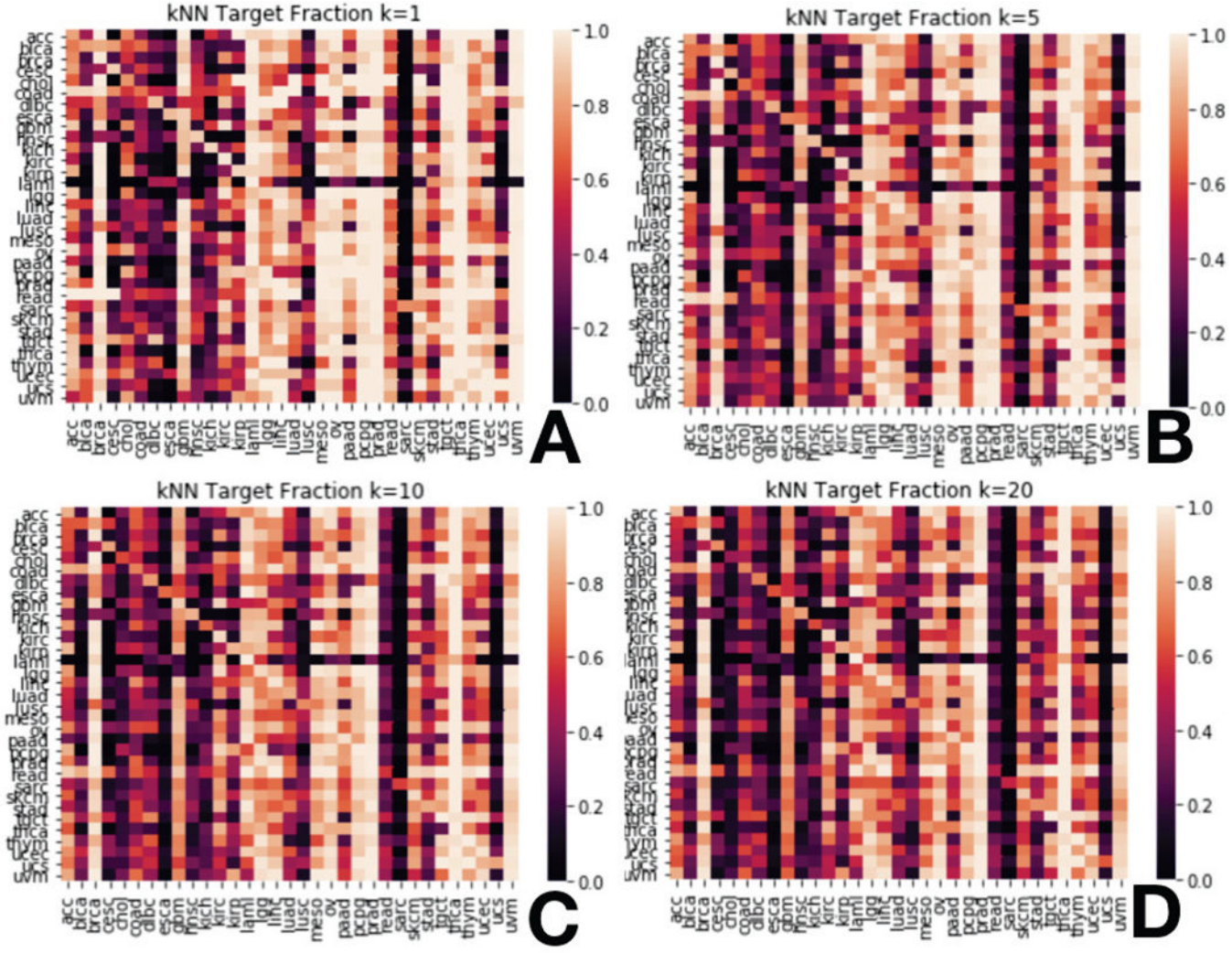

Fig. 4:

The fraction of k nearest neighbors that were in the target domain (the rows of the figures) after decoding from a source domain (the columns of the figures). Some domains were noticeably more difficult to interpolate into. Glioblastoma had strong interpolation results across $k \in [1,5,10,20]$.

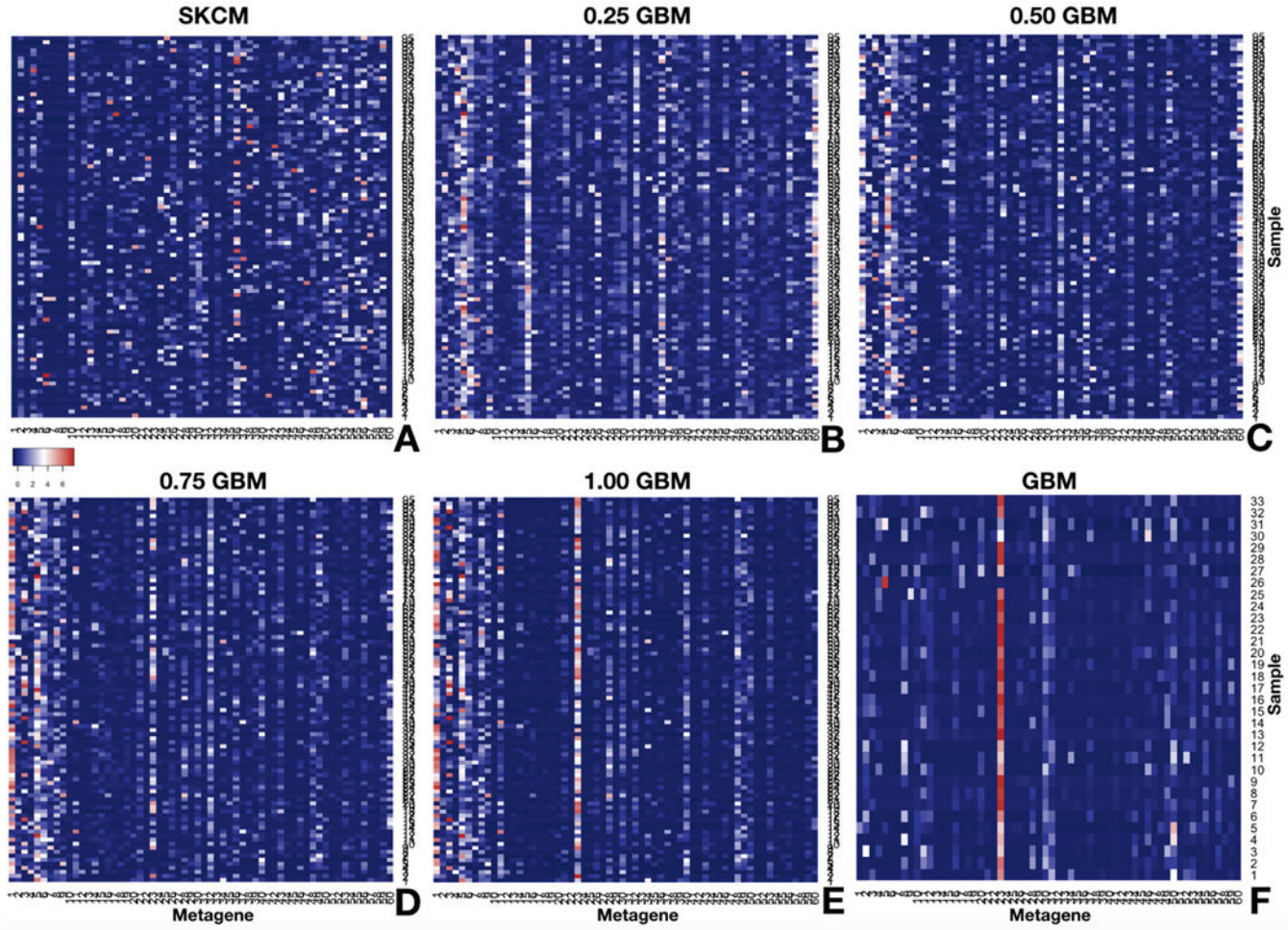

Fig. 5:

Heatmap visualization of the H_j matrices for each interpolation of the SKCM test data set. No row or column reordering was done to keep consistent metagene order across datasets. A full interpolation of SKCM data into GBM data results in a consistent activation of metagene 23 (Figure 5E). This is replicated in H_{GBM} (Figure 5F), which was optimized against the fixed W basis learned for the other 5 datasets.

Table 1:

Results on two classification tasks compared to a random forest baseline.

Algorithm	***Cancer Status* Acc (Train/Test)**	***Domain* Acc (Train/Test)**
Random Forests	**99.60%/98.41%**	**99.65%/95.20%**
UFDN-MSE	—	96.51%/94.10%
Unsupervised UFDN	95.60%/86.14%	—
Semi-supervised UFDN	**99.60%/98.41%**	96.51%/94.10%

Table 2:

The top 15 Gene Ontology Terms enriched in metagene 23

GO ID	Term	p-value
GO:0003676	Nucleic acid binding	5.20E-19
GO:0003735	Structural constituent of ribosome	2.70E-15
GO:0003723	RNA binding	3.90E-14
GO:0003677	DNA binding	1.60E-12
GO:0005198	Structural molecule activity	3.80E-12
GO:0000981	DNA-binding transcription factor activit...	4.70E-12
GO:0003700	DNA-binding transcription factor activit...	3.50E-11
GO:0140110	Transcription regulator activity	2.80E-09
GO:0008376	Acetylgalactosaminyltransferase activity	4.10E-08
GO:0043492	ATPase activity, coupled to movement of...	1.00E-07
GO:0060089	Molecular transducer activity	1.30E-07
GO:0004126	Cytidine deaminase activity	2.10E-07
GO:0019239	Deaminase activity	4.50E-07
GO:0048020	CCR chemokine receptor binding	7.30E-07
GO:0008009	Chemokine activity	8.10E-07