Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes with Variational Autoencoders

Gregory P. Way
Genomics and Computational Biology Graduate Program,
University of Pennsylvania,
Philadelphia, PA 19104, USA
E-mail: gregway@mail.med.upenn.edu

Casey S. Greene*
Department of Systems Pharmacology and Translational Therapeutics
University of Pennsylvania,
Philadelphia, PA 19104, USA
E-mail: csgreene@mail.med.upenn.edu

The Cancer Genome Atlas (TCGA) has profiled over 10,000 tumors across 33 different cancer-types for many genomic features, including gene expression levels. Gene expression measurements capture substantial information about the state of each tumor. Certain classes of deep neural network models are capable of learning a meaningful latent space. Such a latent space could be used to explore and generate hypothetical gene expression profiles under various types of molecular and genetic perturbation. For example, one might wish to use such a model to predict a tumor's response to specific therapies or to characterize complex gene expression activations existing in differential proportions in different tumors. Variational autoencoders (VAEs) are a deep neural network approach capable of generating meaningful latent spaces for image and text data. In this work, we sought to determine the extent to which a VAE can be trained to model cancer gene expression, and whether or not such a VAE would capture biologically-relevant features. In the following report, we introduce a VAE trained on TCGA pan-cancer RNA-seq data, identify specific patterns in the VAE encoded features, and discuss potential merits of the approach. We name our method "Tybalt" after an instigative, cat-like character who sets a cascading chain of events in motion in Shakespeare's "*Romeo and Juliet*". From a systems biology perspective, Tybalt could one day aid in cancer stratification or predict specific activated expression patterns that would result from genetic changes or treatment effects.

Keywords: Deep Learning; Gene Expression; Variational Autoencoder, The Cancer Genome Atlas

1. Introduction

Deep learning has improved the state of the art in many domains, including image, speech, and text processing, but it has yet to make significant enough strides in biomedicine for it to be considered transformative.[1] Nevertheless, several studies have revealed promising results. For instance, Esteva *et al.* used convolutional neural networks (CNNs) to diagnose melanoma from skin images and Zhou and Troyanskaya trained deep models to predict the impact of

*To whom correspondence should be addressed.

non-coding variants.[2,3] However, several domain specific limitations remain. In contrast to image or text data, validating and visualizing learning in biological datasets is particularly challenging. There is also a lack of ground truth labels in biomedical domains, which often limits the efficacy of supervised models. New unsupervised deep learning approaches such as generative adversarial nets (GANs) and variational autoencoders (VAEs) harness the modeling power of deep learning without the need for accurate labels.[4–6]

VAEs and GANs are generative models, which means they learn to approximate a data generating distribution. Through approximation and compression, the models have been shown to capture an underlying data manifold — a constrained, lower dimensional space where data is distributed — and disentangle sources of variation from different classes of data.[7,8] For instance, a recent group trained adversarial autoencoders on chemical compound structures and their growth inhibiting effects in cancer cell lines to learn manifold spaces of effective small molecule drugs.[9,10] Additionally, Rampasek *et al.* trained a VAE to learn a gene expression manifold of reactions of cancer cell lines to drug treatment perturbation.[11] The theoretical basis for modeling cancer using lower dimensional manifolds is established, as it has been previously hypothesized that cancer exists in "basins of attraction" defined by specific pathway aberrations that drive cells toward cancer states.[12] These states could be revealed by data driven manifold learning approaches.

The Cancer Genome Atlas (TCGA) has captured several genomic measurements for over 10,000 different tumors across 33 cancer-types.[13] TCGA has released this data publicly, enabling many secondary analyses, including the training of deep models that predict survival.[14] One data type amenable to modeling manifold spaces is RNA-seq gene expression because it can be used as a proxy to describe tumor states and the downstream consequences of specific molecular aberration. Biology is complex, consisting of multiple nonlinear and often redundant connections among genes, and when a specific pathway aberration occurs, the downstream response to the perturbation is captured in the transcriptome. In the following report, we train and evaluate a VAE on TCGA RNA-seq data. We shall name this model "Tybalt".

2. Methods

2.1. *Model Summary*

VAEs are data driven, unsupervised models that can learn meaningful latent spaces in many contexts. In this work, we aim to build a VAE that compresses gene expression features and reveals a biologically relevant latent space. The VAE is based on an autoencoding framework, which can discover nonlinear explanatory features through data compression and nonlinear activation functions. A traditional autoencoder consists of an encoding phase and a decoding phase where input data is projected into lower dimensions and then reconstructed.[15] An autoencoder is deterministic, and is trained by minimizing reconstruction error. In contrast, VAEs are stochastic and learn the *distribution* of explanatory features over samples. VAEs achieve these properties by learning two distinct latent representations: a mean and standard deviation vector encoding. The model adds a Kullback-Leibler (KL) divergence term to the reconstruction loss, which also regularizes weights through constraining the latent vectors to match a Gaussian distribution. In a VAE, these two representations are learned concurrently

through the use of a reparameterization trick that permits a back propagated gradient.[4]

2.2. *Model Implementation*

We evaluated VAEs in this context because previous work demonstrated that gene expression was amenable to modeling with reconstruction loss, as several studies have previously used autoencoders to extract knowledge from gene expression data.[16–18] We also sought to evaluate the extent to which VAEs characterized the underlying manifold space. VAEs have been shown to generate "blurry" data compared with other generative models, including GANs, but VAEs are also generally more stable to train.[19] We trained our VAE model, Tybalt, with the following architecture: 5,000 input genes encoded to 100 features and reconstructed back to the original 5,000 (Figure 1A). The 5,000 input genes were selected based on highest variability by median absolute deviation (MAD) in the TCGA pan-cancer dataset.

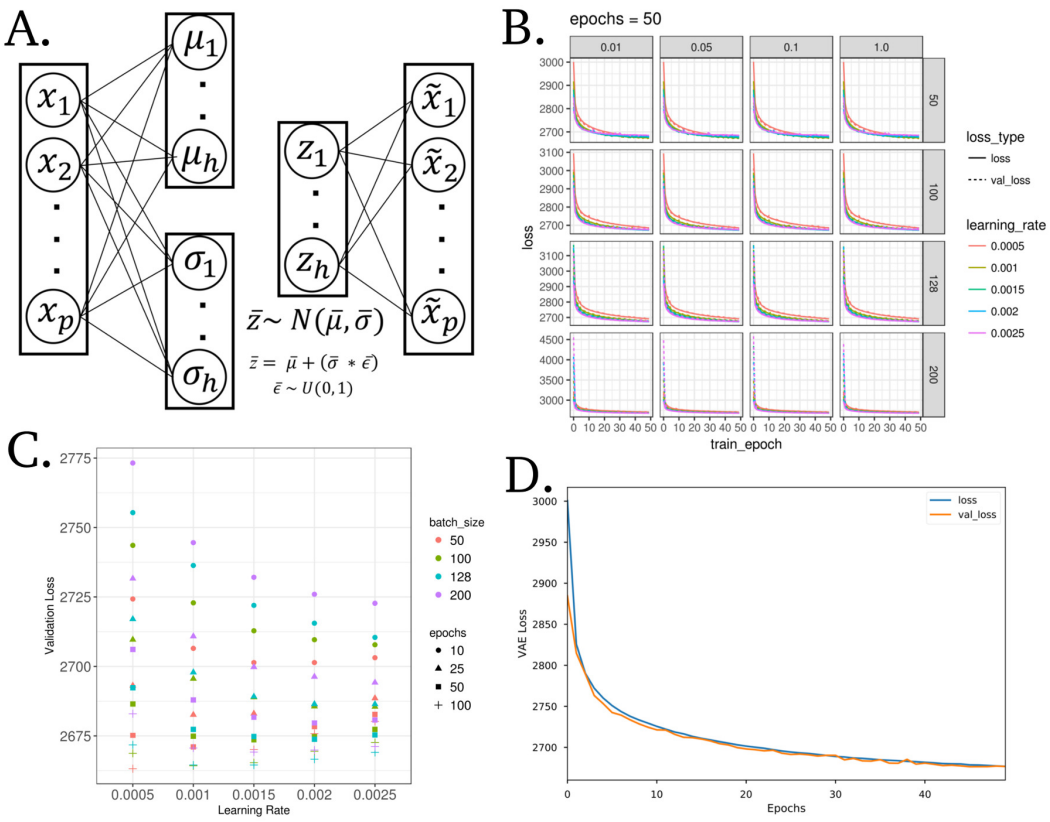

Fig. 1. *A variational autoencoder (VAE) applied to model gene expression data.* **(A)** Model wire diagram of Tybalt encoding a gene expression vector ($p = 5{,}000$) into mean (μ) and standard deviation (σ) vectors ($h = 100$). A reparameterization trick enables learning z, which is then reconstructed back to input ($\tilde{x}$). **(B)** Training and validation VAE loss across training epochs. Shown across vertical and horizontal facets are values of κ and batch size, respectively. **(C)** Final validation loss for all parameters with $\kappa = 1$. **(D)** VAE loss for training and testing sets through optimized model training.

We initially trained Tybalt without batch normalization,[20] but observed that when we included batch normalization in the encoding step, we trained faster and with heterogeneous

feature activation. Batch normalization adds additional feature regularization by scaling activations to zero mean and unit variance, which has been observed to speed up training and reduce batch to batch variability thus increasing generalizability. We trained Tybalt with an Adam optimizer,[21] included rectified linear units[22] and batch normalization in the encoding stage, and sigmoid activation in the decoding stage. We built Tybalt in Keras (version 2.0.6)[23] with a TensorFlow backend (version 1.0.1).[24] For more specific VAE illustrations and walkthroughs refer to an extended tutorial[25] and these intuitive blog posts.[26,27]

2.3. *Parameter Selection*

We performed a parameter sweep over batch size (50, 100, 128, 200), epochs (10, 25, 50, 100), learning rates (0.005, 0.001, 0.0015, 0.002, 0.0025) and warmups (κ) (0.01, 0.05, 0.1, and 1). κ controls how much the KL divergence loss contributes to learning, which effectively transitions a deterministic autoencoder to a VAE.[28,29] For instance, a $\kappa = 0.1$ would add 0.1 to a weight on the KL loss after each epoch. After 10 epochs, the KL loss will have equal weight as the reconstruction loss. We did not observe κ to influence model training (Figure 1B), so we kept $\kappa = 1$ for downstream analyses. We evaluated train and test set loss at each epoch. The test set was a random 10% partition of the full data. In general, training was relatively stable for many parameter combinations, but was consistently worse for larger batches, particularly with low learning rates. Ultimately, the best parameter combination based on validation loss was batch size 50, learning rate 0.0005, and 100 epochs (Figure 1C). Because training stabilized after about 50 epochs, we terminated training early. Training and testing loss across all 50 epochs is shown in Figure 1D. We performed the parameter sweep on a cluster of 8 NVIDIA GeForce GTX 1080 Ti GPUs on the PMACS cluster at The University of Pennsylvania.

2.4. *Input Data*

The input data consisted of level 3 TCGA RNA-seq gene expression data for 10,459 tumors measured by the 5,000 most variably expressed genes. The highest expressing genes were defined by median absolute genes (MAD). In total, there were 33 different cancer-types (including glioblastoma, ovarian, breast, lung, bladder cancer, etc.) profiled, each with varying number of tumors. We accessed RNA-seq data from the UCSC Xena data browser on March 8th, 2016 and archived the data in Zenodo.[30] To facilitate training, we min-maxed scaled RNA-seq data to the range of $0 - 1$. We used corresponding clinical data accessed from the Snaptron web server.[31]

2.5. *Interpretation of Gene Weights*

Much like the weights of a deterministic autoencoder, Tybalt's decoder weights captured the contribution of specific genes to each learned feature.[16–18] For each feature, gene weights followed a normal distribution with many genes having low weight and few genes with high weights at the extreme tails. In order to characterize patterns explained by selected encoded features of interest, we performed overrepresentation pathway analyses separately for both positive and negative high weight genes; defined by greater than 2.5 standard deviations

above or below the mean, respectively. We used WebGestalt,[32] with a background of the 5,000 assayed genes, to perform the analysis over gene ontology (GO) biological process terms.[33] P values are presented after an FDR adjustment.

2.6. *The Latent Space of Ovarian Cancer Subtypes*

Image processing studies have shown the remarkable ability of generative models to mathematically manipulate learned latent dimensions.[34,35] For example, subtracting the image latent representation of a neutral man from a smiling man and adding it to a neutral women, resulted in a vector associated with a smiling woman. We were interested in the extent to which Tybalt learned a manifold representation that could be manipulated mathematically to identify state transitions across HGSC tumors.

We performed an experiment to test whether or not Tybalt learned manifold differences of distinct high grade serous ovarian cancer (HGSC) subtypes. Previously, several groups identified four HGSC subtypes using gene expression.[36–38] The TCGA naming convention of these subtypes is mesenchymal, proliferative, immunoreactive, and differentiated. However, the four HGSC subtypes were not consistently defined across populations; the data suggested the presence of three subtypes or fewer.[39] The study observed that the immunoreactive/mesenchymal and differentiated/proliferative tumors consistently collapsed together when setting clustering algorithms to find 2 subtypes.[39] This observation may suggest the presence of distinct gene expression programs existing on an activation spectrum driving differences in these subtypes. Therefore, we hypothesized that Tybalt would learn the manifold of gene expression spectra existing in differential proportions across these subtypes.

To facilitate a quick search for explanatory features, we subtracted the mean HGSC mesenchymal subtype vector from the mean immunoreactive subtype vector and the mean proliferative subtype vector from the mean differentiated subtype vector. We used tumor subtype assignments provided for TCGA samples in Verhaak *et al.* 2013.[40] If Tybalt learned a biological manifold, this subtraction would result in the identification of biologically relevant features stratifying tumors of specific subtypes with a continuum of expression states.

2.7. *Enabling Exploration through Visualization*

We provide a Shiny app to interactively visualize activation patterns of encoded Tybalt features with covariate information at `https://gregway.shinyapps.io/pancan_plotter/`.

3. Results

Tybalt compressed tumors into a lower dimensional space, acting as a nonlinear dimensionality reduction algorithm. Tybalt learned which genes contributed to each feature, potentially capturing aberrant pathway activation and treatment vulnerabilities. Tybalt was unsupervised; therefore, it could learn both known and unknown biological patterns. In order to determine if the features captured biological signals, we characterized both sample- and gene-specific activation patterns.

3.1. *Tumors were encoded in a lower dimensional space*

The tumors were encoded from original gene expression vectors of 5,000 MAD genes into a lower dimensional vector of length 100. To determine if the sample encodings faithfully recapitulated large, tissue specific signals in the data, we visualized sample-specific Tybalt encoded features (z vector for each sample) by t-distributed stochastic neighbor embedding (t-SNE).[41] We observed similar patterns for Tybalt encodings (Figure 2A) as compared to $0-1$ normalized RNA-seq data (Figure 2B). Tybalt geometrically preserved well known relationships, including similarities between glioblastoma (GBM) and low grade glioma (LGG). Importantly, the recapitulation of tissue-specific signal was captured by non-redundant, highly heterogeneous features (Figure 2C). Based on the hierarchical clustering dendrogram, the features appeared to be capturing distinct signals. For instance, tumor versus normal and patient sex are large signals present in cancer gene expression, but they were distributed uniformly in the clustering solution indicating non-redundant feature activations.

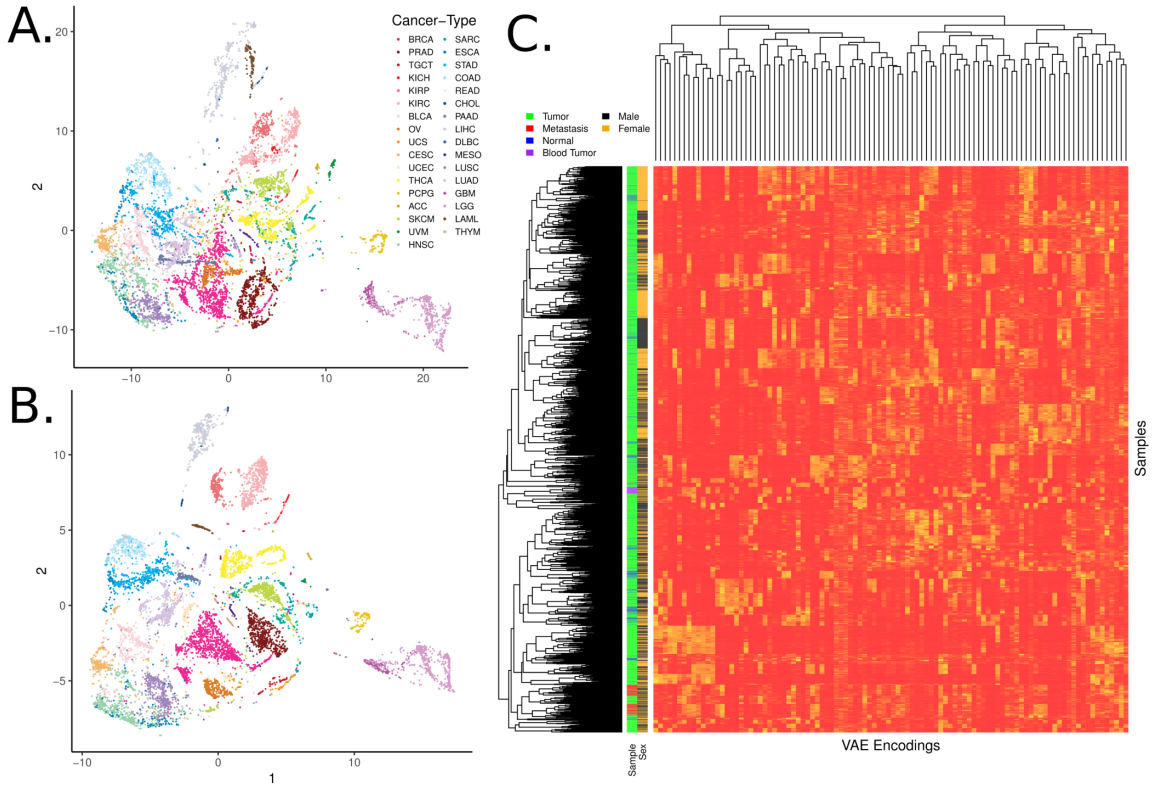

Fig. 2. *Samples encoded by a variational autoencoder retain biological signals.* **(A)** t-distributed stochastic neighbor embedding (t-SNE) of TCGA pan-cancer tumors with Tybalt encoded features. **(B)** t-SNE of 0-1 normalized gene expression features. Tybalt retains similar signals. **(C)** Full Tybalt encoding features by sample heatmap. Shown are different sample-types and patient sex.

3.2. *Features represent biological signal*

Our goal was to train and evaluate Tybalt on its ability to learn biological signals in the data and not to perform a comprehensive survey of learned features. Therefore, we investigated whether or not Tybalt could distinguish patient sex and patterns of metastatic activation. We determined that the model extracted patient sex robustly (Figure 3A). Feature encoding 82 nearly perfectly separated samples by sex. Furthermore, we identified a set of nodes that together identified skin cutaneous melanoma (SKCM) tumors of both primary and metastatic origin (Figure 3B).

The weights used to decode the hidden layer (z vector) back into a high-fidelity reconstruction of the input can capture important and consistent biological patterns embedded in the gene expression data.[16–18] For instance, there were only 17 genes needed to identify patient sex (Figure 3C). These genes were mostly located on sex chromosomes. The two positive weight genes were X inactivation genes *XIST* and *TSIX*, while the negative weight genes were mostly Y chromosome genes such as *EIF1AY*, *UTY*, and *KDM5D*. This result served as a positive control that the unsupervised model was able to construct a feature that described a clearly biological source of variance in the data.

There were several genes contributing to the two encoded features that separated the SKCM tumors (Figure 3D). Several genes existed in the high weight tails of each distribution for feature encodings 53 and 66. We performed an overrepresentation pathway analysis on the high weight genes. In general, several pathways were identified as overrepresented in the set as compared to random (Table 1). The samples had intermediate to high levels of feature encoding 53, which did not correspond to any known GO term, potentially indicating an unknown but important biological process. The samples also had intermediate to high levels of encoding 66 which implicated GO terms related to cholesterol, ethanol, and lipid metabolism. The SKCM samples had consistently high activation of both encoded features, which separated them from other tumors. Nevertheless, more research is required to determine how VAE features could be best interpreted in this context.

Table 1. Significant pathways for nodes separating metastasis/melanoma

Encoding	Tail	Pathway	Adj. p value
53	-	Response to xenobiotic stimulus	$3.6e^{-5}$
53	-	Uronic acid metabolic process	$5.5e^{-5}$
53	+	*No significant pathways identified*	1
66	-	Dorsal/ventral pattern formation	$2.7e^{-1}$
66	-	Forebrain neuron differentiation	$2.7e^{-1}$
66	-	Olfactory bulb interneuron development	$2.7e^{-1}$
66	+	Regulation of intestinal cholesterol absorption	$3.0e^{-2}$
66	+	Ethanol oxidation	$4.0e^{-2}$
66	+	Lipid catabolic process	$4.0e^{-2}$

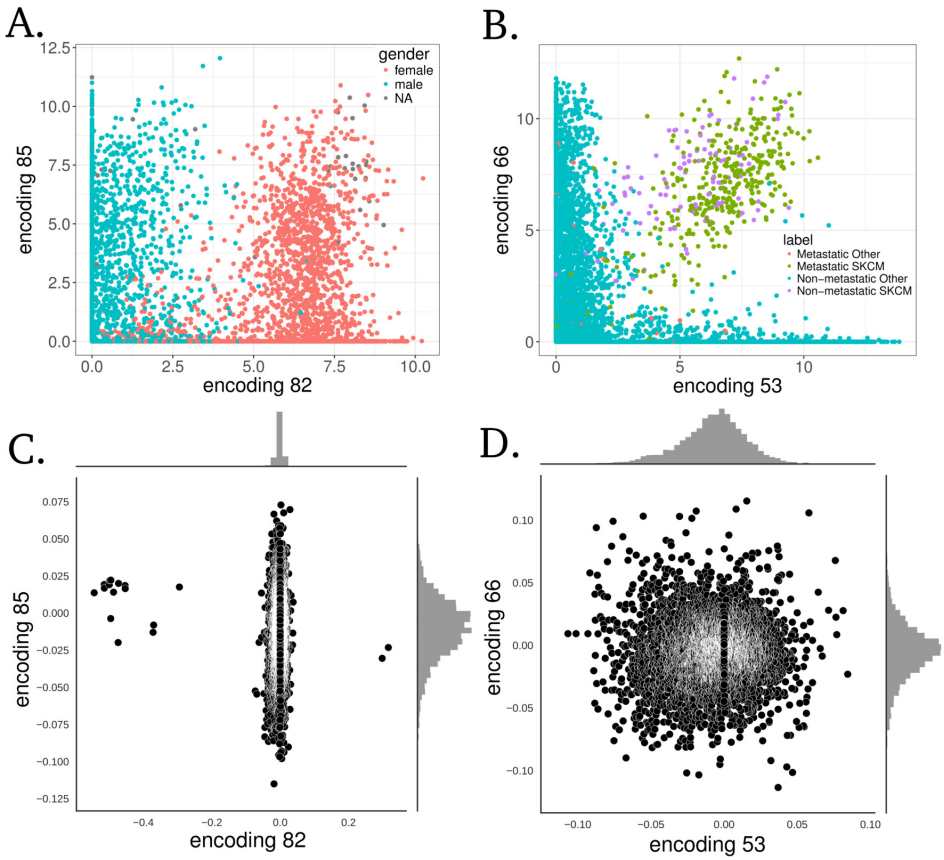

Fig. 3. *Specific Tybalt encoded features capture biological signals.* **(A)** Encoding 82 stratified patient sex. **(B)** Together, encodings 53 and 66 separated metastatic tumors (mostly skin cutaneous melanoma). Distributions of gene coefficients contributing to each plot above for **(C)** patient sex and **(D)** metastases.

3.3. *Interpolating the lower dimensional manifold of HGSC subtypes*

To characterize the largest differences between the mesenchymal/immunoreactive and proliferative/differentiated HGSC subtypes, we performed a series of mean HGSC subtype vector subtractions in Tybalt latent space:

$$\bar{\theta}_k = \frac{\sum_{i=1}^{n} z_{i,1}(i_k = k)}{n_k}, \ldots, \frac{\sum_{i=1}^{n} z_{i,100}(i_k = k)}{n_k} \tag{1}$$

$$\bar{\theta}_{\text{immunoreactive}} - \bar{\theta}_{\text{mesenchymal}} = \bar{\theta}_{\text{immuno-mes}} \tag{2}$$

$$\bar{\theta}_{\text{differentiated}} - \bar{\theta}_{\text{proliferative}} = \bar{\theta}_{\text{diff-prolif}} \tag{3}$$

Where $(i_k = k)$ is an indicator function if sample i has membership with subtype k and z is the encoded layer. The largest feature encoding difference between the mean HGSC mesenchymal and the mean immunoreactive subtype ($\bar{\theta}_{\text{immuno-mes}}$) was encoding 87 (Figure 4A). Encoding 77 and encoding 56 (Figure 4B) also distinguished the mesenchymal and immunoreactive subtypes. The largest feature encoding differences between the mean proliferative and

the mean differentiated subtype ($\bar{\theta}_{\text{diff-prolif}}$) were contributed by encoding 79 (Figure 4C) and encoding 38 (Figure 4D). Interestingly, encoding 38 had high mean activation in both the immunoreactive and differentiated subtypes.

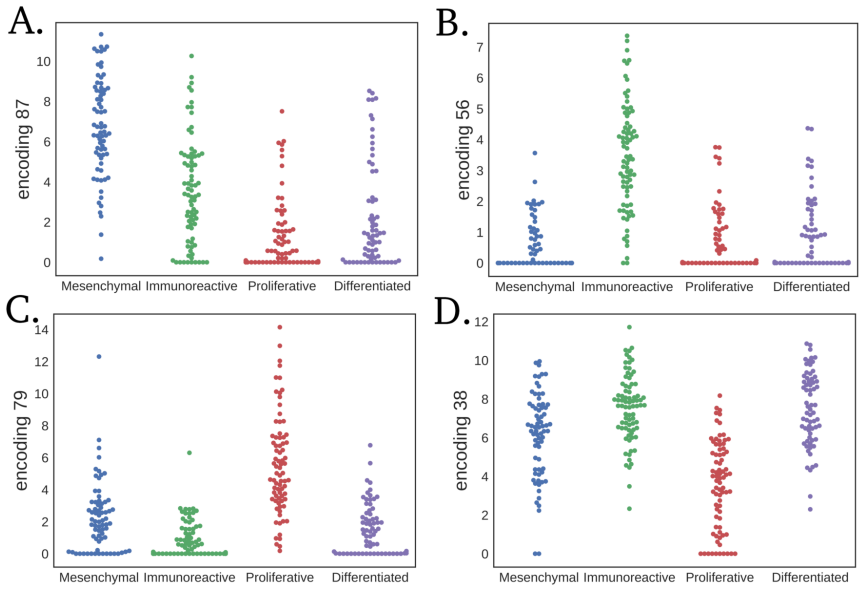

Fig. 4. *Largest mean differences in HGSC subtype vector subtraction for each subtype.* Subtracting the mesenchymal subtype by the immunoreactive results in distribution differences in **(A)** feature encoding 87 and **(B)** encoding 56. Subtracting the proliferative subtype by the differentiated subtype results in differences between **(C)** feature encoding 79 and **(D)** encoding 38.

Samples with high levels of encoding 87 had higher expression of genes associated with cilia (Table 2). The mesenchymal subtype had the highest encoding 87 activation suggesting an association with cilia abundance, which has been observed to be related with autophagy and aberrant hedgehog signaling in epithelial ovarian cancer cells.[42,43] Encoding 56 was associated with microvilli and brush border expression (Table 2), and the immunoreactive subtype displayed the highest activation. Microvilli expression can be induced by hyaluronan,[44] which itself is associated with chemotherapy resistance.[45] It is possible that tumors in the immunoreactive subtype produce differential hyaluronan and are therefore more susceptible to treatment. Encoding 79 is mostly expressed in the proliferative subtype. There were no significant GO terms associated with the positive tail of encoding 79, but pathways involving coagulation and inflammation processes such as fibrinolysis and negative regulation of wound healing were represented in the negative tail. These pathways are important components of ovarian cancer proliferation.[46] Lastly, encoding 38 was observed to have a higher mean in immunoreactive and differentiated as compared to the mesenchymal and proliferative subtypes. Encoding 38 was associated with lipid metabolism. Lipids have been shown to be associated with adipocytes and ovarian tumor growth,[47] hinting towards a potential mechanistic link for the better survival observed in the immunoreactive and differentiated subtypes.[38]

Table 2. Summary of significantly overrepresented pathways separating HGSC subtypes

Encoding	Tail	Pathway	Subtype Enrichment	Adj. p value
87	+	Mesenchymal	Microtubule bundle formation	$< 1.0e^{-15}$
87	+	Mesenchymal	Cilium movement	$< 1.0e^{-15}$
77	+	Mesenchymal	*No significant pathways identified*	1
56	-	Mesenchymal	Macromolecule metabolic process	$6.65e^{-07}$
56	-	Mesenchymal	Dorsal/ventral pattern formation	$4.95e^{-06}$
56	-	Mesenchymal	Keratinization	$1.22e^{-02}$
56	+	Immunoreactive	Regulation of microvillus length	$5.07e^{-04}$
56	+	Immunoreactive	Brush border assembly	$5.07e^{-04}$
56	+	Immunoreactive	Protein glycosylation	$1.27e^{-03}$
77	-	Immunoreactive	Extracellular region	$2.12e^{-03}$
87	-	Immunoreactive	*No significant pathways identified*	1
79	+	Proliferative	*No significant pathways identified*	1
38	-	Proliferative	Cellular lipid catabolic process	$1.21e^{-02}$
38	+	Differentiated	*No significant pathways identified*	1
79	-	Differentiated	Fibrinolysis	$4.21e^{-03}$
79	-	Differentiated	Negative regulation of wound healing	$4.21e^{-03}$

4. Conclusion

Tybalt is a promising model but still requires careful validation and more comprehensive evaluation. We observed that the encoded features recapitulated tissue specific patterns. We determined that the learned features were generally non-redundant and could disentangle large sources of variation in the data, including patient sex and SKCM. Interpretation of the decoding layer weights helped to identify the contribution of different genes and pathways promoting disparate biological patterns. However, interpretation must be performed with caution.

VAEs provide similar benefits as autoencoders, but they also have the ability to learn a manifold with meaningful relationships between samples. This manifold could represent differing pathway activations, transitions between cancer states, or indicate particular tumors vulnerable to specific drugs. We performed initial testing to determine if we could traverse the underlying manifold by subtracting out cancer-type specific mean activations. While we identified several promising functional relationships existing in a spectrum of activation patterns, rigorous experimental testing would be required to draw strong conclusions about the biological implications. The specific subtype associations must be confirmed in independent datasets and the processes must be confirmed experimentally. Further testing is required to confirm that Tybalt catalogued an interpretable manifold capable of interpolation between cancer states. In the future, we will develop higher capacity models and increased evaluation/interpretation efforts to catalog Tybalt encoded RNA-seq expression patterns present in specific cancer-types. This effort will lead to widespread stratification of expression patterns and enable accurate detection of samples who may benefit from specific targeted therapies.

5. Reproducibility

We provide all scripts to reproduce and to build upon this analysis under an open source license at `https://github.com/greenelab/tybalt`.[48]

Acknowledgments

This work was supported by NIH grants R01 CA200854 (CSG) and T32 HG000046 (GPW), as well as GBMF 4552 from the Gordon and Betty Moore Foundation (CSG). We would like to thank Brett K. Beaulieu-Jones for helpful discussions. This is a preprint of an article submitted for consideration in Pacific Symposium on Biocomputing ©2018, World Scientific Publishing Co., http://psb.stanford.edu.

References

1. T. Ching *et al.*, *bioRxiv:10.1101/142760* (May 2017).
2. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun, *Nature* **542**, 115 (February 2017).
3. J. Zhou and O. G. Troyanskaya, *Nature Methods* **12**, 931 (October 2015).
4. D. P. Kingma and M. Welling, *arXiv:1312.6114 [cs, stat]* (December 2013).
5. D. J. Rezende, S. Mohamed and D. Wierstra, *arXiv:1401.4082 [cs, stat]* (January 2014).
6. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, *arXiv:1406.2661 [cs, stat]* (June 2014).
7. I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed and A. Lerchner, *arXiv:1606.05579 [cs, q-bio, stat]* (June 2016).
8. P. Park.
9. A. Kadurin, A. Aliper, A. Kazennov, P. Mamoshina, Q. Vanhaelen, K. Khrabrov, A. Zhavoronkov, A. Kadurin, A. Aliper, A. Kazennov, P. Mamoshina, Q. Vanhaelen, K. Khrabrov and A. Zhavoronkov, *Oncotarget* **8**, 10883 (December 2016).
10. A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper and A. Zhavoronkov, *Molecular Pharmaceutics* (July 2017).
11. L. Rampasek, D. Hidru, P. Smirnov, B. Haibe-Kains and A. Goldenberg, *arXiv:1706.08203 [stat]* (June 2017).
12. S. Huang, I. Ernberg and S. Kauffman, *Seminars in cell & developmental biology* **20**, 869 (September 2009).
13. J. N. Weinstein, E. A. Collisson, G. B. Mills, K. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander and J. M. Stuart, *Nature genetics* **45**, 1113 (October 2013).
14. K. Chaudhary, O. B. Poirion, L. Lu and L. Garmire, *bioRxiv* , p. 114892 (March 2017).
15. P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol, Extracting and Composing Robust Features with Denoising Autoencoders, in *Proceedings of the 25th International Conference on Machine Learning*, ICML '08 (ACM, New York, NY, USA, 2008).
16. J. Tan, M. Ung, C. Cheng and C. S. Greene, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* , 132 (2015).
17. J. Tan, J. H. Hammond, D. A. Hogan and C. S. Greene, *mSystems* **1**, e00025 (February 2016).
18. L. Chen, C. Cai, V. Chen and X. Lu, *BMC Bioinformatics* **17**, p. S9 (January 2016).
19. A. Lamb, V. Dumoulin and A. Courville, *arXiv:1602.03220 [cs, stat]* (February 2016), arXiv: 1602.03220.
20. S. Ioffe and C. Szegedy, *arXiv:1502.03167 [cs]* (February 2015).
21. D. P. Kingma and J. Ba, *arXiv:1412.6980 [cs]* (December 2014).

22. V. Nair and G. E. Hinton, Rectified Linear Units Improve Restricted Boltzmann Machines, in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10 (Omnipress, USA, 2010).
23. F. Chollet and others, *Keras* (GitHub, 2015).
24. M. Abadi *et al.*, *arXiv:1603.04467 [cs]* (March 2016).
25. C. Doersch, *arXiv:1606.05908 [cs, stat]* (June 2016).
26. K. Franz, *Variational Autoencoders Explained*, 2016).
27. H. Saghir, *An intuitive understanding of variational autoencoders without any formula*, 2017).
28. T. Raiko, H. Valpola, M. Harva and J. Karhunen, *J. Mach. Learn. Res.* **8**, 155 (May 2007).
29. C. K. Snderby, T. Raiko, L. Maale, S. K. Snderby and O. Winther, *arXiv:1602.02282 [cs, stat]* (February 2016).
30. G. Way, *Data Used For Training Glioblastoma Nf1 Classifier* (Zenodo, June 2016).
31. C. Wilks, P. Gaddipati, A. Nellore and B. Langmead, *bioRxiv* , p. 097881 (January 2017).
32. J. Wang, S. Vasaikar, Z. Shi, M. Greer and B. Zhang, *Nucleic Acids Research* **45**, W130 (July 2017).
33. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nature Genetics* **25**, 25 (May 2000).
34. A. Dosovitskiy, J. T. Springenberg and T. Brox, Learning to generate chairs with convolutional neural networks (IEEE, June 2015).
35. A. Radford, L. Metz and S. Chintala, *arXiv:1511.06434 [cs]* (November 2015).
36. R. W. Tothill *et al.*, *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **14**, 5198 (August 2008).
37. T. C. G. A. R. Network, *Nature* **474**, 609 (June 2011).
38. G. E. Konecny, C. Wang, H. Hamidi, B. Winterhoff, K. R. Kalli, J. Dering, C. Ginther, H.-W. Chen, S. Dowdy, W. Cliby, B. Gostout, K. C. Podratz, G. Keeney, H.-J. Wang, L. C. Hartmann, D. J. Slamon and E. L. Goode, *Journal of the National Cancer Institute* **106** (October 2014).
39. G. P. Way, J. Rudd, C. Wang, H. Hamidi, B. L. Fridley, G. E. Konecny, E. L. Goode, C. S. Greene and J. A. Doherty, *G3: Genes, Genomes, Genetics* , p. g3.116.033514 (January 2016).
40. R. G. Verhaak *et al.*, *Journal of Clinical Investigation* (December 2012).
41. L. v. d. Maaten and G. Hinton, *Journal of Machine Learning Research* **9**, 2579 (2008).
42. M. Cao and Q. Zhong, *Cilia* **5** (December 2015).
43. D. L. Egeberg, M. Lethan, R. Manguso, L. Schneider, A. Awan, T. S. Jrgensen, A. G. Byskov, L. B. Pedersen and S. T. Christensen, *Cilia* **1**, p. 15 (2012).
44. A. Kultti, K. Rilla, R. Tiihonen, A. P. Spicer, R. H. Tammi and M. I. Tammi, *Journal of Biological Chemistry* **281**, 15821 (June 2006).
45. C. Ricciardelli, M. P. Ween, N. A. Lokman, I. A. Tan, C. E. Pyragius and M. K. Oehler, *BMC Cancer* **13** (December 2013).
46. X. Wang, E. Wang, J. J. Kavanagh and R. S. Freedman, *Journal of Translational Medicine* **3**, p. 25 (June 2005).
47. K. M. Nieman, H. A. Kenny, C. V. Penicka, A. Ladanyi, R. Buell-Gutbrod, M. R. Zillhardt, I. L. Romero, M. S. Carey, G. B. Mills, G. S. Hotamisligil, S. D. Yamada, M. E. Peter, K. Gwin and E. Lengyel, *Nature Medicine* **17**, 1498 (October 2011).
48. G. Way and C. Greene, *greenelab/tybalt: Initial Development Release*, tech. rep., Zenodo (July 2017).