Discriminative Subtyping of Lung Cancers from Histopathology Images via Contextual Deep Learning

Benjamin J. Lengerich*[1], Maruan Al-Shedivat*[1], Amir Alavi[1], Jennifer Williams[1], Sami Labbaki[1], and Eric P. Xing[1,2]

[1]Carnegie Mellon University
[2]Mohamed bin Zayed University of Artificial Intelligence

November 18, 2022

Abstract

Summarizing multiple data modalities into a parsimonious cancer "subtype" is difficult because the most informative representation of each patient's disease is not observed. We propose to model these latent summaries as *discriminative subtypes*: sample representations which induce accurate and interpretable sample-specific models for downstream predictions. In this way, discriminative subtypes, which are shared between data modalities, can be estimated from one data modality and optimized according to the predictions induced in another modality. We apply this approach to lung cancer by training a deep neural network to predict discriminative subtypes from histopathology images, and use these predicted subtypes to generate models which classify adenocarcinoma, squamous cell carcinoma, and healthy tissue based on transcriptomic signatures. In this way, we optimize the latent discriminative subtypes through induced prediction loss, and the discriminative subtypes are interpreted with standard interpretation of transcriptomic predictive models. Our framework achieves state-of-the-art classification accuracy (F1-score of 0.97) and identifies discriminative subtypes which link histopathology images to transcriptomic explanations without requiring pre-specification of morphological patterns or transcriptomic processes.

Introduction

Cancer is a heterogeneous disease which can be phenotypically variable even for tumors of a single tissue [PSE+00, SPT+01, FPS13]. For this reason, subtype discovery — the classification of tumors into biologically- and clinically-relevant groups — is a crucial task to advance the understanding and treatment of cancers. A common approach is to define *molecular subtypes* with a small number of marker genes [vdVHvV+02, RVG+11, CCL+08, CBCB19, BKH+02] identified through immunohistochemistry or gene expression analysis. Characterization of these molecular subtypes is continuously advancing, and much recent work has focused on identifying fine-grained descriptions which comprise previously-characterized subgroups [MBW+13, KMK+13, RPB+19, CGAR12, WYH+10]. While subtypes are continuously being refined through experimentation, a fundamental question remains: how would we know when the optimal cancer subtypes have been defined?

This question inspires us to propose a machine-learning based criterion for optimality of subtyping classifications: a set of subtypes is optimal if it most concisely describes cancers in such a way that downstream predictions require minimal extra information to be maximally accurate. We call these optimal subtypes *discriminative subtypes* and show that they can be estimated through contextual machine learning. In lung cancers, these discriminative subtypes correspond to distinct transcriptomic and morphological patterns, suggesting that different tumors should be understood in light of different pathways.

*These authors contributed equally.
blengeri@mit.edu

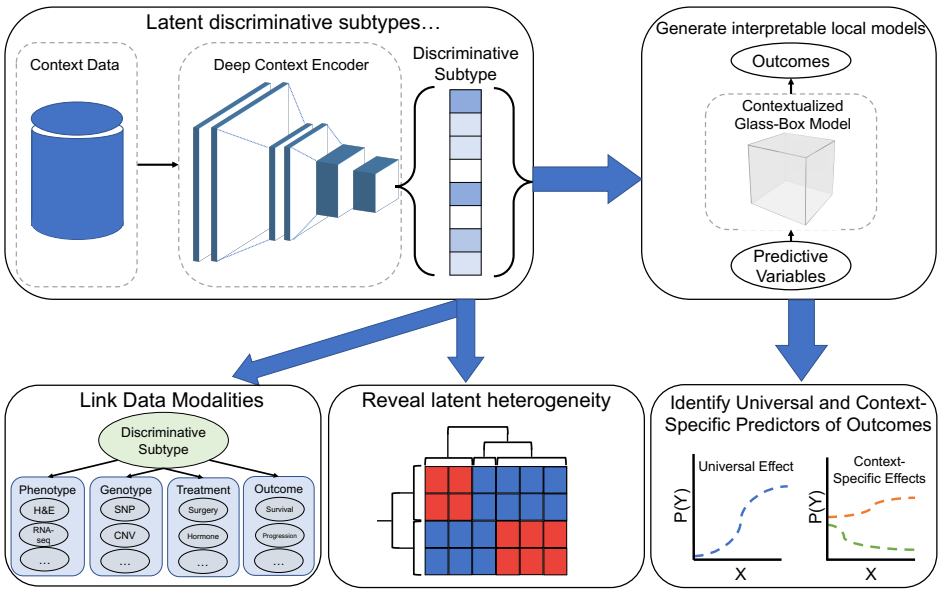

Figure 1: We propose that the most concise description of the many variables regarding oncology patients is an unobserved latent variable, and we seek to estimate this latent *discriminative subtype* which improves the discriminative ability of downstream glass-box predictors. With the optimality of these discriminative subtypes defined by the predictive ability of downstream models, they can be optimized through end-to-end contextual machine learning without pre-specifying any subtypes or factors. This links data modalities, reveals latent heterogeneity, and identifies universal and context-specific predictors of outcomes.

Contributions

We examine three hypotheses surrounding this task of distinguishing LUAD and LUSC tumors: (1) there exist "discriminative subtypes" of lung cancers which improve downstream predictive models, (2) these subtypes can be predicted from histopathology images via contextual deep learning, and (3) the discriminative subtypes correspond to biological processes and morphological patterns. The validation of these hypotheses suggests that macroscopic morphological patterns can be used to predict intra-cellular genomic profiles which predict outcomes, even when the genomic profiles are unknown *a priori*. Beyond lung cancer, discriminative subtypes and contextual deep learning can improve biomedical understanding of cancer subtypes by connecting histopathology images to gene expression patterns in a new end-to-end, fully-differentiable manner which does not require pre-specifying which transcriptomic patterns are of interest.

Discriminative Subtypes

We propose a *discriminative subtype* (Figure 1) as a latent variable which captures the variation in many observable variables, and optimally "contextualizes" predictive tasks by providing a compressed view of relevant information. Discriminative subtypes may correspond to, but are not limited to, previously-characterized molecular subtypes. The optimal descriptions of tumors are multi-faceted; consequently, we would like to automatically identify contexts that summarize the key information of each sample rather than pre-define molecular subtypes.

Discriminative Subtyping of Lung Cancers Lung cancer is common, accounting for 13% of all new cancer cases and over 20% of all cancer deaths [SMJ20]. Lung adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC) are two of the most common subtypes [CTM11] and have important distinctions in clinicopathology [RTGA19], immune subtypes [SKSK18], and therapeutic effectiveness [LCB+13], so distinguishing LUAD from LUSC is an important clinical task.

However, there may not be a single transcriptomic signature which universally distinguishes LUAD/LUSC. Instead, we would like to allow different cellular processes to predict subtype for different samples. To do this,

Level	Model	Data	Accuracy (%)	Macro F1
Patch	**CEN**	**H+T**	**96.18**	**96.97**
	Concatenated	H+T	95.32	93.65
	Ensemble	H+T	94.61	90.23
	Logistic Regression	T	94.05	91.40
	`InceptionV3`	H	69.14	65.85
Sample	**CEN**	**H+T**	**94.51**	**95.29**
	Concatenated	H+T	93.33	93.65
	Ensemble	H+T	92.94	90.23
	Logistic Regression	T	92.16	89.67
	`InceptionV3`	H	80.00	76.76

Table 1: Performance of models on 3-way Normal/LUAD/LUSC classification. We report accuracy on both the "patch" level (where each prediction corresponds to a small patch in the image) and on the "sample" level (where a single prediction is made for the whole-slide image). Some models are designed to operate on transcriptomic data (T), while some operate on histopathology data (H), while others use both forms of data (H+T). CEN (our model) outperforms all existing models.

we use a deep neural network to summarize histopathology "contexts" into discriminative subtype of each sample, and then use this latent discriminative subtype to generate sample-specific coefficients for a logistic regression model which predicts LUAD/LUSC from gene expression. This approach summarizes biological processes as interpretable coefficients of genes, and connects these biological processes to histopathology morphology through latent discriminative subtypes.

As a result, discriminative subtypes go beyond previous studies which have used morphology to estimate the likelihood of individual mutations [ADME18] or gene expression [FJT+20, ZFZM20] (more complete related work is described in Section A); instead, discriminative subtypes are a data-driven approach to learn gene expression patterns that are meaningful in morphological contexts and to summarize biological patterns on a process level. We investigate the utility of this approach to predict subtypes from the histopathology images and transcriptomic data of patients diagnosed with LUAD or LUSC. This approach not only achieves a new state-of-the-art accuracy for LUAD/LUSC classification, but also provides parsimonious descriptions that link morphology and transcriptomics.

Results

Our methodology (detailed in Materials and Methods) uses histopathology images to contextualize transcriptomic models by predicting the discriminative subtype of each sample. This procedure not only improves the accuracy of predictions made by the transcriptomic models, but also provides transcriptomic "archetypes" which can be summarized as biological processes and traced to morphological patterns. This goes beyond previous studies which have used morphology to estimate the likelihood of individual mutations [ADME18] or gene expression [FJT+20, ZFZM20]; here, we allow the model to learn which gene expression patterns are meaningful in each morphological context, allowing us to summarize the biological patterns and tie them to morphology on a process level rather than a mutation level. The goal of this approach is not only to enhance the classification performance, but also to infer meaningful clusters. We first measure predictive accuracy and then we seek to assign biological meaning to the archetypes.

Contextualization Improves Prediction of Subtypes

First, we measure the ability of the model to classify LUAD, LUSC, and healthy samples. Our model (CEN) achieves the best classification results, outperforming all the multi-modal and the uni-modal baselines to achieve a new state-of-the-art performance for classifying LUAD/LUSC/Healthy samples (Tables 1, C.1). The `InceptionV3` deep neural network was used by Coudray et al [COS+18] to achieve state-of-the-art

Arch.	Term ID	Term Name	P-Val
1	KEGG:04071	Sphingolipid signaling pathway	0.037
	KEGG:04310	Wnt signaling pathway	0.049
6	REAC:R-HSA-6802952	Signaling by BRAF and RAF fusions	0.039
8	TF:M06732	Factor: ZNF304	0.023
12	REAC:R-HSA-8939236	RUNX1	0.022
13	TF:M09657_1	Factor: Smad4	0.004
15	GO:0071385	Cellular response to glucocorticoid stimulus	0.013
17	REAC:R-HSA-400206	Regulation of lipid metabolism by PPARα	0.018
18	TF:M04726_1	Factor: REST	0.001
19	TF:M05327_1	Factor: WT1	0.025
21	TF:M01224_1	Factor: P50:RELA-P65	0.034
25	TF:M09611_0	Factor: ER81	0.003
26	KEGG:04215	Apoptosis	0.006
28	GO:0051240	Pos. reg. of multicellular organismal process	0.006
30	REAC:R-HSA-4791275	Signaling by WNT in cancer	0.045

Table 2: Archetypal transcriptomic models with terms significantly enriched ($p < 0.05$).

classification of LUAD/LUSC samples from histopathology images, and our use of this model reproduces the performance reported therein. While this performance is state-of-the-art for histopathology data, a regularized logistic regression model operating on transcriptomic data significantly outperforms the deep neural network, indicating that the transcriptomic assay is a very good source of information regarding tumor subtype. By using both the transcriptomic and histopathology data, the multi-modal models outperform either single modality, and the CEN achieves the highest accuracy of all.

Archetypal Models Correspond to Distinct Biological Processes

The CEN model estimates 32 transcriptomic model archetypes which are used to predict LUAD/LUSC/Healthy labels in various discriminative subtypes. We seek to interpret these archetypes to understand the underlying biological processes. First, we ask whether the model archetypes are redundant or look at different biological processes. Surprisingly, the coefficient vectors of the archetypes are nearly orthogonal (Fig. C.10), indicating that archetypes emphasizes distinct processes.

Next, we use enrichment analysis [STM+05] to identify biological processes which are statistically over-represented in the archetype coefficients. For each archetype, we have 3 vectors of coefficients defining the 3-way (LUAD/LUSC/Healthy) logistic regression model. For each vector of coefficients, we sort the genes by the magnitude of the associated coefficient (top 5 genes for each model archetype for normal, LUAD, and LUSC labels are provided in the Supplement). We search the top 100 genes in this list for enriched biological terms against the background of the remaining 1595 genes using `gProfiler` [RAA+16]. We set a minimum intersection size of 3 genes (the number of genes selected from each term must be at least 3) and a maximum Bonferroni-corrected p-value of 0.05. Terms enriched in the archetypal models for prediction of the "Normal" label (which corresponds to case/control prediction) are shown in Table 2. Similar tables for the LUAD and LUSC models are available in the Supplement. Many of the 32 archetypal models are significantly enriched for markers of cancer, indicating that not only do the discriminative subtypes correspond to distinct *statistical* patterns but also distinct *oncologic* processes.

As a vignette, below we discuss the terms in Table 2 which have a p-value of less than 0.01:

- **13: Factor: Smad4.** SMAD family member 4 (SMAD4) is a necessary component of the transforming growth factor beta (TGFβ) pathway and in this capacity regulates cellular proliferation [YDW+97, ZMD18, ZBZ+98]. SMAD4 has been established as a tumor suppressor in pancreatic, colon and lung cancer [HTK+16, MW12, ZMD18]. Decreased expression of SMAD4 has been identified as a marker of Non-Small Cell Lung Cancer and poor prognosis in LUAD [IBH+12, HTK+16, CSJ+17, MW12].

- **18: Factor: REST.** RE1-silencing transcription factor (REST) is most widely known to repress neuronal genes in non-neuronal cells, but has also been found to be an oncogene or tumor suppressor gene in certain cancers [NPDM13]. In breast cancer, it has been shown to act as a tumor suppressor, and loss of this gene is associated with aggressive breast cancer [WGS+10], while an isoform of REST has been indicated as a specific clinical marker for early detection of small cell lung cancers [CEWQ00, SSI+13].

- **26: Apoptosis.** A hallmark of cancer cells is the ability to limit or avoid cell death induced by apoptosis [HW11]. In lung cancer, multiple genes in the KEGG apoptosis pathway have been found to be correlated with tumorigenesis, chemoresistance, and poor prognosis [LPY+17, PHK13, yYKS+04, DLL+13, TTY+01, JEL+01, JEL+01, Fen05, GED+04]. For example, altered expression of caspases involved in the crucial final steps of apoptosis have been shown to make cells resistant to apoptosis, enabling resistance to chemotherapy [Fen05, PCS+12, JEL+01, TTY+01, yYKS+04].

- **28: Positive Regulation of Multicellular Organismal Process.** Oncogenes activate tumor growth and proliferation processes by up-regulating other development processes, including angiogenesis, cell migration, and metastasis. For example, aggressive cancers achieve metastasis by up-regulating inflammation and cell migration [HW00, HW11].

Archetypal Models Correspond to Distinct Morphologic Patterns

One of the advantages of CEN is that it enables us to tie the morphologic patterns recognized in histopathology images into the biological processes of the transcriptomic models. Towards this end, we visualize representative patches which maximize the predicted influence of archetypal models 12, 13, and 30 in Figure 2 (similar results for the other archetypal models are shown in the Supplement). We can see that the morphology is more similar within the clusters corresponding to archetypes than between the clusters (e.g. archetypal model 12, corresponding to the RUNX1 factor, seems to focus on morphological features of hematopoietic processes and macrophage response, while archetypal model 13, corresponding to the SMAD4 intercellular signaling complex seems to focus on morphological features of irregular multicellular structure) indicating that the transcriptomic patterns used for accurate downstream predictions correspond with morphological changes.

Discriminative Subtypes Display Intratumor Heterogeneity

The spatial diversity of cells within lung tumors has recently been shown to encode prognostic markers [ARR+20]; the CEN model enables us to ask questions regarding the localization of discriminative subtypes. By making predictions for each patch independently, the CEN model allows us to ask whether different locations in the image are indicative of distinct tumor subtypes and if morphological patterns of a particular subtype tend to be clustered in a single location. Firstly, we see that on held-out test data, in 46.3% of patches, the CEN assigns highest weight to a discriminative subtype which is not the discriminative subtype chosen for the full slide by a majority vote of the patches. This indicates that there may either be significant diversity or uncertainty of morphological patterns or tumor subtypes in each tumor. To examine the spatial locality, we visualize the predictions of discriminative subtypes. Discriminative subtypes exhibit spatial locality (Figures 3 and C.1), with nearby patches tending to correspond to the same predicted subtype.

Discussion

This work demonstrates the utility of estimating latent *discriminative subtypes* of lung cancers. We have defined discriminative subtypes as latent variables which provide parsimonious descriptions of a tumor, and these subtypes provide context which optimizes the ability of a downstream model to discriminate

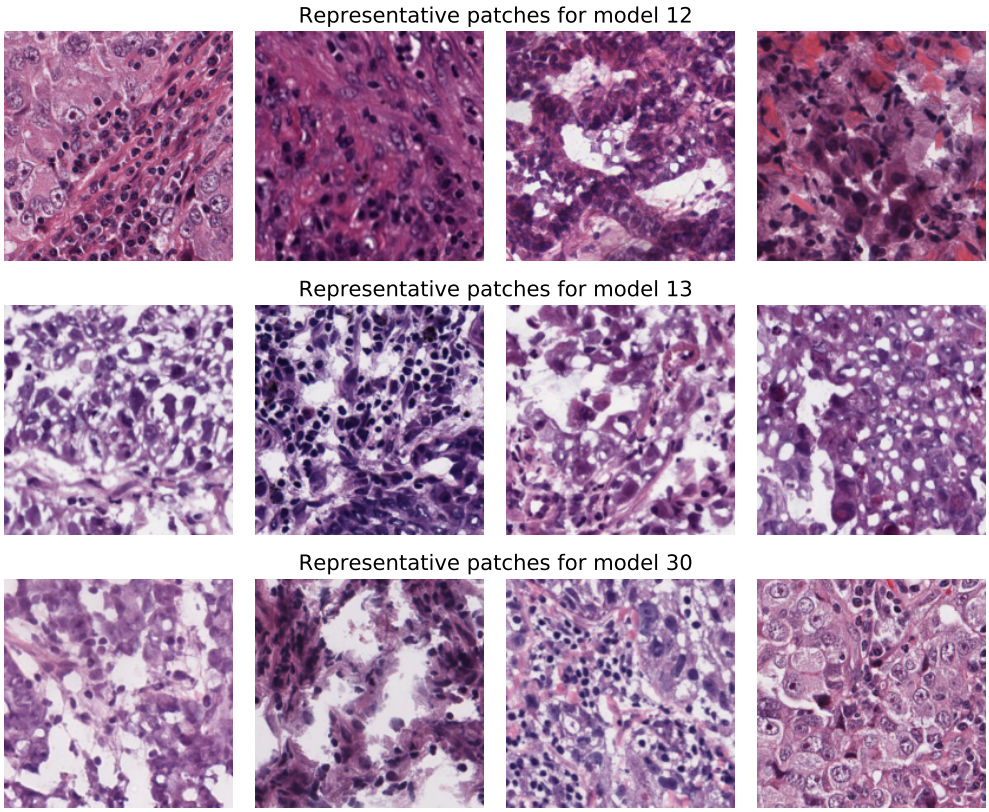

Figure 2: Morphology is homogeneous within each archetype and varies between archetypes. Each row contains the set of patches which made CEN assign the highest weights to the corresponding archetypal model over the entire training dataset. These patches represent archetypal morphological patterns for the archetypal transcriptomic models and link transcriptomic attention with morphology.

between subtypes of tumors. Contextualizing predictive models according to these discriminative subtypes achieves state-of-the-art classification accuracy for LUAD/LUSC/healthy prediction, suggesting that different transcriptomic patterns are most predictive in different contexts.

For machine learning practitioners, this has several implications. Firstly, multi-modal data analysis is critically important in healthcare (often because the technical noise in each measurement can most effectively be reduced by having multiple views of each sample), but multi-modal pipelines are difficult to meld with interpretability. Contextual learning is a promising framework for building interpretable multi-modal approaches. By allowing one data modality to select the important features for another modality, we can achieve both accuracy and interpretability.

This also has biomedical implications. Firstly, because the discriminative subtypes correspond to biological processes, we can surmise that these are distinctive signatures of cancers that may be used to improve personalized medicine beyond known histologic or molecular subtypes. Furthermore, because a deep learning model trained on hisotopathology images can assign meaningful contexts for transcriptomic models, some patterns of transcriptomic aberrations are contained in histopathology data with only basic hematoxylin and eosin (H&E) staining. This concords with several recent works to predict genetic variants from H&E stained histopathology data [PBC+16, ADME18, FJT+20], and encourages future works which discover and analyze morphological patterns of cellular modifications. All in all, this research demonstrates the utility of jointly modeling histopathology and transcriptomic data to identify latent cancer subtypes, and underscores the potential for histopathology images to be used for discovery of fine-grained subtypes.

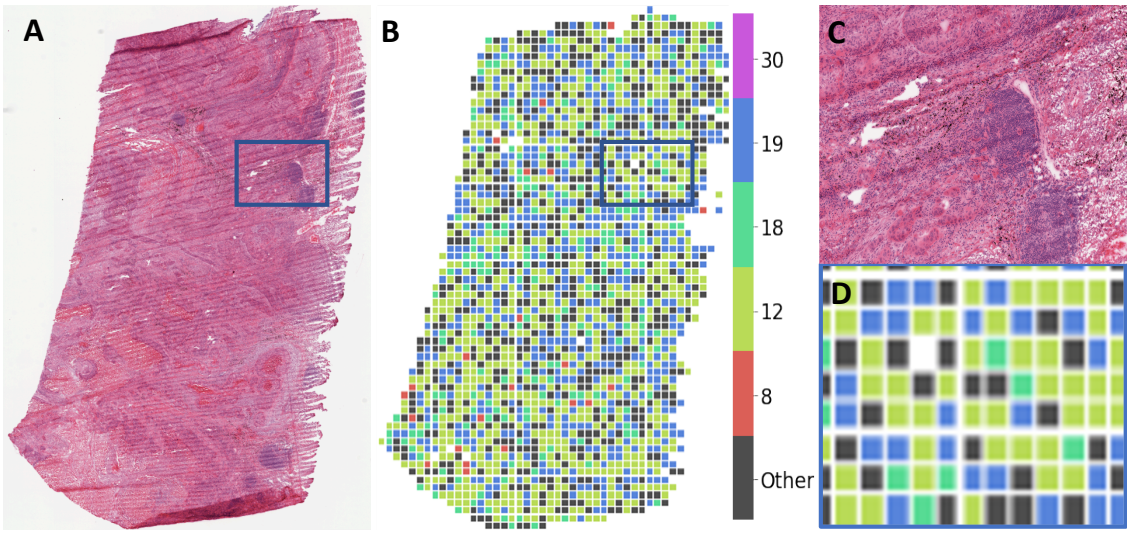

Figure 3: Discriminative subtypes display significant heterogeneity throughout the sample, but cluster spatially. Representative results of patch-level subtype prediction for a LUSC sample held out from CEN training. **(A)** tumor whole-slide image. **(B)** archetypes assigned to each patch by CEN; for visual clarity, we display only the common archetypes 8, 12, 18, 19, and 30, and group all other archetypes under "other". **(C, D)** inset squares outlined in blue in panes A and B.

Materials and Methods

Model Architecture

To combine multi-modal data in an interpretable pipeline, we use a version of Contextual Explanation Networks (CENs) [ASDX20] (Figure 4). The CEN architecture takes data represented by: (i) context variables C, (ii) interpretable variables X, and (iii) target variables Y. In this paper, we use histopathology images as contextual data and gene expression assays as interpretable predictors of the cancer type. The model represents conditional probability of the cancer type given the histopathology and transcriptomic inputs, $\mathbb{P}(Y \mid X, C)$, in the following form:

$$\mathbb{P}(Y \mid X, C) = \int \mathbb{P}(Y \mid X, \theta)\delta(\theta = \phi_w(C))d\theta = \mathbb{P}(Y \mid X, \phi_w(C)), \tag{1}$$

where $\mathbb{P}(Y \mid X, \theta)$ is a linear logistic model that predicts cancer types from gene expression. Note that parameters (or weights) θ of the logistic model are a function of the contextual information, i.e., $\theta = \phi_w(C)$. In other words, CEN produces a sample-specific parameterization of a linear probabilistic model that operates on transcriptomic data based on histopathology.

Generation of parameters for sample-specific linear models is accomplished via a deep context encoder. Aligning with the hypothesis of discriminative subyptes, parameters θ are confined to be a linear combination of a small constant number (K) of "archetypes," denoted $\{\theta_1, \ldots, \theta_K\}$ or $\theta_{1:K}$. For a sample i with context c_i, weights θ_i are:

$$\theta_i = \sum_{k=1}^{K} \alpha_w^k(c_i)\theta_k, \qquad \text{where } \alpha_w^k(c_i) \geq 0,\ k = 1, \ldots, K \text{ and } \sum_{k=1}^{K} \alpha_w^k(c_i) = 1, \tag{2a}$$

where $(\alpha^1, \ldots, \alpha^K)$ is a vector output of the context encoder, which we call the *discriminative subtype*. This dictionary of parameter sets, $\theta_{1:K}$, is estimated jointly with the context encoder, and the entire architecture is trained end-to-end via backpropagation after initialization of the archetypes to random vectors. To summarize, the context encoder is a deep neural network that processes contextual data (histopathology imagery) and

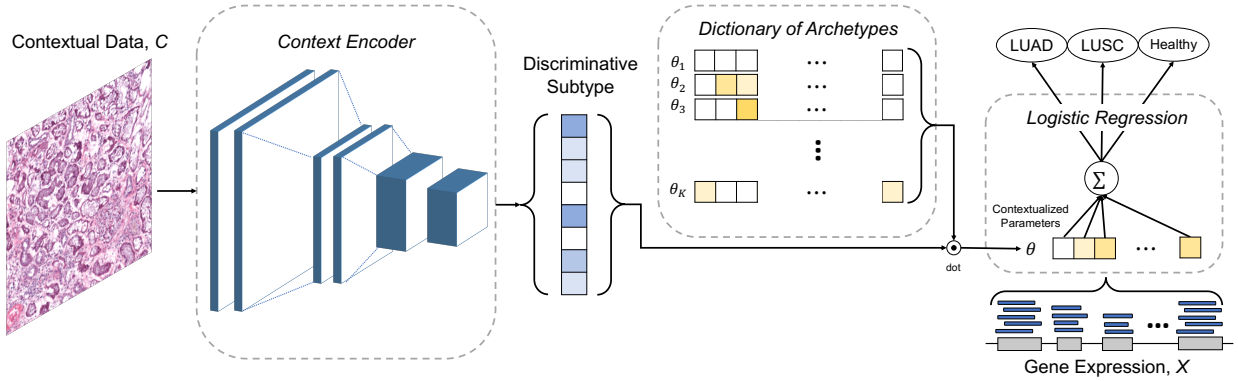

Figure 4: The CEN architecture estimates latent discriminative subtypes by combining "contextual" histopathology images (C) with gene expression (X) to predict sample classifications. The discriminative subtype is predicted from contextual data and forms sample-specific parameters θ by weighting archetypal models stored in a dictionary. The sample-specific logistic regression models estimate classifications from gene expression (X) with sample-specific coefficients, and this classification loss is back-propagated to update both the dictionary of archetypes and the context encoder.

outputs a probability vector of length K that softly selects weights for a linear probabilistic model from a dictionary of archetypes. Architectural decisions, hyperparameters, and baseline methods are available in Section B. `Python` code is available at: `https://github.com/slabbaki/discriminative-subtyping`.

Cohort

We use lung cancer data available in the NCI Genomic Data Commons [GHF+16], which includes resources from both The Cancer Genome Atlas and The Cancer Imaging Atlas to provide multi-modal descriptions of a large number of cancer patients. As of April 2, 2020, this depository included 585 LUAD cases and 504 LUSC cases, along with a variety of cases with cancer of other tissues. From this set, we selected all samples which have both eosin-stained histopathology whole-slide images and transcriptomic RNA-seq profiling. This selects 992 patients, with 1 or 2 whole-slide images for each patient. We split this dataset by patient into training, validation, and testing partitions, so the test set consists of completely held-out patients (size and composition of partitions in Table B.1).

Image Subsampling. We use 20x magnification whole-slide histopathology images. To make these large images suitable for the pre-trained `Inceptionv3` architecture, we split each whole-slide image into non-overlapping 299x299 image patches. The patches were controlled for quality by discarding any patches with more than 25% white pixels, following the procedure of [COS+18]. We treat each patch as a separate sample, with transcriptomic data duplicated over each patch.

Transcriptomic Profiles. The transcriptomic profiling in TCGA captures the expression of over 60,000 distinct transcripts in each sample. We select the 1000 transcripts with the highest variance in the non-lung cancer cases in TCGA. In addition, we augment this set with 695 transcripts corresponding to genes in the Catalogue of Somatic Mutations in Cancer (COSMIC, [BDF+04]).

References

[ADME18] Jordan Ash, Gregory Darnell, Daniel Munro, and Barbara Engelhardt. Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology. *bioRxiv*, 2018.

[ARR+20] Khalid AbdulJabbar, Shan E Ahmed Raza, Rachel Rosenthal, Mariam Jamal-Hanjani, Selvaraju Veeriah, Ayse Akarca, Tom Lund, David A Moore, Roberto Salgado, Maise Al Bakir, et al. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nature Medicine*, pages 1–9, 2020.

[ASDX20] Maruan Al-Shedivat, Avinava Dubey, and Eric P Xing. Contextual explanation networks. *J. Mach. Learn. Res.*, 21:194–1, 2020.

[AYHvdS16] Ahmed M Alaa, Jinsung Yoon, Scott Hu, and Mihaela van der Schaar. Personalized risk scoring for critical care patients using mixtures of gaussian process experts. *ICML 2016 Workshop on Computational Frameworks for Personalization*, 2016.

[BDF+04] Sally Bamford, Emily Dawson, Simon Forbes, Jody Clements, Roger Pettett, Ahmet Dogan, A Flanagan, Jon Teague, P Andrew Futreal, Michael R Stratton, et al. The cosmic (catalogue of somatic mutations in cancer) database and website. *British journal of cancer*, 91(2):355–358, 2004.

[BKH+02] D. G. Beer, S. L. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*, 8(8):816–24, 2002.

[BNC+15] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155, 2015.

[BWSH+11] D R Baldwin, B White, M Schmidt-Hansen, A R Champion, and A M Melder. Diagnosis and treatment of lung cancer: summary of updated NICE guidance. *BMJ*, 342, 2011.

[CBCB19] Eric A Collisson, Peter Bailey, David K Chang, and Andrew V Biankin. Molecular subtypes of pancreatic cancer. *Nature Reviews Gastroenterology & Hepatology*, 16(4):207–220, 2019.

[CCL+08] Simon J Crabb, Maggie CU Cheang, Samuel Leung, Taina Immonen, Torsten O Nielsen, David D Huntsman, Chris D Bajdik, and Stephen K Chia. Basal breast cancer molecular subtype predicts for lower incidence of axillary lymph node metastases in primary breast cancer. *Clinical breast cancer*, 8(3):249–256, 2008.

[CEWQ00] Judy M Coulson, Jodie L Edgson, Penella J Woll, and John P Quinn. A splice variant of the neuron-restrictive silencer factor repressor is expressed in small cell lung cancer: a potential role in derepression of neuroendocrine genes and a useful clinical marker. *Cancer research*, 60(7):1840–1844, 2000.

[CGAR12] Network Cancer Genome Atlas Research. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–25, 2012.

[CHG+19] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

[COS+18] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyo, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, 2018.

[CSJ+17] Jean Chiou, Chia-Yi Su, Yi-Hua Jan, Chih-Jen Yang, Ming-Shyan Huang, Yung-Luen Yu, and Michael Hsiao. Decrease of FSTL1-BMP4-Smad signaling predicts poor prognosis in lung adenocarcinoma but not in squamous cell carcinoma. *Scientific reports*, 7(1):9830, aug 2017.

[CTM11] Charles S Dela Cruz, Lynn T Tanoue, and Richard A Matthay. Lung cancer: epidemiology, etiology, and prevention. *Clinics in chest medicine*, 32(4):605–644, 2011.

[CZH+17] Jun Cheng, Jie Zhang, Yatong Han, Xusheng Wang, Xiufen Ye, Yuebo Meng, Anil Parwani, Zhi Han, Qianjin Feng, and Kun Huang. Integrative Analysis of Histopathological Images and Genomic Data Predicts Clear Cell Renal Cell Carcinoma Prognosis. *Cancer Research*, 77(21):e91—-e100, 2017.

[Daw17] Mark A Dawson. The cancer epigenome: Concepts, challenges, and therapeutic opportunities. *Science*, 355(6330):1147–1152, 2017.

[DLL+13] Xin Dong, Dong Lin, Chris Low, Emily A. Vucic, John C. English, John Yee, Nevin Murray, Wan L. Lam, Victor Ling, Stephen Lam, Peter W. Gout, and Yuzhuo Wang. Elevated expression of birc6 protein in non-small-cell lung cancers is associated with cancer recurrence and chemoresistance. *Journal of Thoracic Oncology*, 8(2):161–170, 2013.

[Fen05] Dean A Fennell. Caspase Regulation in Non{\textendash}Small Cell Lung Cancer and its Potential for Therapeutic Exploitation. *Clinical Cancer Research*, 11(6):2097–2105, 2005.

[FJT+20] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, 1(8):800–810, 2020.

[FPS13] R Fisher, L Pusztai, and C Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479, 2013.

[GED+04] A M Groeger, V Esposito, A De Luca, R Cassandro, G Tonini, V Ambrogi, F Baldi, R Goldfarb, T C Mineo, A Baldi, and E Wolner. Prognostic value of immunohistochemical expression of p53, bax, Bcl-2 and Bcl-xL in resected non-small-cell lung cancers. *Histopathology*, 44(1):54–63, 2004.

[GHF+16] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

[GSCM+19] Arkadiusz Gertych, Zaneta Swiderska-Chadaj, Zhaoxuan Ma, Nathan Ing, Tomasz Markiewicz, Szczepan Cierniak, Hootan Salemi, Samuel Guzman, Ann E Walts, and Beatrice S Knudsen. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Scientific Reports*, 9(1):1483, 2019.

[GWC+13] A Goldhirsch, E P Winer, A S Coates, R D Gelber, M Piccart-Gebhart, B Thürlimann, H-J Senn, and Panel Members. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Annals of oncology : official journal of the European Society for Medical Oncology*, 24(9):2206–2223, sep 2013.

[HKT+19] Jie Hao, Sai Chandra Kosaraju, Nelson Zange Tsaku, Dae Hyun Song, and Mingon Kang. PAGE-Net: Interpretable and Integrative Deep Learning for Survival Analysis Using Histopathological Images and Genomic Data. In *Biocomputing 2020*, pages 355–366. WORLD SCIENTIFIC, nov 2019.

[HPL14] Josie Hayes, Pier Paolo Peruzzi, and Sean Lawler. MicroRNAs in cancer: Biomarkers, functions and therapy, 2014.

[HTK+16] S M Haeger, J J Thompson, S Kalra, T G Cleaver, D Merrick, X-J Wang, and S P Malkoski. Smad4 loss promotes lung cancer formation but increases sensitivity to DNA topoisomerase inhibitors. *Oncogene*, 35(5):577–586, 2016.

[HW00] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[HW11] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation, mar 2011.

[IBH+12] Marcin Imielinski, Alice H. Berger, Peter S. Hammerman, Bryan Hernandez, Trevor J. Pugh, Eran Hodis, Jeonghee Cho, James Suh, Marzia Capelletti, Andrey Sivachenko, Carrie Sougnez, Daniel Auclair, Michael S. Lawrence, Petar Stojanov, Kristian Cibulskis, Kyusam Choi, Luc De Waal, Tanaz Sharifnia, Angela Brooks, Heidi Greulich, Shantanu Banerji, Thomas Zander, Danila Seidel, Frauke Leenders, Sascha Ansén, Corinna Ludwig, Walburga Engel-Riedel, Erich Stoelben, Jürgen Wolf, Chandra Goparju, Kristin Thompson, Wendy Winckler, David Kwiatkowski, Bruce E. Johnson, Pasi A. Jänne, Vincent A. Miller, William Pao, William D. Travis, Harvey I. Pass, Stacey B. Gabriel, Eric S. Lander, Roman K. Thomas, Levi A. Garraway, Gad Getz, and Matthew Meyerson. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6):1107–1120, sep 2012.

[IKK+20] Osamu Iizuka, Fahdi Kanavati, Kei Kato, Michael Rambeau, Koji Arihiro, and Masayuki Tsuneki. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Scientific Reports*, 10(1):1504, 2020.

[Ina17] Kentaro Inamura. Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification. *Frontiers in Oncology*, 7:193, 2017.

[JEL+01] Bertrand Joseph, Jessica Ekedahl, Rolf Lewensohn, Philippe Marchetti, Pierre Formstecher, and Boris Zhivotovsky. Defective caspase-3 relocalization in non-small cell lung carcinoma. *Oncogene*, 20(23):2877–2888, 2001.

[KI18] Daisuke Komura and Shumpei Ishikawa. Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018.

[KMK+13] Hyun Seok Kim, Saurabh Mendiratta, Jiyeon Kim, Chad Victor Pecot, Jill E Larsen, Iryna Zubovych, Bo Yeun Seo, Jimi Kim, Banu Eskiocak, Hannah Chung, et al. Systematic identification of molecular subtype-selective vulnerabilities in non-small-cell lung cancer. *Cell*, 155(3):552–566, 2013.

[KTY+15] Marieke Lydia Kuijjer, Matthew Tung, GuoCheng Yuan, John Quackenbush, and Kimberly Glass. Estimating sample-specific regulatory networks. *arXiv preprint arXiv:1505.06440*, 2015.

[LCB+13] Neal I Lindeman, Philip T Cagle, Mary Beth Beasley, Dhananjay Arun Chitale, Sanja Dacic, Giuseppe Giaccone, Robert Brian Jenkins, David J Kwiatkowski, Juan-Sebastian Saldivar, Jeremy Squire, Erik Thunnissen, and Marc Ladanyi. Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors: guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, 8(7):823–859, jul 2013.

[LdPM+19] Marta Lucchetta, Isabelle da Piedade, Mohamed Mounir, Marina Vabistsevits, Thilde Terkelsen, and Elena Papaleo. Distinct signatures of lung cancer types: aberrant mucin O-glycosylation and compromised immune response. *BMC Cancer*, 19(1):824, 2019.

[LMP17] Daniel Lopez-Martinez and Rosalind Picard. Multi-task neural networks for personalized pain recognition from physiological signals. *arXiv preprint arXiv:1708.08755*, 2017.

[LPY+17] Guangbo Liu, Fen Pei, Fengqing Yang, Lingxiao Li, Amit Dipak Amin, Songnian Liu, J Ross Buchan, and William C Cho. Role of Autophagy and Apoptosis in Non-Small-Cell Lung Cancer. *International journal of molecular sciences*, 18(2):367, feb 2017.

[LWJ+16] Xiaoping Liu, Yuetong Wang, Hongbin Ji, Kazuyuki Aihara, and Luonan Chen. Personalized characterization of diseases using sample-specific networks. *Nucleic acids research*, 44(22):e164–e164, 2016.

[LZY+17] Xin Luo, Xiao Zang, Lin Yang, Junzhou Huang, Faming Liang, Jaime Rodriguez-Canales, Ignacio I Wistuba, Adi Gazdar, Yang Xie, and Guanghua Xiao. Comprehensive Computational Pathological Image Analysis Predicts Lung Cancer Prognosis. *Journal of Thoracic Oncology*, 12(3):501–509, 2017.

[MAK+07] Hojin Moon, Hongshik Ahn, Ralph L Kodell, Songjoon Baek, Chien-Ju Lin, and James J Chen. Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artificial intelligence in medicine*, 41(3):197–207, 2007.

[MBW+13] Hiroko Masuda, Keith A Baggerly, Ying Wang, Ya Zhang, Ana Maria Gonzalez-Angulo, Funda Meric-Bernstam, Vicente Valero, Brian D Lehmann, Jennifer A Pietenpol, Gabriel N Hortobagyi, et al. Differential response to neoadjuvant chemotherapy among 7 triple-negative breast cancer molecular subtypes. *Clinical cancer research*, 19(19):5533–5540, 2013.

[MOP+18] Siyuan Ma, Shuji Ogino, Princy Parsana, Reiko Nishihara, Zhirong Qian, Jeanne Shen, Kosuke Mima, Yohei Masugi, Yin Cao, and Jonathan A Nowak. Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. *Genome biology*, 19(1):142, 2018.

[MSC+19] Hassan Muhammad, Carlie S. Sigel, Gabriele Campanella, Thomas Boerner, Linda M. Pak, Stefan Büttner, Jan N. M. IJzermans, Bas Groot Koerkamp, Michael Doukas, William R. Jarnagin, Amber Simpson, and Thomas J. Fuchs. Towards unsupervised cancer subtyping: Predicting prognosis using a histologic visual dictionary, 2019.

[MW12] Stephen P Malkoski and Xiao-Jing Wang. Two sides of the story? Smad4 loss in pancreatic cancer versus head-and-neck cancer. *FEBS letters*, 586(14):1984–1992, jul 2012.

[MYA+18] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee A D Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970—-E2979, 2018.

[NPDM13] Sara Negrini, Ilaria Prada, Rosalba D'Alessandro, and Jacopo Meldolesi. Rest: an oncogene or a tumor suppressor? *Trends in cell biology*, 23(6):289–295, 2013.

[PBC+16] Vlad Popovici, Eva Budinska, Lenka Capkova, Daniel Schwarz, Ladislav Dusek, Josef Feit, and Rolf Jaggi. Joint analysis of histopathology image features and gene expression in breast cancer. *BMC Bioinformatics*, 17:1–9, 2016.

[PCS+12] I Paul, A D Chacko, I Stasik, S Busacca, N Crawford, F McCoy, N McTavish, B Wilson, M Barr, K J O'Byrne, D B Longley, and D A Fennell. Acquired differential regulation of caspase-8 in cisplatin-resistant non-small-cell lung cancer. *Cell Death & Disease*, 3(12):e449–e449, 2012.

[PHK13] Milind M. Pore, T. J. Hiltermann, and Frank A. E. Kruyt. Targeting apoptosis pathways in lung cancer. *Cancer letters*, 332(2):359–368, May 28 2013. Copyright - Copyright Elsevier Limited May 28, 2013; Last updated - 2015-04-10.

[PSE+00] Charles M Perou, Therese Sørlie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, et al. Molecular portraits of human breast tumours. *nature*, 406(6797):747–752, 2000.

[RAA+16] Jüri Reimand, Tambet Arak, Priit Adler, Liis Kolberg, Sulev Reisberg, Hedi Peterson, and Jaak Vilo. g: Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, 44(W1):W83–W89, 2016.

[RPB+19] Charles M Rudin, John T Poirier, Lauren Averett Byers, Caroline Dive, Afshin Dowlati, Julie George, John V Heymach, Jane E Johnson, Jonathan M Lehman, David MacPherson, et al. Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nature Reviews Cancer*, 19(5):289–297, 2019.

[RR17] Martin Reck and Klaus F Rabe. Precision Diagnosis and Treatment for Advanced Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, 377(9):849–861, aug 2017.

[RTGA19] Valeria Relli, Marco Trerotola, Emanuela Guerra, and Saverio Alberti. Abandoning the Notion of Non-Small Cell Lung Cancer. *Trends in Molecular Medicine*, 25(7):585–594, 2019.

[RVG+11] Sara Ricardo, André Filipe Vieira, Renê Gerhard, Dina Leitão, Regina Pinto, Jorge F Cameselle-Teijeiro, Fernanda Milanezi, Fernando Schmitt, and Joana Paredes. Breast cancer stem cell markers cd44, cd24 and aldh1: expression distribution within intrinsic molecular subtype. *Journal of clinical pathology*, 64(11):937–946, 2011.

[Sch10] Stuart J Schnitt. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Modern Pathology*, 23(2):S60–S64, 2010.

[SCM19] Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. Deep neural network models for computational histopathology: A survey, 2019.

[SGLGK22] Artem Shmatko, Narmin Ghaffari Laleh, Moritz Gerstung, and Jakob Nikolas Kather. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nature Cancer*, 3(9):1026–1038, Sep 2022.

[SIVA17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[SKSK18] Jeong-Sun Seo, Ahreum Kim, Jong-Yeon Shin, and Young Tae Kim. Comprehensive analysis of the tumor immune micro-environment in non-small cell lung cancer for efficacy of checkpoint inhibitor. *Scientific Reports*, 8(1):14576, 2018.

[SMJ20] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(1):7–30, 2020.

[SPT+01] Therese Sørlie, Charles M Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001.

[SSI+13] Masahito Shimojo, Yoshie Shudo, Masatoshi Ikeda, Tomoyo Kobashi, and Seiji Ito. The small cell lung cancer-specific isoform of re1-silencing transcription factor (rest) is regulated by neural-specific ser/arg repeat-related protein of 100 kda (nsr100). *Molecular Cancer Research*, 11(10):1258–1268, 2013.

[SSZF19] Wengong Si, Jiaying Shen, Huilin Zheng, and Weimin Fan. The role and mechanisms of action of microRNAs in cancer drug resistance. *Clinical Epigenetics*, 11(1):25, 2019.

[STM+05] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[TBB+15] William D Travis, Elisabeth Brambilla, Allen Burke, Alexander Marx, and Andrew G Nicholson. *WHO classification of tumours of the lung, pleura, thymus and heart*. International Agency for Research on Cancer, 2015.

[TTY+01] Tetsuya Takata, Fumihiro Tanaka, Tomoko Yamada, Kazuhiro Yanagihara, Yosuke Otake, Yozo Kawano, Tatsuo Nakagawa, Ryo Miyahara, Hiroki Oyanagi, Kenji Inui, and Hiromi Wada. Clinical significance of caspase-3 expression in pathologic-stage I, nonsmall-cell lung cancer. *International Journal of Cancer*, 96(S1):54–60, jan 2001.

[vdVHvV+02] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25):1999–2009, 2002.

[VKVK19] Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. Maximum likelihood estimation for learning populations of parameters. In *International Conference on Machine Learning*, pages 6448–6457, 2019.

[VSS+17] Michel E Vandenberghe, Marietta L J Scott, Paul W Scorer, Magnus Söderberg, Denis Balcerzak, and Craig Barker. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Scientific Reports*, 7(1):45938, 2017.

[WCY+18] Shidan Wang, Alyssa Chen, Lin Yang, Ling Cai, Yang Xie, Junya Fujimoto, Adi Gazdar, and Guanghua Xiao. Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Scientific Reports*, 8(1):10393, 2018.

[WGS+10] Matthew P Wagoner, Kearney TW Gunsalus, Barry Schoenike, Andrea L Richardson, Andreas Friedl, and Avtar Roopra. The transcription factor rest is lost in aggressive breast cancer. *PLoS genetics*, 6(6), 2010.

[WTL+19] Jason W Wei, Laura J Tafe, Yevgeniy A Linnik, Louis J Vaickus, Naofumi Tomita, and Saeed Hassanpour. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific Reports*, 9(1):3358, 2019.

[WYH+10] M. D. Wilkerson, X. Yin, K. A. Hoadley, Y. Liu, M. C. Hayward, C. R. Cabanski, K. Muldrew, C. R. Miller, S. H. Randell, M. A. Socinski, A. M. Parsons, W. K. Funkhouser, C. B. Lee, P. J. Roberts, L. Thorne, P. S. Bernard, C. M. Perou, and D. N. Hayes. Lung squamous cell carcinoma mrna expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin Cancer Res*, 16(19):4864–75, 2010.

[YDW+97] J M Yingling, M B Datto, C Wong, J P Frederick, N T Liberati, and X F Wang. Tumor suppressor Smad4 is a transforming growth factor beta-inducible DNA binding protein. *Molecular and Cellular Biology*, 17(12):7019–7028, 1997.

[yYKS+04] Jin young Yoo, Chi Hong Kim, So Hyang Song, Byoung Yong Shim, Youn Ju Jeong, Meyung Im Ahn, Suji Kim, Deog Gon Cho, Min Seop Jo, Kyu Do Cho, Hong Joo Cho, Seok Jin Kang, and Hoon Kyo Kim. Expression of caspase-3 and c-myc in non-small cell lung cancer. *Cancer research and treatment : official journal of Korean Cancer Association*, 36(5):303–307, oct 2004.

[YZB+16] Kun-Hsing Yu, Ce Zhang, Gerald J Berry, Russ B Altman, Christopher Ré, Daniel L Rubin, and Michael Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7(1):12474, 2016.

[ZBZ+98] S Zhou, P Buckhaults, L Zawel, F Bunz, G Riggins, J L Dai, S E Kern, K W Kinzler, and B Vogelstein. Targeted deletion of Smad4 shows it is required for transforming growth factor

beta and activin signaling in colorectal cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, 95(5):2412–2416, mar 1998.

[ZFZM20] Sanguo Zhang, Yu Fan, Tingyan Zhong, and Shuangge Ma. Histopathological imaging features-versus molecular measurements-based cancer prognosis modeling. *arXiv preprint arXiv:2006.03702*, 2020.

[ZMD18] Ming Zhao, Lopa Mishra, and Chu-Xia Deng. The role of TGF-β/SMAD4 signaling in cancer. *International journal of biological sciences*, 14(2):111–123, jan 2018.

[ZYL+16] X. Zhu, J. Yao, X. Luo, G. Xiao, Y. Xie, A. Gazdar, and J. Huang. Lung cancer survival prediction from pathological images and genetic data — an integration study. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1173–1176, 2016.

A Related Work

Molecular profiling has increased understanding of the pathology of cancer [RR17, TBB+15, BWSH+11, CBCB19], allowing pathologists to use phenotypic molecular and morphological data to identify increasingly specific cancer subtypes [Ina17, GWC+13, Sch10], but subtype identification has been hindered by the lack of knowledge of the biological processes underlying tumor growth [HW11, Daw17, HPL14, SSZF19]. In previous work, this was circumvented by using observed phenotypic and/or genotypic data to define cancer subtypes [KI18, SCM19, IKK+20, WTL+19, WCY+18, GSCM+19, CHG+19, MYA+18, MSC+19]. Similarly, rising interest in linking morphological patterns of cancer (e.g., histopathology images) with intra-cellular mechanisms [SGLGK22] has inspired many attempts to use deep learning to predict genetic mutations from morphology [FJT+20, VSS+17, ADME18, PBC+16]. While these approaches have proven that transcriptomic changes are predictable from detectable morphological patterns, these approaches are limited to pre-specified transcriptomic patterns. and the features identified capture only a subset of the biological processes underlying tumor growth and relate to only pre-specified tasks. Our work is the first to define cancer subtypes according to their implications on downstream discriminative models. This definition enables us to learn patient-specific models for downstream prediction tasks, linking data modalities and discovering meaningful patterns without pre-specification of genomic patterns.

Challenges of Interpretable Multi-modal Predictions A key challenge in multi-modal predictions is interpretability. Previous work [COS+18, WCY+18, GSCM+19, WTL+19, LZY+17, YZB+16, LdPM+19] built interpretable predictive models of only one modality, such as histologic patterns which are interpretable to pathologists. Multimodal models increase predictive capabilities, but interpretability has been hindered by the complexity of the models [ZYL+16, PBC+16, HKT+19, CZH+17, MYA+18]. The main way these models have been interpreted is through retrospective analyses of the features selected in their neural networks [HKT+19, MYA+18], and it is unclear how to use this type of interpretation to inform clinical decisions. In contrast, the CEN approach links modalities by directly generating interpretable sample-specific models.

Sample-Specific Model Parameters In this work, we identify discriminative subtypes by estimating a basis set of models which can be utilized to generate model parameters specific to each sample. Interest in such *sample-specific* model parameters for cancer analysis has grown in recent years, as increasing evidence has pointed to fine-grained subtypes which do not form discrete clusters [MOP+18, BNC+15]. Discovering these individualized molecular profiles could lead to substantial advances in the diagnosis and treatment of disease, in addition to refining our biological understanding of the mechanisms behind different diseases. Despite the recent surge of interest in personalized modeling [AYHvdS16, LMP17, MAK+07, KTY+15, LWJ+16, VKVK19], basic statistical challenges remain unanswered and deserve further study. In this work, we contextualize such personalized models with imaging data. We hope that this idea of contextualization can spur further development of such patient-specific modeling.

B Implementation Details

Architectural Components To encode histopathology images, we follow [COS+18] and use `InceptionV3` [SIVA17] architecture as context encoder. Since this architecture naturally operated on images of size 299x299, we sliced the larger histopathology images into non-overlapping patches of size 299x299, filtering out patches that had more than 25% of background whitespace.

For the interpretable probabilistic model, we used logistic regression (LR) that predicted cancer type from gene expression data. Predictions were computed both on the level of patches (in which case patches from the same slide were assigned the same slide-level labels and gene expression data) and on the level of slides (using majority vote over the corresponding patches). Other choices and hyperparameters are detailed in Section B, and

Baselines We compare against both uni-modal and multi-modal baselines. Firstly, we compare against the two uni-modal parts of the CEN architecture: logistic regression on transcriptomic profiles and the convolutional neural network on histopathology images. In addition, we compare performance against two multi-modal baselines. Multi-modal baseline 1 ("Concatenated") is the same architecture used for the state-of-the-art subtype prediction in [MYA+18], which builds a predictor on concatenation of the RNA-seq features and the output of the InceptionV3 network, while multi-modal baseline 2 ("Ensemble") is an ensemble (i.e., a weighted combination) of the two uni-modal models.

Data. The data was retrieved from TCGA on April 16, 2019 and preprocessed using the pipeline released by [COS+18]. The images are subject to a non-deterministic data augmentation similar to the one implemented in [COS+18]. We perform a log-transform and mean-centering of the transcriptomics data and select only those features with a standard deviation greater than 1e-2.

Partition	**Healthy**		**LUAD**		**LUSC**		**Total**	
	patches	slides	patches	slides	patches	slides	patches	slides
Training	59,530	89	658,498	562	722,396	518	1,440,424	1,169
Validation	10,478	17	132,101	120	156,698	116	299,277	253
Test	11,899	18	118,521	124	183,946	113	314,366	255

Table B.1: Dataset sizes by counts of slides and patches. The slides are divided into patches.

Architectures. To the `Inceptionv3` archiecture, we added a top dense layer of 128 neurons with dropout (keep probability=0.7). The Contextual Explanation Network parameters have been optimized via a grid-search on the dictionary size along with the regularization of the contextual layer. The final model uses a dictionary size of 32 and ℓ_1 regularization parameter of 1e-5.

Training. We trained all models for 5 epochs with a batch size of 64 on 1 NVIDIA® Tesla® P100 GPU. The model is optimised using RMSProp with parameters $\rho = 0.9$, momentum $= 0.9$, $\epsilon = 1.0$. The choice of the checkpoints is based on a best validation loss criterion.

C Extended Results

C.1 Classification Results

Table C.1 provides the slide-level confusion matrices for all models on the held-out test data. These results, analogous to the accuracy metrics reported in Table 1 of the main text, demonstrate the improved accuracy of CEN over existing models for LUAD/LUSC/Healthy classification.

	CEN			Logistic Regression			InceptionV3		
	Healthy	**LUAD**	**LUSC**	**Healthy**	**LUAD**	**LUSC**	**Healthy**	**LUAD**	**LUSC**
Healthy	18	0	0	15	1	2	12	5	1
LUAD	1	117	6	3	115	6	4	102	18
LUSC	0	7	106	0	8	105	1	22	90

	Concatenated			Ensemble		
Healthy	**LUAD**	**LUSC**	**Healthy**	**LUAD**	**LUSC**	
Healthy	17	1	0	15	1	2
LUAD	1	115	8	3	115	6
LUSC	0	7	106	0	6	107

Table C.1: Slide-level confusion matrices for baselines and CEN. Rows correspond to ground truth labels and columns to predictions made by the corresponding models.

C.2 Intra-tumor Heterogeneity

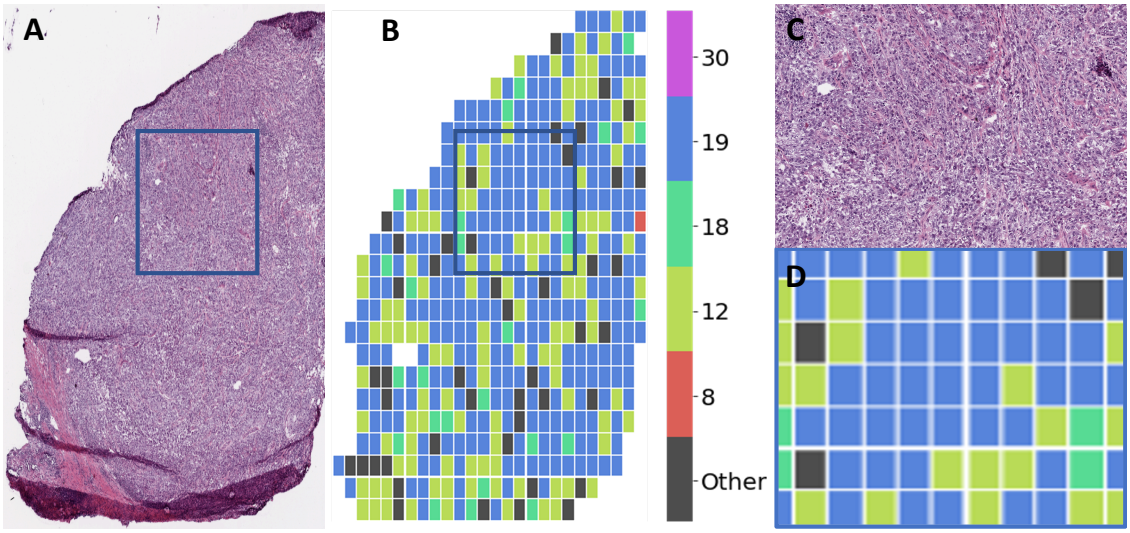

Figure C.1: Discriminative subtypes display significant heterogeneity throughout the sample and cluster spatially. Representative results of patch-level subtype prediction for a LUAD sample held out from CEN training. **(A)** shows the tumor whole-slide image. **(B)** displays the archetype which CEN assigned to each patch; for visual clarity, we display only the common archetypes 8, 12, 18, 19, and 30, and group all other archetypes under "other". **(C)** inset square outlined in blue in panes A and B. **(D)** inset for a matched control sample from healthy tissue surrounding the tumor.

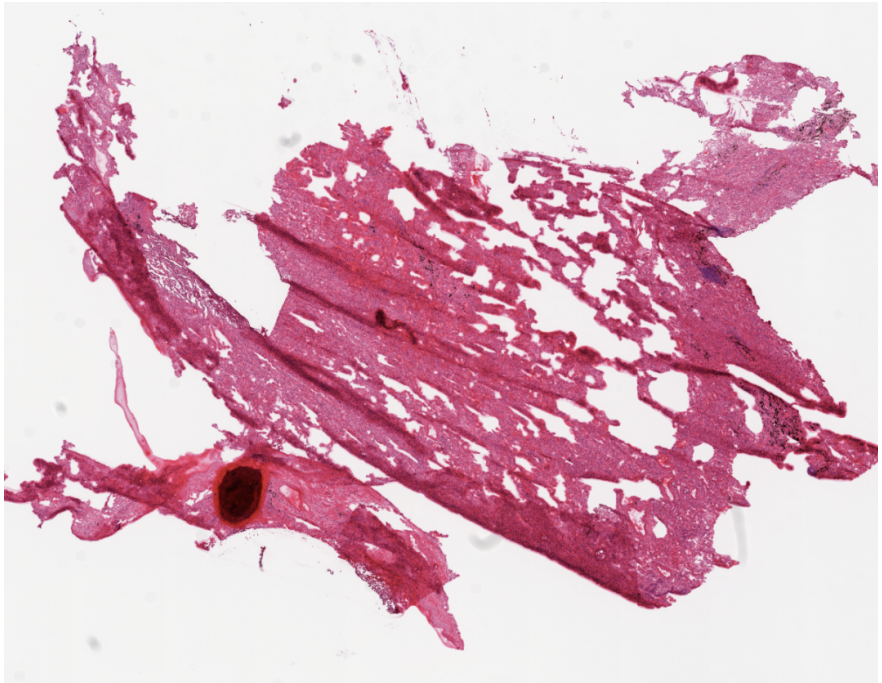

Figure C.2: Matching control sample of healthy tissue surrounding the LUSC tumor in Figure 3.

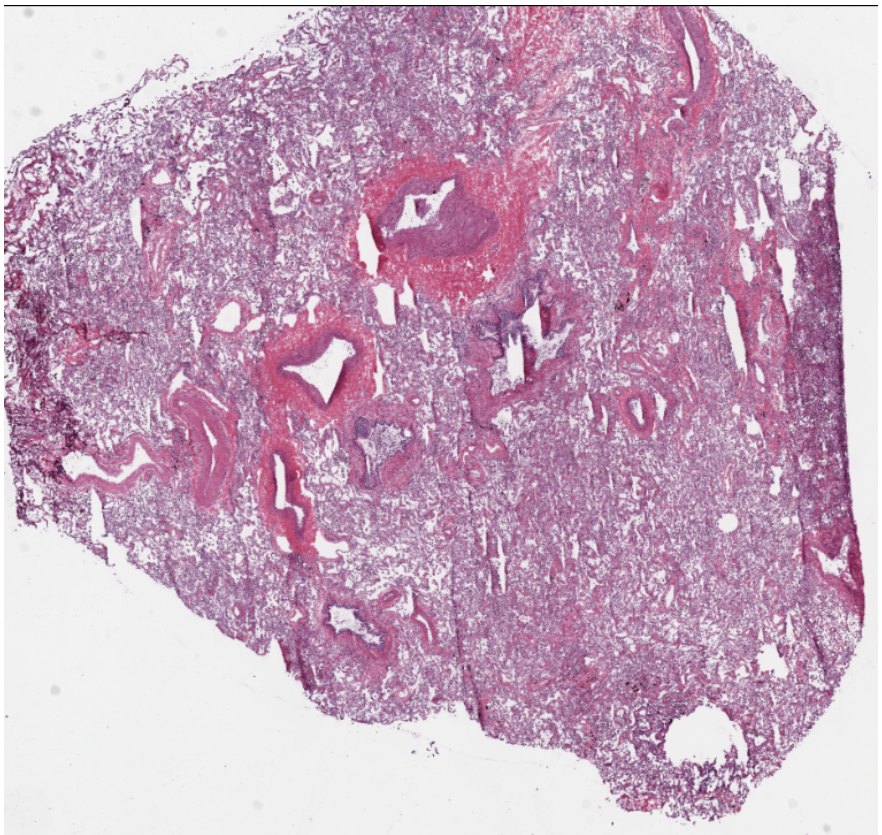

Figure C.3: Matching control sample of healthy tissue surrounding the LUAD tumor in Figure C.1.

C.3 Morphological Patterns of Archetypes

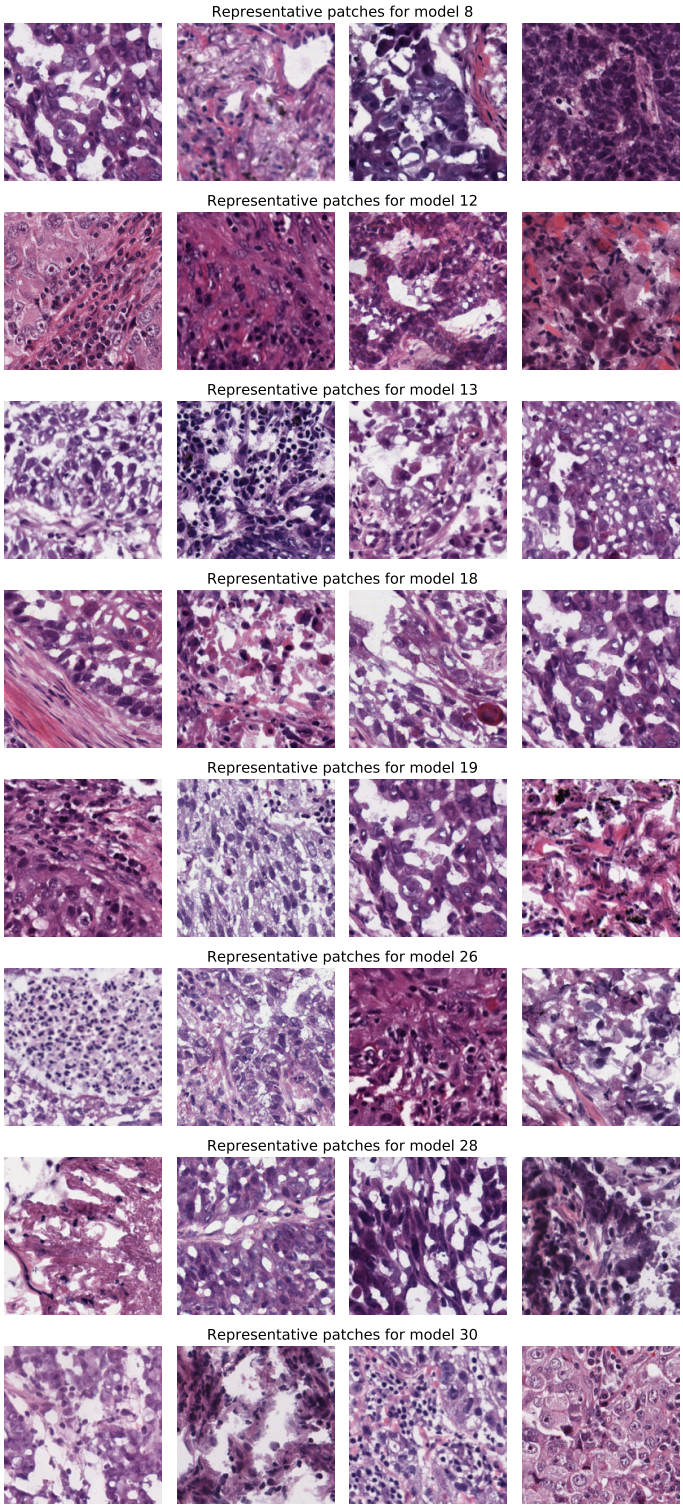

Figure C.4: Top four patches that made CEN assign the highest weights to corresponding archetypes.

C.4 Analysis of the Discriminative Subtyping Hypothesis

Our discussion started with a hypothesis that gene expression data must be interpreted with respect to patient-specific contexts that may correspond to different subtypes. The contextual learning approach we proposed to use in this work allowed us to discover latent subtypes, which we termed *discriminative* since each of the subtypes corresponded to an interpretable linear model that could accurately discriminate between different classes (cancer types in our case) based on transcriptomics.

However, if samples of the same type always mapped to the same linear model, while samples of different types mapped to different models, this would have been troublesome—the biological interpretation of the archetypal coefficients would have been confounded by the assignment of different models to different classes. As shown in Figure C.5, this is not the case: the distribution over the weights assigned to archetypes by the context encoder conditional on the target label are nearly identical across all classes. Thus, indeed linear models that correspond to each of the subtypes are ultimately discriminating between the classes.

While class-conditional distributions over the discovered subtypes are nearly identical, as shown in Figures C.6, C.7, C.8, C.9 different tumor stages do show different distributions of discriminative subtypes. This may reflect the tendency of different subtypes to progress at different rates, or different efficacy of treatment for different subtypes.

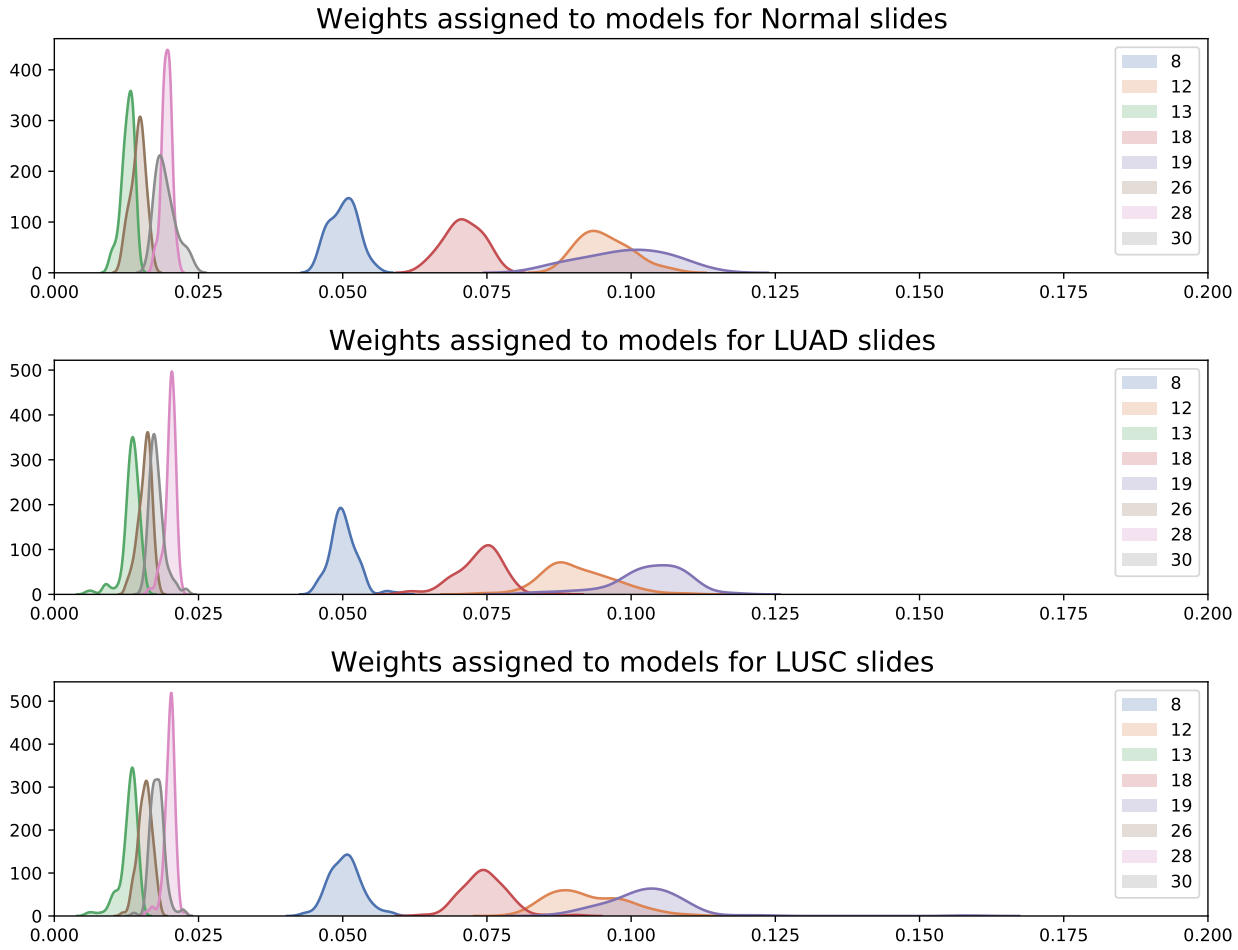

Figure C.5: Class-conditional distributions over the archetype weights assigned by CEN, visualized for healthy, LUAD, and LUSC slides. There is only slight variation between distributions of the weights assigned to different archetypes, and the relative ordering between the archetypes is preserved.

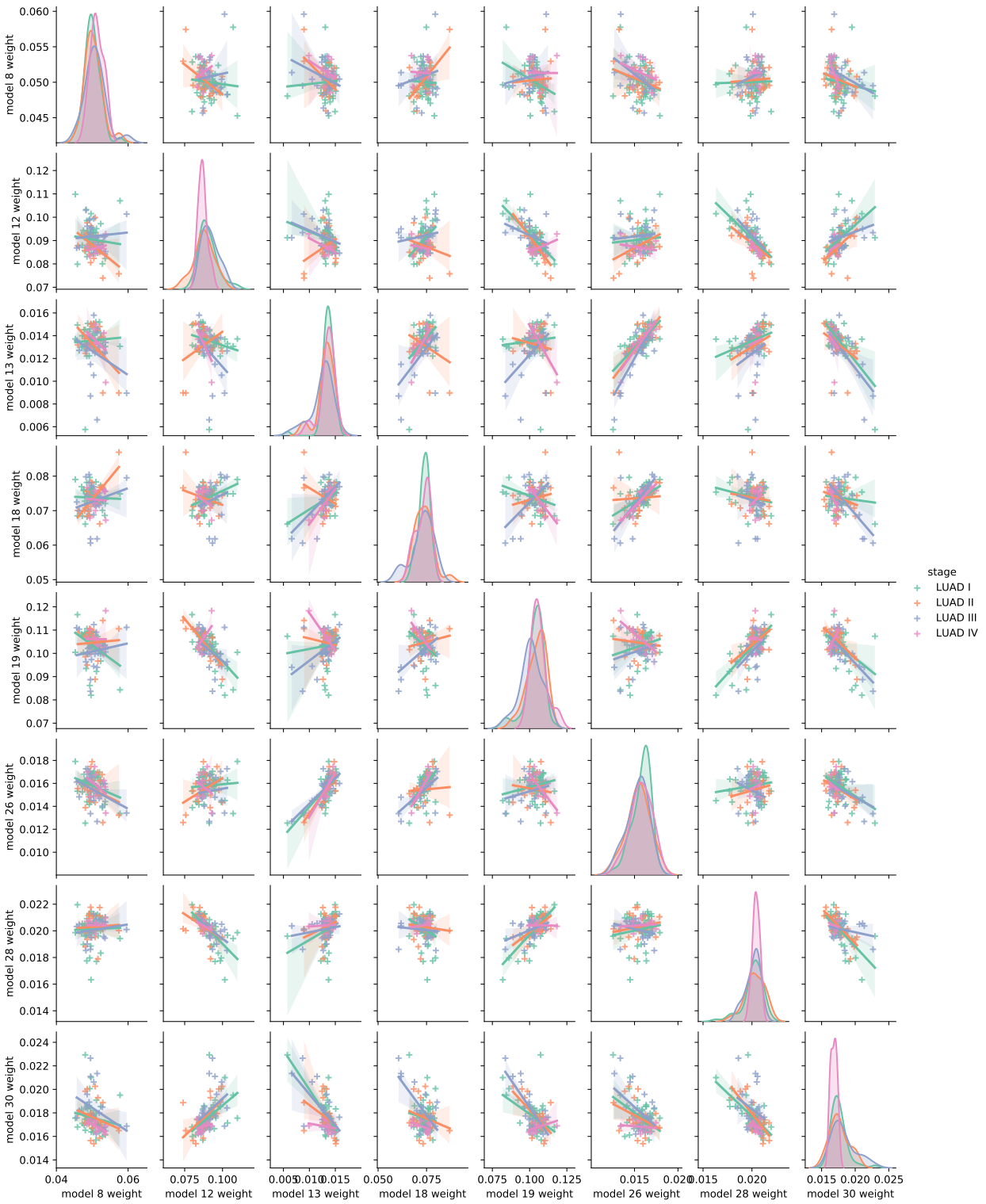

Figure C.6: Distribution of weights assigned to 8 different models in the dictionary for different stages of LUAD. Densities on the diagonal indicate variation of the assigned weights for each stage of LUAD. Off-diagonal plots showcase pairwise dependencies between model weights for different stages of LUAD.

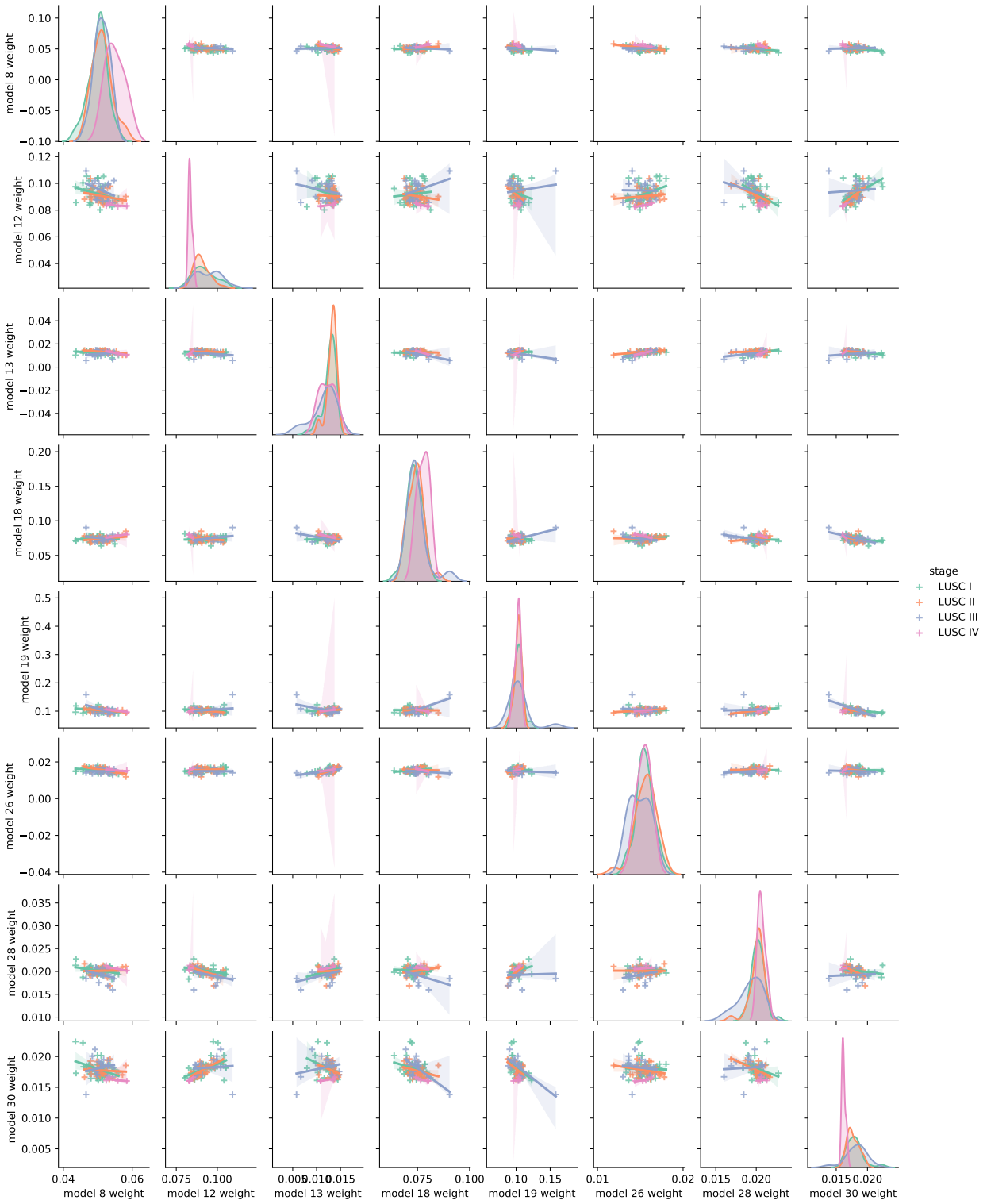

Figure C.7: Distribution of weights assigned to 8 different models in the dictionary for different stages of LUSC. Densities on the diagonal indicate variation of the assigned weights for each stage of LUSC. Off-diagonal plots showcase pairwise dependencies between model weights for different stages of LUSC.

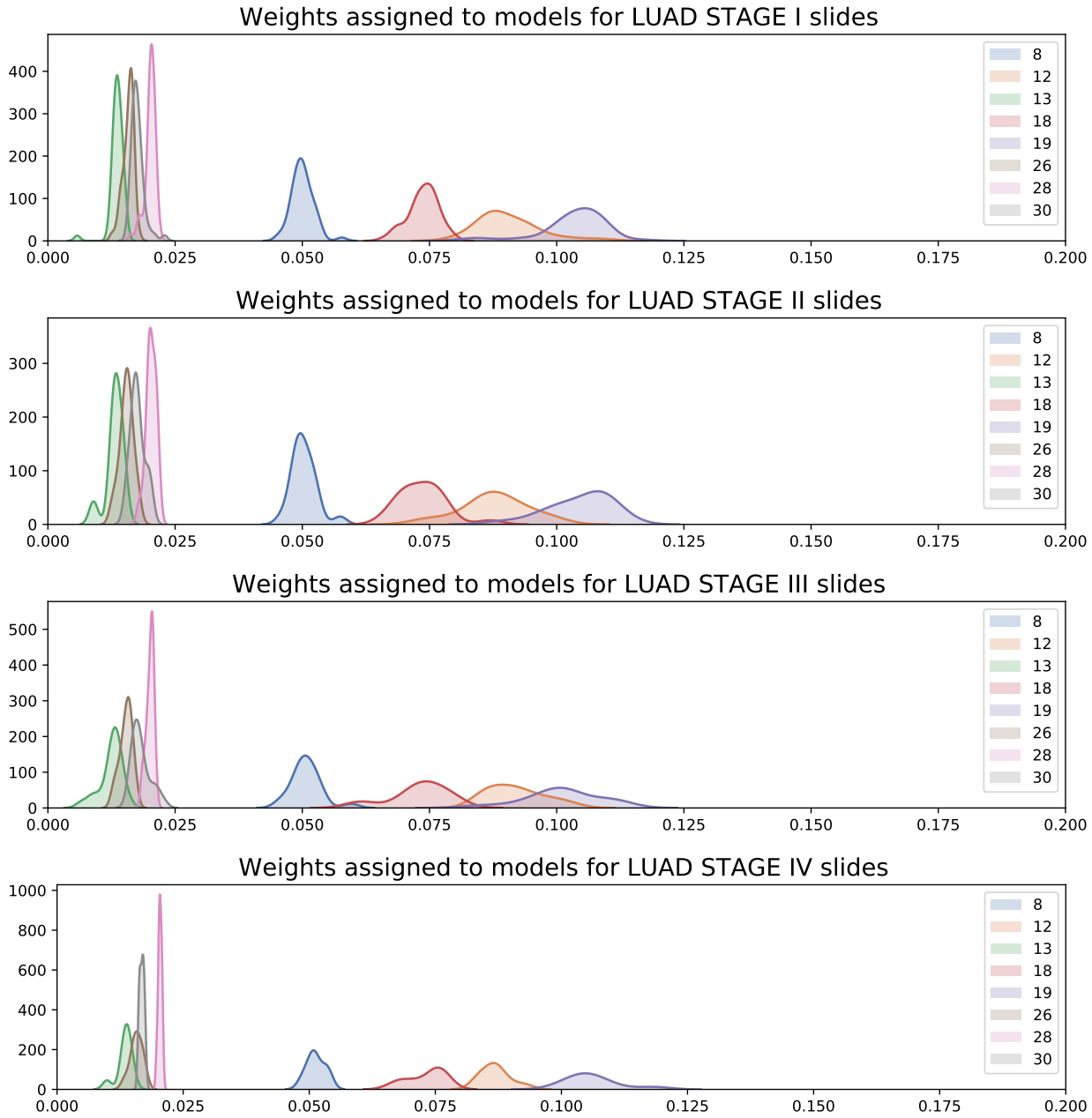

Figure C.8: Distributions over the weights assigned to different models in the CEN dictionary visualized for LUAD slides of stages I-IV.

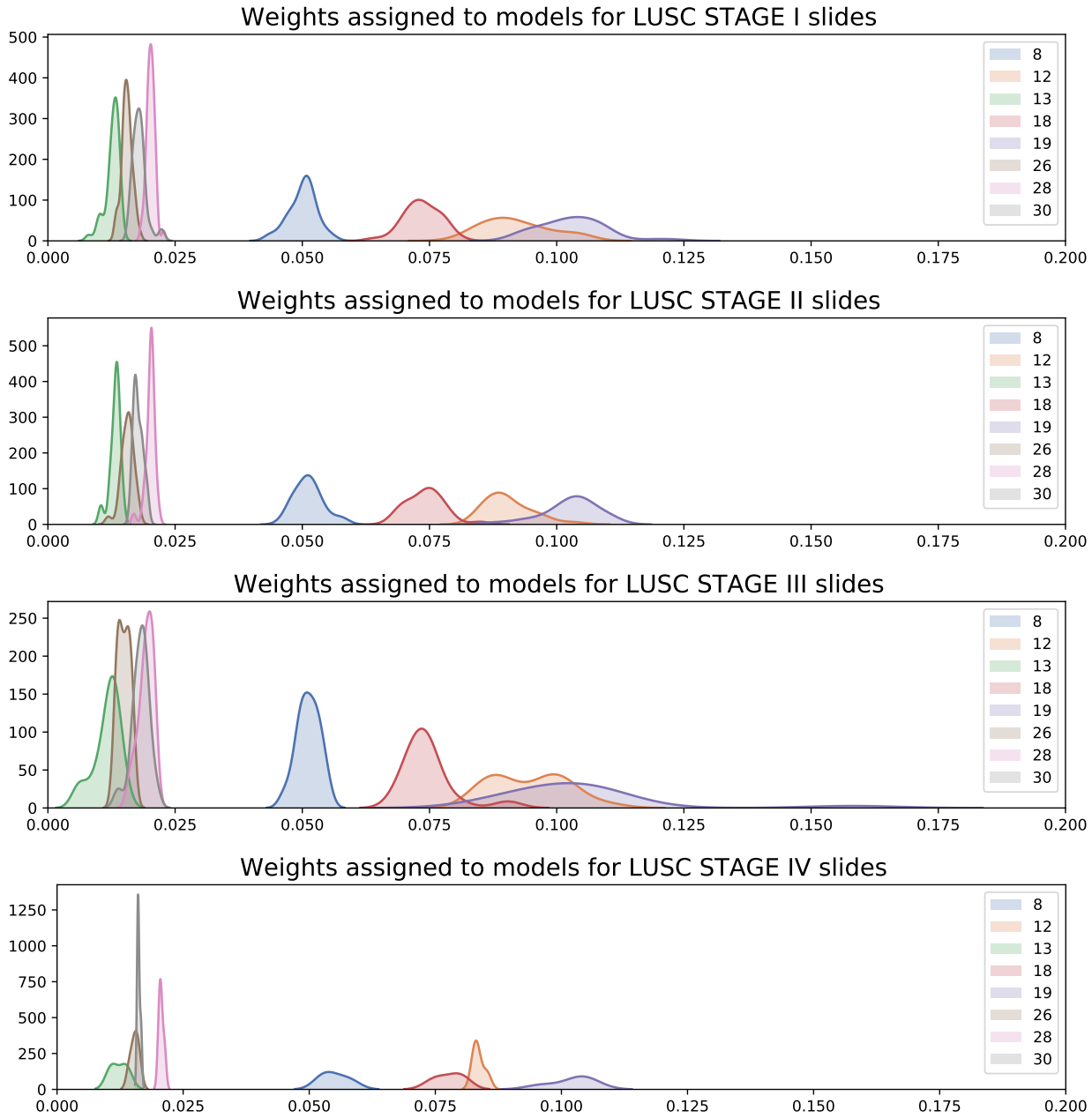

Figure C.9: Distributions over the weights assigned to different models in the CEN dictionary visualized for LUSC slides of stages I-IV.

C.5 Archetype Analysis

C.5.1 Enrichment Analyses

Archetype	Term ID	Term Name	P-Value
5	GO:0120038	plasma membrane bounded cell projection part	0.038
7	TF:M07385_1	Factor: Lhx2	0.008
11	GO:0030101	natural killer cell activation	0.011
11	KEGG:05321	Inflammatory bowel disease (IBD)	0.001
11	KEGG:04659	Th17 cell differentiation	0.003
16	CORUM:3075	UTX-MLL2/3 complex	0.003
22	TF:M00693_1	Factor: E12	0.012
23	HPA:008020_13	cerebral cortex; glial cells[Supported,High]	0.050
27	TF:M01596	Factor: Gli3	0.000
27	TF:M07455_0	Factor: Gli2	0.001
30	GO:0006259	DNA metabolic process	0.017

Table C.2: Terms enriched in LUAD archetypal models ($p < 0.05$).

Archetype	Term ID	Term Name	P-Value
0	REAC:R-HSA-6805567	Keratinization	0.000
2	TF:M02095_0	Factor: Fra-1	0.032
3	HPA:050020_02	spleen; cells in white pulp	0.022
6	GO:0007626	locomotory behavior	0.020
7	REAC:R-HSA-9006927	Signaling by Non-Receptor Tyrosine Kinases	0.035
8	TF:M05884_0	Factor: znf439	0.017
9	GO:0031435	mitogen-activated protein kinase kinase kinase binding	0.011
12	GO:0051962	positive regulation of nervous system development	0.034
15	CORUM:2815	BRCA1-BARD1-BACH1-DNA damage complex II	0.029
16	REAC:R-HSA-4791275	Signaling by WNT in cancer	0.029
20	REAC:R-HSA-2173789	TGF-beta receptor signaling activates SMADs	0.043
24	TF:M08982_0	Factor: YY1	0.020
26	GO:0008038	neuron recognition	0.010

Table C.3: Terms enriched in LUSC archetypal models ($p < 0.05$).

C.5.2 Influential Genes in Archetypes

Archetype	Top 5 Genes
0	FOXO4, FNBP1, CLTC, PAX7, AC103769.1
1	RNA5SP160, NONAME, ALDH2, C10orf99, ZFHX3
2	ZNF29P, NONAME, PCBP1, FOXO3, AL353626.2
3	UGT2A3, SRGAP3, VTRNA1-2, EP300, AL645939.2
4	RF00019, EP300, SNORD72, NONAME, IGHV3-62
5	RHOH, MTCP1, TFG, CCR7, SDHD
6	AKT3, PRAC1, HOXD13, MAP2K1, MIR6730
7	NONAME, FOXL2, REG1A, KTN1, MIR643
8	MIR6835, ASXL2, AC147055.1, RF00019, RNU6-100P
9	MIR374B, IGLV5-48, NONAME, TNFRSF17, BNIP3P15
10	MYCL, PLCG1, DDX3X, AL133370.1, SRGAP3
11	MIR548AR, AC015813.4, EIF3E, SNORA22C, STAG2
12	RF00410, LMO1, RNU6-130P, BCL6, RF00019
13	BMPR1B-DT, AC245041.1, RNA5SP268, RNU6V, RSPO3
14	FGA, PTPRB, NONAME, MIR7152, KAT6B
15	NONAME, MIR320A, EML4, ZFY, FOXO3
16	DPP10-AS1, RNU6-824P, NFATC2, RNU6-892P, RF00019
17	AL645939.2, MIR196A1, APOH, RNU7-75P, SNORA20
18	RNU6-403P, MIR6728, SNORA38, TRAF7, NCOA1
19	RARA, IGKJ5, NONAME, RNA5SP477, RNU6-90P
20	EBF1, ABL2, AC136428.1, RNU6-407P, REG3A
21	CBLC, PHGR1, MICOS10P3, NCOA1, AL162413.1
22	AC111000.2, PDCD1LG2, LCK, S100A7, APOA2
23	NONAME, BCL7A, POLD1, RF00019, MAML2
24	RF02119, RNY4P25, PRAC2, AL512363.1, RF01518
25	SH3GL1, EPS15, RNU12-2P, MIR6740, RNU5A-1
26	AC008649.2, MIR5091, PSPHP1, SNORA58B, EPS15
27	RF02272, AC129926.1, MIR3657, FAT3, NONAME
28	NONAME, MIR4677, STAT3, TMSB10P2, IGLJ3
29	OLIG2, MIR5091, SLC17A4, CBFA2T3, EPAS1
30	KCNJ5, RNA5SP298, UGT2B4, RNU6-315P, RF00019
31	NONAME, GMPS, AC007849.1, WFDC12, CASP9

Table C.4: Most influential genes in each case/control archetypal models.

Archetype	Top 5 Genes
0	KRT6C, SPRR3, HRAS, NONAME, SPRR1A
1	IGLJ1, SNORA74A, PML, IGLV2-5, DAXX
2	EZH2, NONAME, IGHV1OR21-1, RF00019, AL354766.2
3	BCL11A, FANCE, SAA2-SAA4, NONAME, IGF2BP2
4	IGHV3-35, SCGB1A1, HOXC13, MUTYH, SNORD48
5	LGALS7, MIR4713HG, NONAME, TSHR, ROBO2
6	NONAME, TM4SF5, TSC1, AL645924.1, RF00265
7	RF02272, BTG1, CNTRL, IGKV2D-28, RNY4P19
8	TGFBR2, KIF5B, KMT2A, AR, RNA5SP39
9	FANCE, TGFBR2, RAD17, CLTCL1, CBFB
10	TSC2, AP000439.2, SNORD116-24, BCOR, RNU6-460P
11	DROSHA, ATR, CDK6, MLF1, DCTN1
12	SERPINB4, BRIP1, IDH2, CIC, CLP1
13	TFRC, RNU7-75P, TCL1A, AC010280.1, CD79B
14	ZBTB16, NUMA1, AR, LASP1, RNA5SP301
15	MIR6071, RNU7-186P, MITF, HERPUD1, MPL
16	CRCT1, AC025575.2, MIR8071-1, HMGA2, KLF4
17	NFE2L2, C5orf66-AS1, TLX3, HOXD13, NONAME
18	SPRR2E, KEAP1, EIF1AY, IGFL1P1, IGHV7-56
19	FOXO4, RAD17, MIR5690, HIP1, AKAP9
20	AC018659.2, LINC01351, RF00019, ARAF, PRKAR1A
21	ZRSR2, RF00019, NCOA1, USP6, MIR6859-4
22	KRT6B, RFWD3, RNU6-353P, RF02272, RNU6-48P
23	KNSTRN, SPRR2E, MIR6832, HEY1, AC008663.1
24	AC245100.3, NONAME, RF00019, MIR4685, RF00019
25	DDR2, AC015922.1, NONAME, MDM2, MT1P3
26	ZNF521, MIR342, JUN, RNU1-67P, NONAME
27	ZBTB16, RERG-IT1, NONAME, RNU6-377P, RNU6-483P
28	ACKR3, FBXO11, IGKV1-39, POLG, DCC
29	AL139415.2, MT-TL2, GATA2, ZRSR2, NONAME
30	FANCE, NONAME, SRSF3, FANCF, TET1
31	AC112206.3, GMPS, AP000851.2, RNA5SP477, NONAME

Table C.5: Most influential genes in each LUAD archetypal models.

Archetype	Top 5 Genes
0	TRIM27, WFDC5, AP002498.1, RNU6-944P, CLTC
1	MUC1, FBXW7, RN7SL8P, MIR27B, AC135068.2
2	IL6ST, RNA5SP21, KRAS, MIR4668, RF00019
3	CACNA1D, ARHGEF38-IT1, RF00019, IGLV2-28, TRIM27
4	PHOX2B, IGLV3-16, SDC4, RNU4-21P, RNU6-652P
5	FOXP1, SDHD, NONAME, AL139317.2, MALAT1
6	AC004080.2, AC131532.1, SGK1, AC010280.1, MLLT1
7	ITK, ALDH2, ZNF429, RNY1, C20orf85
8	NUTM1, SNORA58B, MB21D2, ARID2, RNVU1-4
9	AC092569.1, AL354766.2, MIR5188, RNU6-834P, TRAF7
10	KNSTRN, MTOR, RAD21, IRS4, ALK
11	SCGB2A1, SFRP4, REG1A, POU2AF1, NONAME
12	USP44, RNY4P25, TCEA1, CLTC, TFEB
13	CIITA, AL157388.1, MITF, AL645939.2, RNU6-720P
14	BCL11B, SNORA54, SERPINA11, RABEP1, NONAME
15	SPRR2A, NONAME, CDKN2A, ZFY-AS1, SRGAP3
16	FBXW7, RNU6-944P, KDR, UGT2B4, PREX2
17	IGHV3-38, NONAME, AC016769.3, USP44, TP53
18	IL6ST, RNU1-47P, ESR1, NONAME, GP2
19	NFE2L2, FGFR3, CRCT1, RF00334, SETBP1
20	SPRR2E, TFRC, MIR8055, RF00019, CRISP3
21	RIPPLY2, HOXC13-AS, GUCA2B, NONAME, AC092755.1
22	RNU6-944P, IGLV3-16, RF00019, RF00001, HIST1H4I
23	IGHV3-71, ESR1, RPL10, COL1A1, PHGR1
24	CACNA1D, TRIP11, HNF1A, RNU6-1010P, MIR374B
25	SPRR2F, ARID2, MYH11, NONAME, TMEM127
26	AMTN, AL390729.1, RF01518, DEK, PRCC
27	CALML5, DDX5, ASXL2, NFIB, RNU6-323P
28	MIR126, DDX6, RNU6-1189P, GNAQ, CUL3
29	MYCL, RF00019, SIRPA, POLE, RNU6-892P
30	AC026462.3, MIR3178, ERG, MIR4512, SCGB1A1
31	AC092569.1, RNU5D-1, RNU6-4P, SNORA2B, MIR27A

Table C.6: Most influential genes in each LUSC archetypal models.

C.5.3 Orthogonality of Archetypes

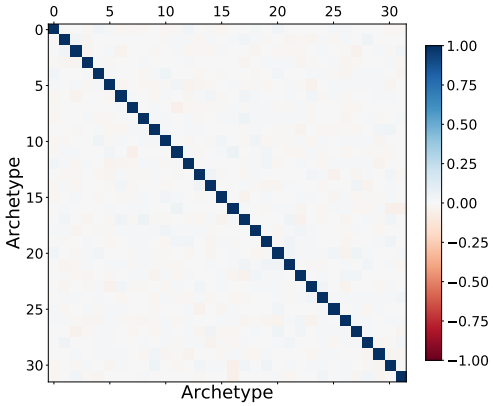

Figure C.10: Kendall Tau similarity of ranks of transcripts selected by each entry in the dictionary. All off-diagonal comparisons have similarity less than 0.1, indicating that archetypes are orthogonal.