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]

[1]Carnegie Mellon University

June 25, 2020

Abstract

When designing individualized treatment protocols for cancer patients, clinicians must synthesize the information from multiple data modalities into a single parsimonious description of the patient's personal disease. However, such a description of a patient is never observed. In this work, we propose to model these patient descriptions as latent *discriminative subtypes*—sample representations which can be learned from one data modality and used to contextualize predictions based on another data modality. We apply contextual deep learning to learn these sample-specific discriminative subtypes from lung cancer histopathology imagery. Based on these subtypes, we produce sample-specific transcriptomic models which accurately classify samples as adenocarcinoma, squamous cell carcinoma, or healthy tissue (F1 score of 0.97, outperforming previous state-of-the-art multimodal approaches). Combining these data modalities in a single pipeline not only improves the predictive accuracy, but also gives biological interpretations of the discriminative subtypes and ties the phenotypic patterns present in histopathology images to biological processes.

1 Introduction

In many biological settings, inter-sample heterogeneity is critical to understanding the complex biological processes under study. For example, in the analysis of cancer gene expression assays, each patient in a cohort may have a different somatic mutation. As these mutations adjust the baseline gene expression levels of each patient, the observed gene expression data must be interpreted with respect to patient-specific "contexts." These contexts can be described as *molecular subtypes* which are associated with patient outcomes [1] and can inform the decision-making process behind surgery and radiation therapy [2]. Our understanding of molecular subtypes is still advancing, and much recent work has focused on identifying fine-grained descriptions which comprise previously-characterized subgroups (e.g. in breast [3] or lung cancers [4]).

If the perfect contexts were annotated (e.g., if molecular subtypes were to correspond exactly to primary tissue sites), we could simply train a different probabilistic model for each context. However, meaningful contexts are often complex and possibly unknown. Simply increasing the capacity of the predictive model to handle multiple contexts is typically not feasible due to limited sample size and an emphasis on fitting interpretable models. Instead, we would like to use a model which captures the complexity of different contexts while retaining the domain knowledge and interpretability of structured probabilistic models.

*blengeri@cs.cmu.edu
†alshedivat@cs.cmu.edu
‡aalavi@cs.cmu.edu
§jlw1@cs.cmu.edu
¶slabbaki@andrew.cmu.edu
‖epxing@cs.cmu.edu

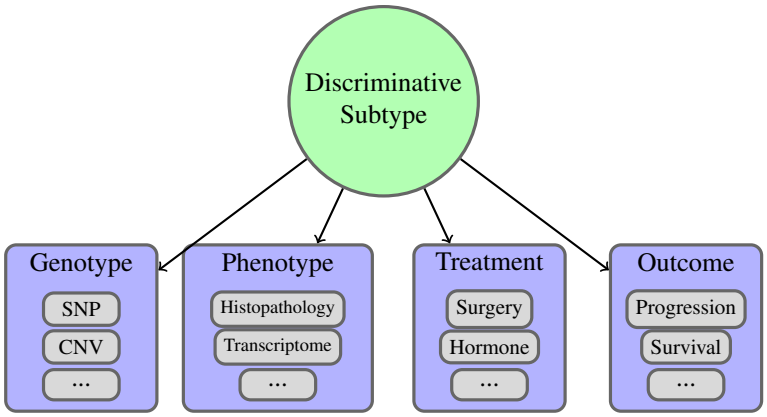

Figure 1: The most concise description of the many variables regarding oncology patients is an unobserved latent variable. In this work, we seek to estimate this latent variable as a *discriminative subtype* which improves the discriminative ability of downstream predictors.

To do so, we introduce the notion of a *discriminative subtype* (Figure 1). A discriminative subtype is a latent variable which captures the variation in many observable variables, and can be used for a variety of downstream tasks. Discriminative subtypes may correspond, but are not limited, to previously-characterized molecular subtypes.

In this work, we will use a deep neural network to estimate the discriminative subtype for each sample based on contextual data. We will use this subtype information to generate the parameters for an interpretable model for downstream tasks. This two-stage procedure allows us to analyze transcriptomic data in the context of complex context data such as diagnostic images. Specifically, we investigate the capacity of this architecture to predict disease subtypes from the histopathology images and transcriptomic data of patients diagnosed with two types of lung cancer: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC).

1.1 Lung Cancers

Lung cancer is one of the top ten most common cancers for males and females, and is estimated to account for roughly 13% of all new cancer cases[5]. For males, it is second only to prostate cancer, and for females it is second only to breast cancer. In terms of mortality, it is estimated to be the most lethal cancer for both sexes, accounting for 23% of cancer deaths in males and 22% of cancer deaths in females for the year 2020 [5]. It is estimated that in 2020 in the United States, 228,820 new cases of lung cancer will be diagnosed and 135,720 people will die from lung cancer [5].

The most common category of lung cancer is Non-Small Cell Lung Cancer (NSCLC), accounting for 85% of cases. LUAD and LUSC are subtypes within this NSCLC category, and are the two most common types of lung cancers, accounting for 38.5% and 20% of all lung cancer cases respectively [6]. When lung cancer is diagnosed, it is categorized into one of four stages based on the TNM grading system (based on (T) the size of the tumor, (N) the spread of cancer to lymph nodes, and (M) metastasis) where a lower stage number means that there is less cancer in the patient. Early detection is critical in lung cancer: in the U.S., 61% of NSCLC patients diagnosed at the localized stage survived beyond five years compared to only 6% of those diagnosed at the distant stage.[1]

While LUAD and LUSC are both major subtypes of lung cancer, they have important distinctions in terms of their clinicopathology, tumor microenvironments, and molecular organization. For example, LUSC is associated with a history of smoking, while LUAD may occur in smokers but is the most common form of lung cancer in non-smokers [7]. These differences impact the best therapeutic approaches for each subtype. For example, epidermal growth factor receptor (EGFR) and anaplastic lymphoma kinase (ALK) are two commonly mutated genes in LUAD and are targets for current therapies, while they are infrequently mutated in LUSC [8]. In addition, LUAD and LUSC tumors each have different immune subtypes (tumor subtypes characterized by differences in infiltrating immune cell types and immunogenic expression), and these differences have been shown to have an effect on the efficacy of immune

[1] cancer.org

therapies (immune checkpoint blockade therapy) [9]. Given this diversity of pathologies, estimating a discriminative subtype that encapsulates these variations will be essential for machine learning tasks such as cancer type prediction.

1.2 Contributions and Generalizable Insights

As described above, effective machine learning systems for oncology require both the integration of data modalities (e.g. temporal, demographic, radiological, histopathological data) and model interpretability. The generic form of this problem is a significant open problem in machine learning. In this manuscript, we demonstrate that we can apply contextual deep learning to simultaneously achieve accuracy and interpretability for classification of lung cancers. This case study brings to light the following insights that apply in general to the problem of using machine learning techniques on multi-modal patient data:

- In applications with heterogeneous samples, understanding the context in which predictions are made is important to making accurate predictions. In this work, we show that sample-specific contexts can improve downstream predictions even when the contexts are latent and must be inferred from data (Section 5.2). Towards this end, we show that contextual deep learning is a promising tool for estimating these latent variables and linking different data modalities in a single pipeline.
- For transcriptomic analysis of lung cancer samples, we show that meaningful contexts can be inferred from histopathology samples. These meaningful contexts form *discriminative subtypes* (Section 5.5), which are not themselves correlated with the cancer type, provide meaningful context in which simple models can accurately label sample types.
- Finally, the learned transcriptomic models for different discriminative subtypes tend to give high attention to different biological processes (Section 5.3). This suggests that the sample-specific transcriptomic models learned from our discriminative subtypes are capturing the heterogenity of the biological processes underlying lung cancer. This highlights the promise of approaches which use contextual deep learning to learn biologically explainable/meaningful models for heterogenous data.

2 Related Work

Molecular profiling has increased understanding of the pathology of cancer [10–13]. This understanding allows pathologists to use phenotypic molecular and morphological data to identify increasingly specific cancer subtypes [14–16]. Improved cancer subtype identification enables clinicians to devise more effective individualized treatment plans for patients with cancer. The machine learning community is increasingly working to assist pathologists further improve cancer subtype identification to better inform treatment based off molecular and morphological data [17–26]. This common goal of better informing treatment has also led the machine learning community to work towards creating patient-specific interpretable and explainable predictive models to provide actionable insights to clinicians about the prognosis of their patient's disease [24, 27–29].

2.1 Cancer Subtype Identification

Identifying cancer subtypes is crucial to devising individualized treatment protocols for patients with cancer, but subtype identification has been hindered by the lack of knowledge of the biological processes underlying tumor growth [30–33]. In previous work, this was circumvented by using observed phenotypic and/or genotypic data to define cancer subtypes [17–25]. These papers defined specific cancer subtypes based off phenotypic and/or genotypic features which were meaningful for specific downstream tasks such as predicting healthy versus cancer. Recently, a new definition of subtype was introduced based off mutations predicted from histopathology images within a clinically accepted cancer type [34]. This approach is promising, as it could assist pathologists detect gene mutations to inform treatment. While all of these definitions of a cancer subtype are useful, the features they identify only capture a subset of the aspects of the biological processes underlying tumor growth related to a specific task or data type. Our work is the first to our knowledge to associate discriminative models with cancer subtypes, which we call *discriminative subtypes*. This definition enables us to learn accurate patient-specific models for downstream prediction tasks.

2.2 Downstream Prediction

The key challenge to downstream prediction tasks is to use multimodal data to generate predictions that are interpretable and explainable to clinicians. In previous work, interpretable predictive models have been based off only one type of data, such as histologic patterns which are interpretable to pathologists [20–22, 34–37]. Multimodal models have leveraged different types of data to increase predictive capabilities, but their interpretability has been hindered by the complexity of the models [24, 27, 38–40]. The only way some of these complex models have been interpreted is with a retrospective analysis of the features selected in their neural networks, and it is unclear how to use this type of interpretation to inform clinical decisions [24, 39]. Our approach allows us to generate predictive models linking histopathology images and molecular data which are accurate, interpretable, and explainable to clinicians.

2.3 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 has grown in recent years, as increasing evidence has pointed to fine-grained subtypes which do not form discrete clusters [41, 42]. 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 [43–48], basic statistical challenges remain unanswered and deserve further study. Patient-specific analysis of cancer patients has mainly focused on two areas: detection of personalized biomarkers [28, 29, 49, 50], and inference of personalized regulatory networks [51, 52]. 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 efforts that appear promising.

3 Methods

To combine multi-modal data in an interpretable pipeline, we use Contextual Explanation Networks (CENs) [53], illustrated in Figure 2. CEN architecture assumes that the data is represented by: (i) context variables, denoted C, (ii) semantically meaningful variables, denoted X, and (iii) target variables, denoted Y. Here, we will use histopathology images as contextual data and gene expression assays as semantically meaningful predictors of the cancer type. The model will represent 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 information. 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 architecture produces a sample-specific parameterization of a linear probabilistic model that operates on transcriptomic data based on the sample's contextual data (i.e., based on histopathology).

Generation of parameters for sample-specific linear models is accomplished via a context encoder represented by a convolutional neural network (Figure 2). To reduce model complexity, parameters θ are further confined to be a linear combination of a small constant number (K) of "archetypes," denoted $\{\theta_1, \ldots, \theta_K\}$ or $\theta_{1:K}$. Specifically for a sample i with context c_i, weights θ_i are computed as follows:

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

where $(\alpha^1, \ldots, \alpha^K)$ is a vector output of the context encoder, which we call *discriminative subtype*. This dictionary of parameter sets, $\theta_{1:K}$, is learned jointly with the context encoder, and the entire architecture is trained end-to-end via backpropagation. To summarize, the context encoder is a deep neural network that processes contextual data (histopathology imagery) and outputs a probability vector of length K that softly selects weights for a linear probabilistic model from a dictionary of archtetypes.

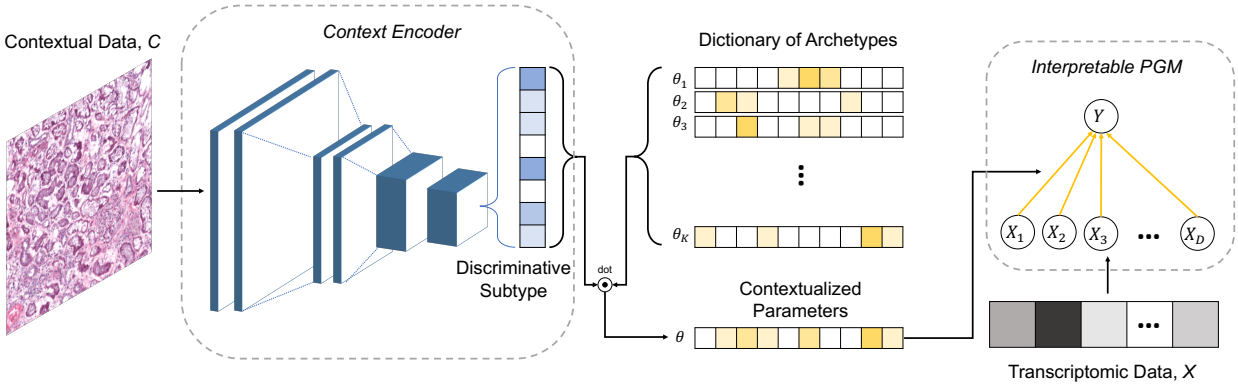

Figure 2: A CEN architecture combines "contextual" histopathology data (C) with gene expression data (X) through intermediate, sample-specific interpretable linear probabilistic models. We call the output of the context encoder the discriminitative subtype, as it is used to form sample-specific parameters θ by weighting the archetypal models stored in a dictionary. Finally, the formed sample-specific linear models operate on the transcriptomic data (X) to estimate the final target Y (in our experiments, cancer type).

3.1 Architectural Components

To encode histopathology imagery, following [34], we used `InceptionV3` [54] architecture as our context encoder. Since this architecture naturally operated on images of size 299x299, to utilize it on the large histopathology images, we sliced the images into non-overlapping patches of size 299x299, filtering out patches that had more than 25% of background.

Further, for the interpretable probabilistic part of CEN, 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). Minor model choices and hyperparameters are detailed in Section S1.

3.2 Baseline Methods

To benchmark the performance of the CEN model, 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 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 [24]. It predicts labels from a concatenation of the RNA-seq features and the output of the InceptionV3 network that encodes histopathology. Multi-modal baseline 2 ("Ensemble") is an ensemble (i.e., a weighted combination) of the two uni-modal models described above.

4 Cohort

For this investigation, we use lung cancer data available in the NCI Genomic Data Commons [55], which includes both TCGA and TCIA resources to provide multi-modal descriptions of a large number of cancer patients. As of April 2, 2020, this depository includes 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 reduces the dataset to a total of 992 patients, with 1 or 2 whole-slide image for each patient. We selected these datatypes because inference of discriminative subtype is most clinically-useful if it can be done from phenotypic data. In addition, we look forward to the availability of transcriptomic data from multiple time points to make dynamic predictions which track tumor progression.

We split this dataset according to patient ID into training, validation, and testing partitions. The size and composition of each partition is shown in Table 1.

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

Feature Choices

Image Subsampling. For the image analysis, we use the 20x magnification whole-slide images. To make these large images suitable for the pretrained Inception 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 [34]. 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. To reduce this dimensionality, 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, [56]). This leaves us with a final set of 1695 transcriptomic features.

5 Results

5.1 Evaluation Approach

The goal of our study is not only to evaluate the predictive ability of the proposed multi-modal approach, but also to study the biological meaning and clinical utility of the inferred clusters. Towards that end, first, we measure predictive accuracy compare performance of our model against unimodal approaches: `InceptionV3` trained to predict cancer type from histopathology only [34] and logistic regression that predicts cancer type from transcriptomic data only. Then, we seek to assign biological meaning to the learned model archetypes, find descriptive histopathology images for each archetype, and analyze the discovered discriminative subtypes.

5.2 Contextualization Improves Prediction

First, we measure the ability of the model to discriminate between LUAD, LUSC, and healthy samples. As shown in Table 2, the multi-modal CEN achieves the best classification results, outperforming both the multi-modal and the uni-modal baselines. Our results for the `InceptionV3` model come very close to the performance reported by [34]. Based on these accuracy and macro F1 metrics, we can see that most of the signal to distinguish the cancer types is held in the transcriptomic data, but a single linear transcriptomic model is not flexible enough to achieve the best accuracy over the diverse cohort.

5.3 Archetypal Models Correspond to Biologically Meaningful Processes

Next, we turn to interpret the learned models. A natural question is to what degree the different model archetypes are redundant or look at different biological processes. Surprisingly, the model archetypes coefficient vectors are nearly all orthogonal (Fig. S1), indicating that the archetype emphasize distinct biological processes.

Table 2: Predictive performance of models on the Normal/LUAD/LUSC classification task. Transcriptomic data (T) was not available for all samples with histopathological data (H), and thus CEN and Logistic Regression models were trained on fewer data points than the InceptionV3 baseline, although the achieve higher predictive performance.

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

	CEN			Logistic Regression			InceptionV3		
	Norm	**LUAD**	**LUSC**	**Norm**	**LUAD**	**LUSC**	**Norm**	**LUAD**	**LUSC**
Norm	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		
	Norm	**LUAD**	**LUSC**	**Norm**	**LUAD**	**LUSC**
Norm	17	1	0	15	1	2
LUAD	1	115	8	3	115	6
LUSC	0	7	106	0	6	107

To identify these biological meanings, we use enrichment analysis to identify biological processes statistically over-represented in the model archetypes. For each archetype, we have 3 vectors of coefficients defining the 3-way logistic regression model. For each vector, we sort the genes by the magnitude of the associated coefficient. The top 5 genes for each archetype of the model for normal, LUAD, and LUSC labels are listed in Tables S1, S2, S3, respectively. We search the top 100 genes in this order for enriched biological terms against the background of the remaining 1595 genes using gProfiler[57]. 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 4. Similar tables for the LUAD and LUSC models are available in Tables S4, S5. Here we see that a large number of the 32 archetypal models are significantly enriched for a diverse set of processes which are markers for cancer.

As a vignette, below we discuss the significance terms in Table 4 which have a p-value of less than 0.01:

- **13: Factor: Smad4.** Intersection: CRNKL1, MYH11, ZRSR2.
  SMAD family member 4 (SMAD4) is a necessary component of the transforming growth factor beta (TGFβ) pathway and in this capacity regulates proliferation [58–60]. SMAD4 has been established as a tumor suppressor in pancreatic, colon and lung cancer [59, 61, 62]. Decreased expression of SMAD4 has been identified in NSCLC, and has been shown to be correlated with poor prognosis in LUAD [61–64].

- **18: Factor: REST.** Intersection: FGFR1, NCOA1, CRTC1, CBLB, DGCR8, PRDM16, BAP1, OTOS, FOXL2, ETV5.

Table 4: Terms enriched in case/control archetypal models ($p < 0.05$).

Archetype	Term ID	Term Name	P-Val
1	KEGG:04071 KEGG:04310	Sphingolipid signaling pathway Wnt signaling pathway	0.037 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 GO:0010629	RUNX1 regulates transcription of genes involved in differentiation of HSCs negative regulation of gene expression	0.022 0.039
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 - multiple species	0.006
28	GO:0051240	positive regulation of multicellular organismal process	0.006
30	REAC:R-HSA-4791275	Signaling by WNT in cancer	0.045

RE1-silencing transcription factor (REST) is most widely known to repress neuronal genes in non-neuronal genes, but has also been found to be an oncogene or tumor suppressor gene in certain cancers as well [65]. In breast cancer, it has been shown to act as a tumor suppressor, and loss of this gene is associated with aggressive breast cancer [66]. In lung cancers, an isoform of REST has been indicated as a specific clinical marker for early detection of small cell lung cancers [67, 68].

- **26: Apoptosis - multiple species.** Intersection: BIRC3, BIRC6, CASP3.
  A hallmark of cancer cells is their ability to limit or avoid cell death induced by apoptosis [30]. In NSCLC, multiple genes in the KEGG apoptosis pathway have been found to be correlated with tumorigenesis, chemoresistance, and poor prognosis [69–74, 74–76]. As an example, altered expression and/or localization of caspases which are involved in the crucial last steps of the apoptosis pathway have been shown to make cells resistant to apoptosis enabling cancer to progress and making cells resistant to chemotherapy [71, 73–75, 77]. Both cancer progression and resistance to chemotherapy increase the likelihood of a poor prognosis.

- **28: Positive Regulation of Multicellular Organismal Process.** Intersection: PTPRC, CRTC1, CPEB3, COL1A1, IL6ST, PIM1, LEF1, BCL10, RUNX1, ATP1A1, STAT3, MYD88, SETD2, BCL9L, SRC, MIR126.
  This Gene Ontology term is broadly defined as those processes which up-regulate or activate the progress of organismal processes [78, 79]. In cancer, oncogenes activate tumor growth and proliferation processes by up-regulating other development processes, including angiogenesis, cell migration, and metastasis. As an example, aggressive cancers achieve metastasis by up regulating the genes involved in inflammation and migration of cells, which include genes that produce cell-to-cell or cell-to-extra-cellular-membrane adhesion molecules [30, 80].

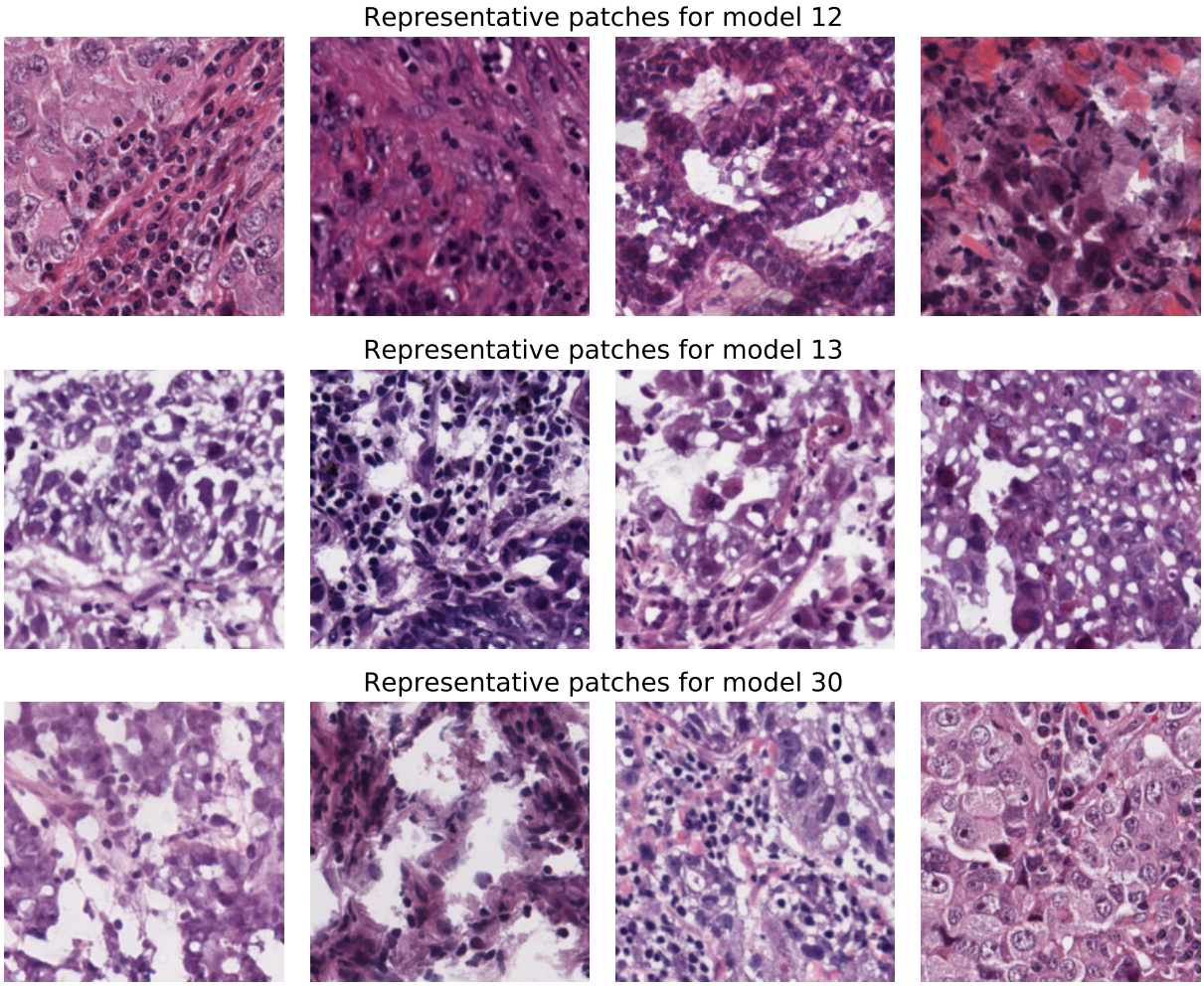

Figure 3: Patches that made CEN assign the highest weights to the corresponding models in the dictionary. Morphology of the patches is homogeneous within model-specific cluster and varies between the clusters.

5.4 Morphologic Patterns of Archetypal Models

One of the advantages of contextual deep learning is that it enables us to tie the morphology patterns recognized in histopathology images into the biological processes used to describe the transcriptomic models. Towards this end, we visualize representative patches which maximize the influence of archetypal models 12, 13, and 30 in Figure 3 (similar results for the other archetypal models are shown in Figure S6 in the Appendix). We can see that the morphology is more similar within the clusters corresponding to archetypes than between the clusters, indicating that the transcriptomic patterns used for accurate downstream predictions have a correspondence with morphological changes on a larger scale.

5.5 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

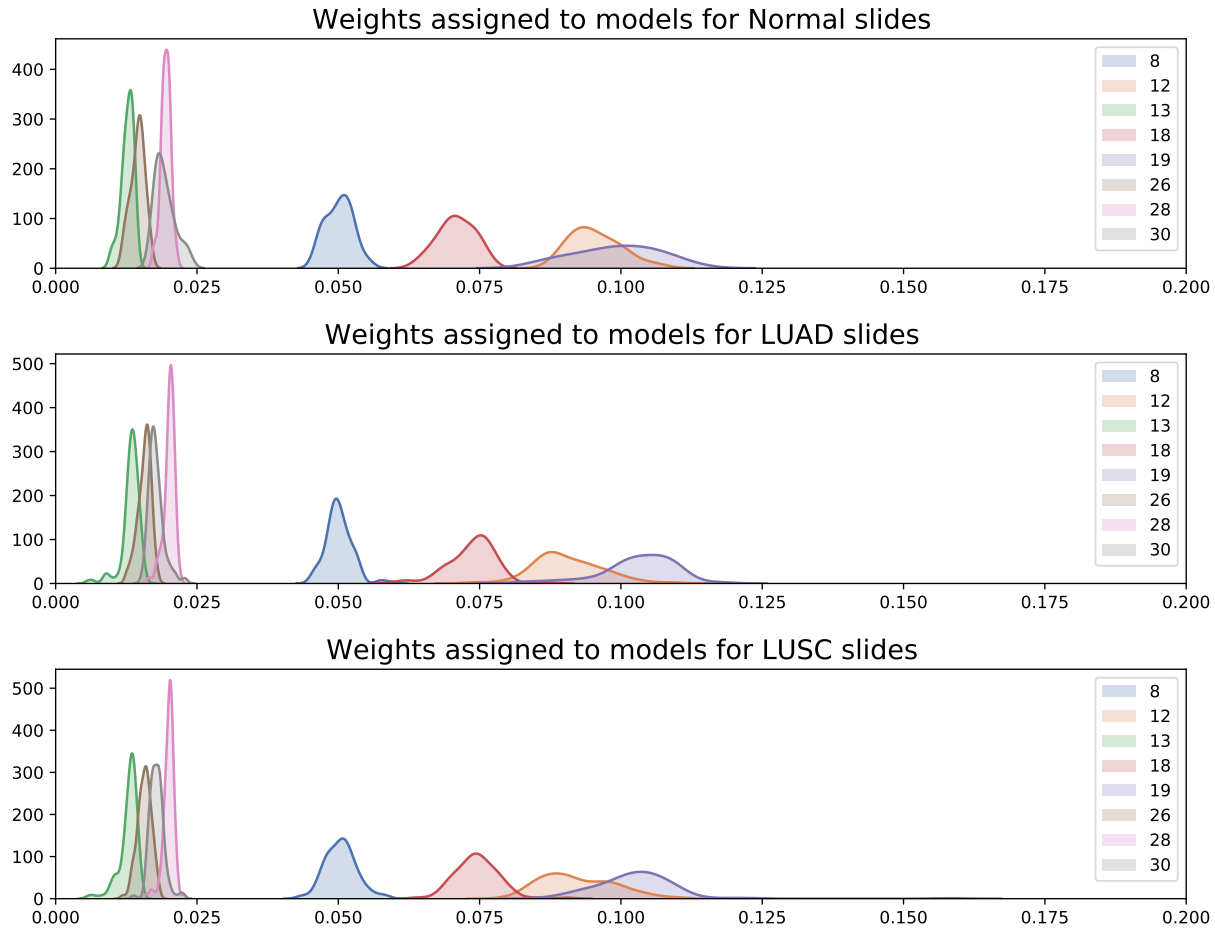

Figure 4: Class-conditional distributions over the weights assigned to different archetypes by CEN, visualized for healthy, LUAD, and LUSC slides. Observe only slight variation between distributions of the weights assigned to different archetypes, while the relative ordering between the archetypes being strictly preserved.

on transcriptomic information.

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 (Section 5.3) would have been confounded by the assignment of different models to different classes. As shown in Figure 4, 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.

We note that while class-conditional distributions over the discovered subtypes are nearly identical, as shown in Figures S2, S3, S4, S5 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.

6 Discussion

In this work, we have used contextual deep learning to estimate discriminative subtypes of lung cancers. These discriminative subtypes improve prediction of cancer type from transcriptomic data by allowing the sample-specific

models to pay attention to different biological processes for different samples.

For machine learning practitioners, this has several implications. Firstly, multi-modal data analysis is critically important in healthcare applications (often because the technical noise in each measurement can only 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 which can be used to design these interpretable multi-modal approaches. By allowing one data modality to select the model to be used on another modality, we can achieve both accuracy and interpretability. This is extremely important for practitioners who want their machine learning models to be used in areas such as healthcare and public policy.

In addition, this work has clinical implications. Firstly, because the discriminative subtypes correspond to biologically-meaningful 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, in this study, a deep learning model trained on images was able to assign meaningful contexts for transcriptomic models, indicating that 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 [38, 81, 82], and we are excited to see future works improve on the connection between these datatypes to discover and analyze morphological patterns of cellular modifications. Future work in this area will enable the machine learning community to assist pathologists. Lastly, this approach of patient-specific modeling by contextualization has clinical relevance because it suggests that the inter-sample heterogeneity which hampers clinical predictive power may be overcome by more flexible modeling. Future work in this area may lead to machine learning practitioners being able to assist clinicians by providing patient-specific actionable insights about the prognosis of their patient's disease.

Limitations. While our work advances our understanding of machine learning and healthcare, we note some important limitations. First, we only considered transcriptomic and histopathology data in our model. It would be beneficial to add known clinical confounders for lung cancer to our model such as smoking, age, and environment to add more relevant prognostic information to our model and enable prediction of more clinically-relevant outcomes such as treatment responses. Second, the labels for our histopathology data relied on the whole slide level labels provided by TCGA. While other works [34] have also used this labeling scheme, it adds noise to the labels because not every patch within a cancerous slide depicts cancerous tissue. Third, the distribution of healthy, LUAD, LUSC samples were imbalanced because we only used samples with both transcriptomic and histopathologic data and TCGA collected less transcriptomic data from healthy subjects. Future work should consider validating these results with an external balanced dataset. Lastly, we have focused on only one prediction task of classification. While this highlights the promise of contextual deep learning, we are excited for future work to use our approach to predict stage, survival, and optimal treatments.

Despite these limitations, our results suggest that with contextual deep models we can learn a context called a discriminative subtype from histopathological data and this discriminative subtype can produce sample-specific transcriptomic models which accurately classify LUAD, LUSC, and healthy tissue. Our work takes another step towards creating patient-specific interpretable predictive models of disease. We look forward to future work to expand on the use of contextual deep models to learn from multi-modal patient data.

7 Conclusions

In this paper, we have demonstrated the usefulness of estimating *discriminative subtypes* of lung cancers. These discriminative subtypes are latent variables which provide parsimonious descriptions of many aspects of a patient's pathology. We call these subtypes *discriminative* because they are used as contexts under which to make downstream predictions. We find that contextual modeling with discriminative subtypes improves classification accuracy in the prediction of different types of lung cancer. In addition, we are able to interpret the model archetypes through enrichment of biological processes. The archetypes are enriched for a variety of distinct processes, suggesting that different lung cancer samples should be classified according to different reasons. We also confirm that the selection of these archetypes is not based on the cancer type but rather provides meaningful context for our simpler models to then accurately label sample types. Finally, we link the biological processes to morphological patterns by identifying the patches which maximize the selection of each archetypal model. This research demonstrates the utility of jointly

modeling histopathology and transcriptomic data, and underscores the potential for histopathology images to be used for discovery of fine-grained subtypes. We look forward to future work to expand on these relationships between macroscopic morphological patterns and microscopic cellular processes.

Acknowledgements

The authors thank Ziv Bar-Joseph and Kayhan Batmanghelich for helpful discussions and comments on an early draft of the manuscript. This work was supported in part by NIH R01GM114311. B.L. was supported in part by the CMLH Fellowship. M.A. was supported in part by the Google PhD Fellowship. A.A. was supported in part NIH grants 1R01GM122096 and OT2OD026682. J.W. was supported by NIH T32 training grant T32 EB009403 as part of the HHMI-NIBIB Interfaces Initiative.

References

[1] 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.

[2] 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.

[3] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lonning, and A. L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–74, 2001.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

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

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Benjamin J Lengerich, Bryon Aragam, and Eric P Xing. Personalized regression enables sample-specific pan-cancer analysis. *Bioinformatics*, 34(13):i178–i186, 2018.

[29] Benjamin J. Lengerich, Bryon Aragam, and Eric P Xing. Learning sample-specific models with low-rank personalized regression. In *Advances in Neural Information Processing Systems (In Press)*, 2019.

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

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

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

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

[47] 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.

[48] 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.

[49] Thanos P Mourikis, Lorena Benedetti, Elizabeth Foxall, Julianne Perner, Matteo Cereda, Jesper Lagergren, Michael Howell, Christopher Yau, Rebecca Fitzgerald, Paola Scaffidi, et al. Patient-specific detection of cancer genes reveals recurrently perturbed processes in esophageal adenocarcinoma. *bioRxiv*, page 321612, 2018.

[50] Adham Beykikhoshk, Thomas P Quinn, Samuel C Lee, Truyen Tran, and Svetha Venkatesh. Deeptriage: Interpretable and individualised biomarker scores using attention mechanism for the classification of breast cancer sub-types. *bioRxiv*, page 533406, 2019.

[51] Kristina L Buschur, Maria Chikina, and Panayiotis V Benos. Causal network perturbations for instance-specific analysis of single cell and disease samples. *bioRxiv*, page 637710, 2019.

[52] Fattaneh Jabbari, Shyam Visweswaran, and Gregory F Cooper. An empirical investigation of instance-specific causal bayesian network learning. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2582–2585. IEEE, 2019.

[53] Maruan Al-Shedivat, Avinava Dubey, and Eric P Xing. Contextual explanation networks. *arXiv preprint arXiv:1705.10301*, 2017.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[79] Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.

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

[81] 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*, page 458711, 2018.

[82] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vohringer, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *bioRxiv*, page 813543, 2019.

S1 Implementation Details

To allow reproducibility of the results, we describe in this section the implementation details. The code to reproduce our analysis will be also made publicly available.

Data. The data was retrieved from TCGA on April 16, 2019 and preprocessed using the pipeline released by [34]. The images are subject to a non-deterministic data augmentation similar to the one implemented in [34]. We were also using mean centered log-genomics features keeping only those having a standard deviation greater than 1e-2, which result in selection of 1695 genes.

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 $\rho = 0.9$, $\mathrm{momentum} = 0.9$, $\epsilon = 1.0$. The choice of the checkpoints is based on a best validation loss criterion.

Table S1: Most influential genes in each case/control archetypal models.

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 S2: Most influential genes in each LUAD 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 S3: Most influential genes in each LUSC 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 S4: Terms enriched in LUAD archetypal models ($p < 0.05$).

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

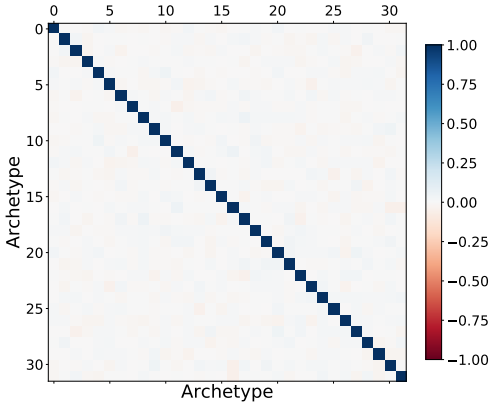

Figure S1: Kendall Tau similarity of ranks of transcripts selected by each entry in the dictionary. Nearly all off-diagonal comparisons have similarity less than 0.1, indicating that the models are orthogonal.

Table S5: Terms enriched in LUSC 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

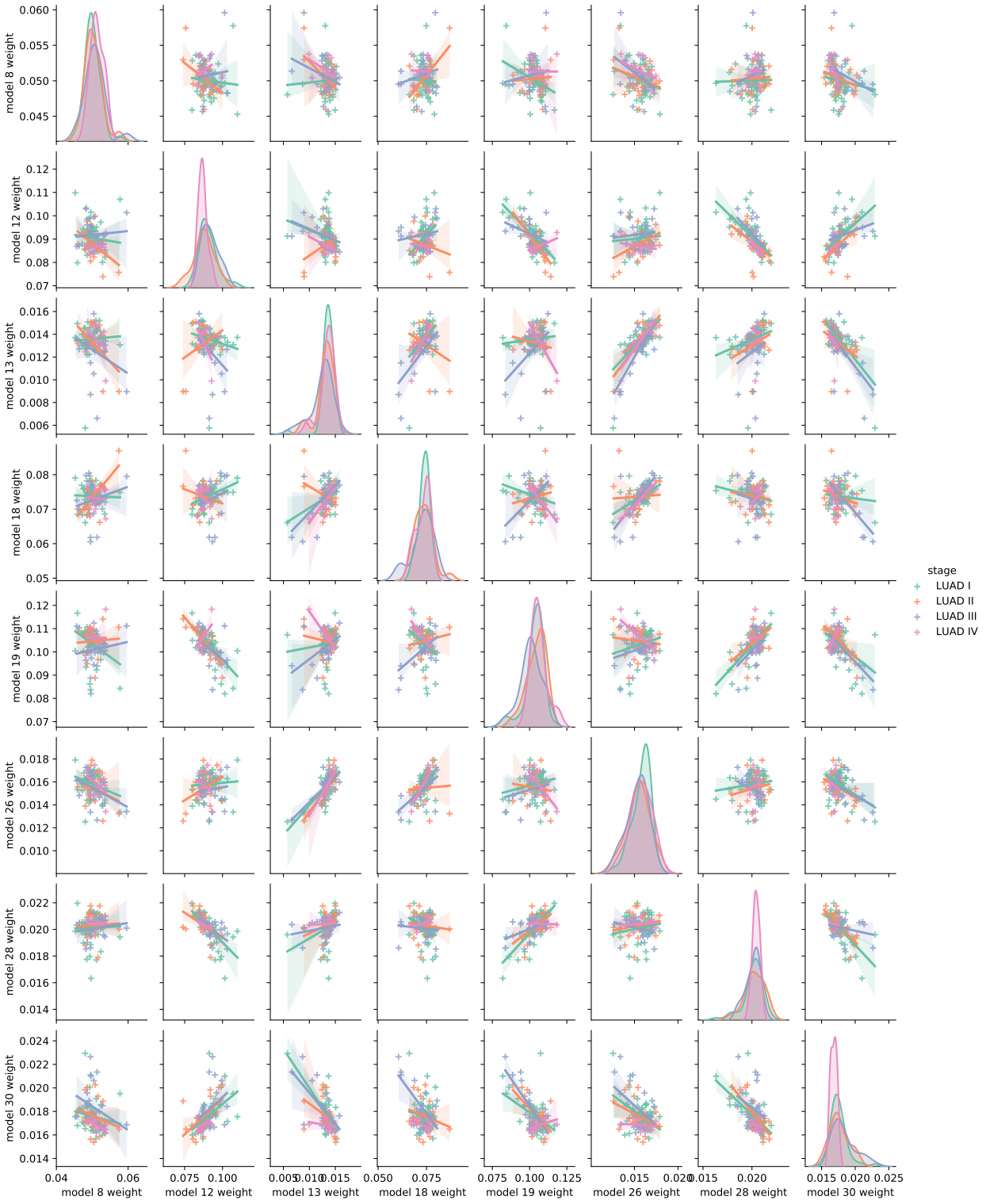

Figure S2: Distribution of weights assigned to 8 different models in the dictionary for different stages of LUAD. Densities on the diagonal indicate variation of the assigned weights for each stage of LUAD. Off-diagonal plots showcase pairwise dependencies between model weights for different stages of LUAD.

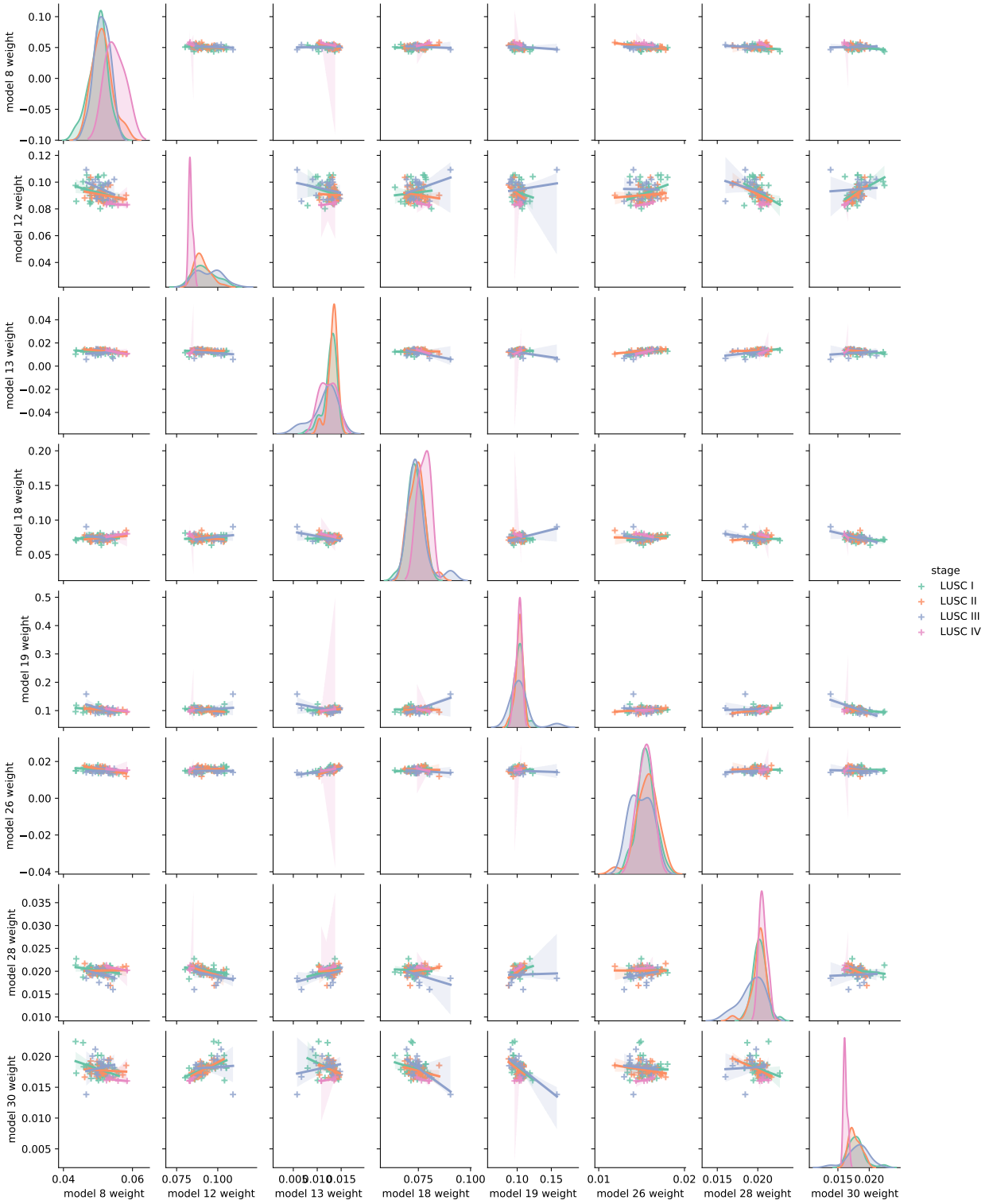

Figure S3: Distribution of weights assigned to 8 different models in the dictionary for different stages of LUSC. Densities on the diagonal indicate variation of the assigned weights for each stage of LUSC. Off-diagonal plots showcase pairwise dependencies between model weights for different stages of LUSC.

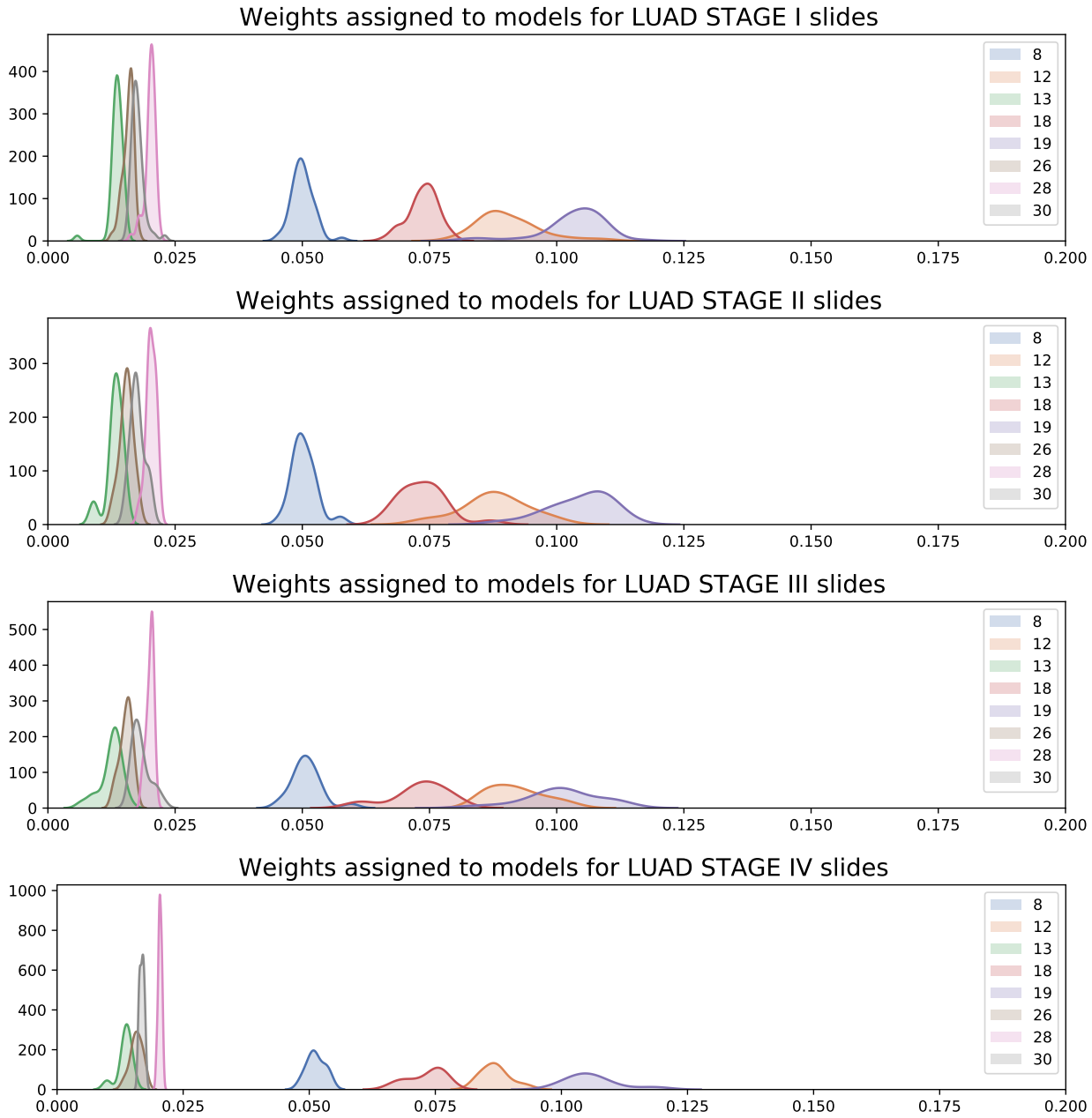

Figure S4: Distributions over the weights assigned to different models in the CEN dictionary visualized for LUAD slides of stages I-IV.

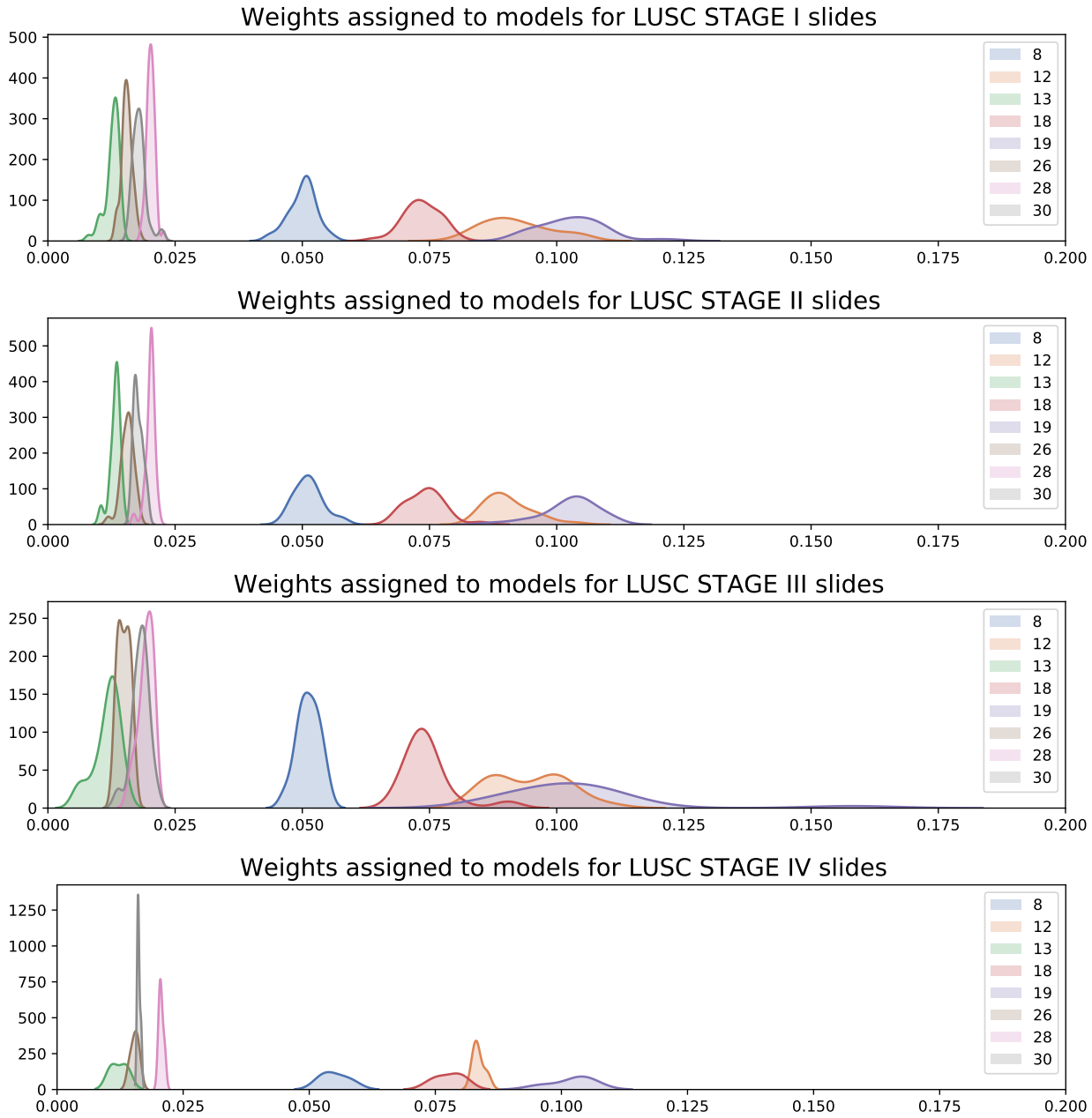

Figure S5: Distributions over the weights assigned to different models in the CEN dictionary visualized for LUSC slides of stages I-IV.

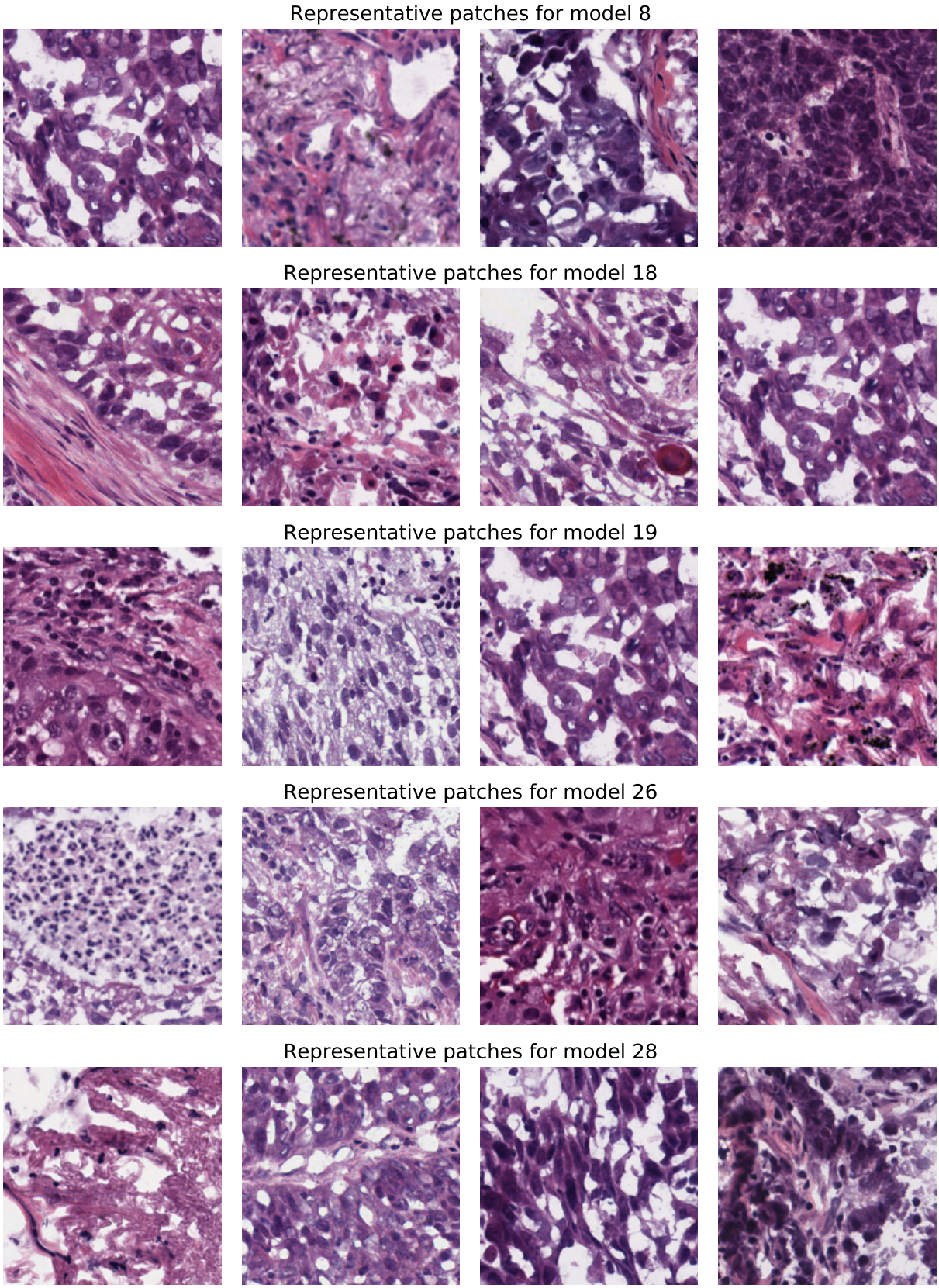

Figure S6: Top four patches that made CEN assign the highest weights to the corresponding models 8, 18, 19, 26, 28 in the dictionary.