Deep Learning Features Encode Interpretable Morphologies within Histological Images

Ali Foroughi pour
Jackson Laboratory
Brian White
Jackson Laboratory
Jonghanne park
Jackson Laboratory
Todd Sheridan
Jackson Laboratory
Jeffrey Chuang (jeff.chuang@jax.org)
Jackson Laboratory

Research Article

Keywords:

Posted Date: February 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-865341/v2

DEEP LEARNING FEATURES ENCODE INTERPRETABLE MORPHOLOGIES WITHIN HISTOLOGICAL IMAGES

BY ALI FOROUGHI POUR[1] AND BRIAN WHITE[1] AND JONGHANNE PARK[1] AND TODD B. SHERIDAN[1,2] AND JEFFREY H. CHUANG[1,3]

[1]*The Jackson Laboratory for Genomic Medicine, 10 Discovery Dr. Farmington, CT, 06032* *ali.foroughipour@jax.org* *jeff.chuang@jax.org*

[2]*Department of Pathology, Hartford hospital, 80 Seymour St, Hartford, CT, 06106*

[3]*Department of Genetics and Genome Sciences, UCONN Health, Farmington, CT, 06032*

ABSTRACT. Convolutional neural networks (CNNs) are revolutionizing digital pathology by enabling machine learning-based classification of a variety of phenotypes from hematoxylin and eosin (H&E) whole slide images (WSIs), but the interpretation of CNNs remains difficult. Most studies have considered interpretability in a post hoc fashion, e.g. by presenting example regions with strongly predicted class labels. However, such an approach does not explain the biological features that contribute to correct predictions. To address this problem, here we investigate the interpretability of H&E-derived CNN features (the feature weights in the final layer of a transfer-learning-based architecture). While many studies have incorporated CNN features into predictive models, there has been little empirical study of their properties. We show such features can be construed as abstract morphological genes ("mones") with strong independent associations to biological phenotypes. Many mones are specific to individual cancer types, while others are found in multiple cancers especially from related tissue types. We also observe that mone-mone correlations are strong and robustly preserved across related cancers. Importantly, linear mone-based classifiers can very accurately separate 38 distinct classes (19 tumor types and their adjacent normals, AUC=97.1% $\pm$ 2.8% for each class prediction), and linear classifiers are also highly effective for universal tumor detection (AUC=99.2% $\pm$ 0.12%). This linearity provides evidence that individual mones or correlated mone clusters may be associated with interpretable histopathological features or other patient characteristics. In particular, the statistical similarity of mones to gene expression values allows integrative mone analysis via expression-based bioinformatics approaches. We observe strong correlations between individual mones and individual gene expression values, notably mones associated with collagen gene expression in ovarian cancer. Mone-expression comparisons also indicate that immunoglobulin expression can be identified using mones in colon adenocarcinoma and that immune activity can be identified across multiple cancer types, and we verify these findings by expert histopathological review. Our work demonstrates that mones provide a morphological H&E decomposition that can be effectively associated with diverse phenotypes, analogous to the interpretability of transcription via gene expression values. Our work also demonstrates mones can be interpreted without using a classifier as a proxy.

1. Introduction. Deep learning has become an important methodology for analyzing biomedical images, and in particular for analyzing hematoxylin and eosin (H&E) stained whole slide images (WSIs). Deep neural networks have achieved classification accuracies higher than classical machine learning models [1]. However, they are black-boxes that do not directly reveal the morphological features they associate with labels, a significant concern for mechanistic analysis and clinical decision making [2]. Identification of biologically meaningful morphological features may be confounded by image artifacts [3], such as

blurring, noise, and lossy image compression [4]. Tissue damage, image quality, and dataset-specific artifacts have also been suggested to affect feature representation and prediction accuracy of neural networks (see [1], [5], and [6]). Given the impact of such artifacts on deep learning-based predictors, it is of critical importance to be able to decompose CNNs into features that can be biologically interpreted.

The majority of models for visualizing, analyzing, and interpreting CNNs reveal "where" a network is "looking" to make its prediction, rather than revealing "what" information in the region of interest is important. Some methods output pixel patterns that affect the value of a neuron in a deep network (see [7]). However, such techniques tend to output different predictive regions, can be difficult to validate, or have been suggested to be "fragile", i.e. extremely sensitive to small perturbations of the image [8]. Optimizing conventional deep learning techniques, such as self-attention, to identify regions informative of class labels is a current theme in digital pathology (see [9, 10] for examples). While most methods assess deep feature representations as a whole, recent work suggests deep learning features cluster together and encode distinct morphologies [11]. Other recent works have focused on visualizing individual deep learning features as heatmaps [12]. Finally, as the majority of interpretation methods have focused on identifying regions predictive of class labels, they requires a trained classifier and cannot be directly used in pipelines that employ unsupervised feature learning. Unlike natural image analysis [13], biomedical image analysis is complemented by additional data modalities, such as multiplexed imaging, single cell and bulk sequencing, and clinical information citep ray2014information, kong2011integrative. These data may aid in interpreting the deep feature representations of the H&E slide. However, models integrating these diverse modalities are needed. The feasibility of doing so is supported by work establishing the connection between modalities, for example by using CNNs to predict expression values of specific genes from H&E images (see [14, 15, 16] for examples). Because of the architectural complexity of CNNs, it has often been assumed that CNN-based decompositions of images into features are not interpretable. However, there has been little empirical study of this question, e.g. by testing whether CNN-derived features are correlated with simple biological features such as gene expression values. In this work, we investigate the interpretability of CNN-derived image features. Prior works ([17, 1]) have referred to these by various names (e.g. features, fingerprints) whose use is not specific to biological image analysis. For clarity and because they represent morphological features in many ways analogous to genes, we refer to them as mones (i.e. "morphological genes"). We find that mones share statistical similarities with gene expression data, and hence, a mone can be conceptualized as an abstract gene with some expression value. Individual mones have strong linear associations with phenotypic features, making them directly interpretable, which we demonstrate in several analyses. We demonstrate that many mones can distinguish cancer tumors from adjacent normal slides. These mones can be linearly combined for reliable prediction of both pan-cancer and tissue-specific cancer phenotypes. Mone-mone correlation analysis identifies clusters of highly correlated mones within cancer types, and these correlations are strongly preserved among cancers from related tissues. The similarity of mone values to gene expression data allows immediate use of many interpretable bioinformatics tools and machine learning models to identify the underlying biology of morphologies encoded by CNNs. For example, integrative mone-gene expression correlation analysis reveals that collagen content and immune infiltration are linearly associated with morphologies encoded by mones in several cancer types, and we confirm these relationships by expert histopathological review. Our studies confirm individual deep learning features encode distinct and identifiable morphology, and demonstrate the power of mones for computationally deconstructing cancer images into interpretable biological features.The linear analysis of individual deep learning features versus expression values or interpretable morphology provides a simple and effective approach to interpreting deep learning models in biological image analysis, notably without the need for a trained classifier.

2. Results. We analyzed the InceptionV3 [18] features of tiles from whole slide images of The Cancer Genome Atlas (TCGA) [1] for 19 cancers (see Supplementary File 1 for the full list). We used features derived from this architecture because predictive models based on Inception have shown high accuracy for identifying phenotypes in prior studies [1]. We hereafter denote each of the 2048 outputs of the global average pooling layer of the Inception V3 network as mones (morphological genes). We use this terminology because mones have analogies to genes with individual expression values. Tile level mones can be combined to construct slide level mones (see methods). Unless otherwise stated, in the studies below "mone" refers to a slide level characterization. Figure 1 provides an overview of interpretive mone analyses and their connection to current interpretation techniques in the field.

2.1. *Individual mones differentiate phenotypes.* We first investigated to what extent individual mones can differentiate phenotypes, focusing on TCGA tumor/normal slide comparisons. We initially identified individual mones with significant differences in distribution between breast cancer (BRCA) tumor and adjacent normal slides (>1800 mones were statistically significant with FDR$< 5\%$). A clustermap of the top 100 such mones was able to clearly separate these classes (see methods, Figure 2a, and Supplementary Figure 1, clustermap AUC=89%, rand score =96%, and adjusted rand score=85%). These results were typical of mone behaviors in many tumor types – in any given tumor type, many mones were able to separate frozen tumor from normal slides (see methods and Supplementary File 1). We applied several statistical methods to test the robustness of such mones. In each cancer, at least 40% of mones were statistically significant irrespective of the statistical test used (FDR$= 5\%$, see Supplementary File 1), and 75.7% of mones significant by at least one method were significant by all tests (see methods, see Figure 2b). A smaller subset of these mones showed strong effect sizes, as identified by optimal Bayesian filter (OBF) [19] statistics (see Methods). 22% of all mone-cancer pairs met this criterion based on distributional differences between tumor and normal slides (see Supplementary Figure 2). As an example, mone 983 is a tumor marker with a distributional difference between frozen tumor and adjacent normal breast cancer (BRCA) slides (Figure 2c). It also behaves similarly in several other tumor types (Supplementary Figure 3). It strongly correlates with cell density in frozen BRCA slides (see methods, see Figures 2d-e, Pearson r=0.69, p-value<1e-200, see Supplementary File 2) and significantly though with moderate magnitude in FFPE BRCA tumor slides (Pearson r=0.24, p-value=1.9e-13). Mone 983 has higher correlation with cell density in FFPE LUAD slides (Pearson r=0.61, p-value=2.2e-34) than frozen LUAD slides (Pearson r=0.28, p-value=4.1e-18, see Supplementary File 2). To further test whether mones exhibit consistent behavior in different cancer types, we analyzed the four cancer families of [1]: pan-GYN, pan-KIDNEY, pan-LUNG, and pan-GI (see Supplementary File 3 for cancers in each family). More than half of the mones with distributional differences in each cancer type also have distributional differences in all cancers of the family (see Supplementary File 3). Although cancer-specific mones are uncommon, such mones still show clear distributional differences between tumor and normal (see Supplementary Figure 4). Interestingly, most mones distinguishing tumor from normal in both LUAD and LUSC also distinguish between the LUAD and LUSC cancers (see Supplementary File 3), suggesting quantitative values are important. Individual mones can also distinguish frozen from FFPE slides (see Supplementary File 4). We observed the ability of some mones to distinguish between tumor and normal is impacted by differences between frozen and FFPE modalities (421 ± 112 across all cancers, see methods, see Supplementary File 4), though the majority of mones behave similarly in frozen and FFPE.

2.2. *Mone clusters provide robust encodings of cancer phenotypes.* We next investigated to what extent mones are independent or encode behaviors together. We did this by calculating pairs of mones that were significantly correlated, for each cancer type. We analyzed this first by restricting to frozen slides (to avoid Simpson's paradox), and second by combining frozen and FFPE slides (to avoid false correlations due to frozen-specific artifacts). The results were generally robust between the two methods—a few cancers had a non-trivial difference in the ratio of correlated mone pairs ascertained by the two methods (ESCA, KICH, SARC, and PRAD $22.3\% \pm 2.22\%$), while the remaining cancers had small differences ($7.8\% \pm 3.9\%$, see Supplementary File 6). Therefore we subsequently analyzed frozen samples only unless otherwise specified. Overall, we found that correlated mones are prevalent among tumor slides (see Figures 2f and 2g, and Supplementary File 6). For example, 83.9% and 88.7% of mone-mone pairs are correlated within LUAD and within LUSC, respectively (see methods and Supplementary File 6). We observed similar results for other cancers: $68.8\% \pm 13\%$ of all mone-pairs were statistically significant across cancers (see Supplementary File 6). Remarkably, we observed that pairwise correlations are preserved within cancer families – more than 45% of mone-pairs statistically significant in one cancer are significant in all cancers of the family (pan-GYN, pan-KIDNEY, pan-LUNG, and pan-GI, see Supplementary Figure 5). For example, the mone-mone correlations in lung adenocarcinoma and lung squamous cell carcinoma are nearly identical (Figures 2f and 2g). We also calculated mone-mone correlations in the normal slides associated with each cancer type, hypothesizing that the difference in mone-mone correlations between tumor and normal might be important to distinguishing tumor from normal images. However, these differential mone correlations are weaker and less preserved across cancers (see Supplementary File 6 and Supplementary Figure 5).

Different cancers can be distinguished by different sets of mones. For example, while mone 983 separates tumor and normal slides of BRCA, it does not differentiate COAD tumor and normal slides (see Supplementary Figure 3). We identified 105 mones that are highly correlated with mone 983 in BRCA (Null: $|r| \leq 0.5$, FDR$= 0.1\%$, see methods), among which 52, 14, and 29 also differentiate tumor from normal slides in COAD, READ, and STAD, respectively (t-test FDR<0.1%). The first principal component of these COAD-overlapping mones (explaining 41% of the variance across COAD samples) strongly correlates with cell density in frozen slides (Cellpose cellularity estimates: Pearson r=0.22, p-value=2.2e-11, HoverNet cellularity estimates: Pearson r=0.43, P-value=1.9e-47, see Methods and Supplementary File 2). Thus the high cellularity in BRCA [20, 21] and COAD [22, 23] involve incompletely overlapping mone sets.

2.3. *Linear models of mones can detect and distinguish tumors.* We investigated linear models of mones for predicting phenotypes, as they allow direct interpretation of mone values. We observed that linear models of mones can efficiently distinguish tumor from adjacent normal slides, as well as the cancer type from which they are derived (19 cancers, 38 classes, see methods, see Figure 3a-c). We tested two linear models, multi-class linear discriminant analysis [MLDA, One versus Rest (OVR)-AUC=$97.1\% \pm 4.6\%$, see Supplementary File 7] and multinomial logistic regression with LASSO penalty (LR-LASSO, OVR-AUC=$97.1\% \pm 4.2\%$, see Supplementary File 7). MLDA encodes mone patterns indicative of class labels into a low dimensional space (i.e. the number of classes - 1), yielding t-SNE visualizations with improved interpretability over naïve t-SNE (compare Figures 3a,b and Supplementary Figures 6a,b). LR-LASSO, on the other hand, is a linear model based on small mone sets, so its regression coefficients can be interpreted directly with risks incurred by each mone. Although CNN methods typically use difficult-to-interpret fully connected layers at the classification step, we found that efficiently designed linear models

can replace fully connected layers while still achieving high prediction AUCs. Combining tumor probabilities of the LR-LASSO classifier we obtained a universal tumor detector with extremely high AUC ($99.2\% \pm 0.12\%$, see methods), out-performing the fully deep learning model of [24] (Reported AUC=0.95 ± 0.02). Furthermore, LR-LASSO is effective at cross-classification similar to CNNs with fully connected classification layers [1], i.e., LR-LASSO trained to distinguish tumor/normal for one cancer type can distinguish tumor/normal for other cancer types as well (Figure 3d and Supplementary Figure 6c). While the LR-LASSO model has smaller average AUC (0.84) compared to the fully deep learning model of [1] (0.88), logistic regression is more interpretable than a multi-layer perceptron. Our slide level tumor detectors also produce meaningful tile level predictions. Independent review by our pathology team supports most tumor regions having high tumor probability, and most non-tumor regions have low tumor probability in these images, with the cases of misclassification tending to be prediction of non-tumor regions to be tumor (see Supplementary Figure 7 for examples). These results indicate that the LR-LASSO slide level tumor markers are effective at the tile level.

2.4. *Mones have interpretable correlations with gene expression.* We next investigated whether mones are linearly associated with gene expression values, as this could provide transcriptional interpretability for mones (see Methods for data stratification and pre-processing). We observed many mones significantly correlated with individual genes. Across five cancers analyzed (OV, COAD, KIRC, LUAD, and LUSC) between 83 (LUSC) and 1797 (KIRC) mones were associated with at least one gene, whereas between 332 (LUSC) and 16474 (KIRC) genes were associated with at least one mone (methods, Supplementary File 8). We then analyzed several cases of particular interest.

2.4.1. *Mones encode collagen content.* We first used unsupervised analysis to study clusters of correlated mone-gene pairs. We identified a cluster of highly correlated mones and collagen genes in OV (Figure 4a). These mone values can be efficiently combined for association with phenotype using PCA (PC-1 explains 63% of the variance). Histopathological review by our pathology team confirmed tiles with high PC-1 values as typically rich in collagen (see Figure 4d and Supplementary Figure 8), and tiles with low PC-1 values as having low collagen but increased cellularity. These mone-gene associations may be clinically relevant, as high expression of collagen genes correlates with multi-drug resistance [25] and poor prognosis in ovarian cancers [26]. Mone 1062–one of the mones in the identified cluster– was additionally highly correlated with ECM2, THBS1 and THBS2, which have been suggested to play a significant role in ovarian cancer drug resistance and metastasis [26, 27, 28].

2.4.2. *Mones encode immune infiltration.* Supervised correlation analysis using fixed gene sets can also be used to test if mones encode morphological features associated with a biological phenotype. We tested whether mones encode immune infiltration in pan-GI (COAD, READ, and STAD) and pan-LUNG (LUAD and LUSC) cancers, with immune related gene sets taken from [29] and [30] (see Methods and Supplementary File 9). We identified 19 mones significantly correlating (some positive, some negative) with immune related genes in pan-GI cancers (Figure 4b and Supplementary File 10). These mones have significant correlations with each other ($|r|=0.56 \pm 0.13$), and 14 mones significantly correlate with more than one immune gene. LAIR1, LCP2, MS4A4A, and CCR1 correlate with 15, 9, 12, and 15 mones, respectively. All 19 mones are significantly correlated with prior TCGA estimates of leukocyte fraction [31] (FDR<0.05,$|r|=0.29 \pm 0.07$) and HoverNet estimates of immune cell quantity from the H&E images (see methods, FDR<0.05, $|r|=0.45 \pm$

0.12) in COAD. Mone 179 had strong positive correlations with both leukocyte fraction (r=0.2) and HoverNet estimates (r=0.65). Histopathological review confirmed that mone 179 differentiates amongst COAD tiles according to their level of immune infiltration (see Figure 4e and Supplementary Figure 9). PC-1 of the 19 mones has significant correlation with leukocyte fraction (r=0.4, p-value=2.4e-35) and HoverNet estimates (r=0.17, p-value=2.5e-17). Histopathological review validated that PC-1 also differentiates COAD tiles according to differing levels of immune infiltration (see Supplementary Figure 10). Thus PC-1 efficiently combines mones and provides a stronger separation than individual mones. We observed similar correlations, but smaller in magnitude, between mones and local immune cytolytic activity genes in lung cancers (Figure 3c and Supplementary File 10). We observed stronger correlations in LUAD than LUSC (LUAD $|r|=0.20 \pm 0.05$, LUSC $|r|=0.12 \pm 0.05$). We identified 31 mones significantly correlating (some positive, some negative) with immune related genes in LUAD. 7 mones correlated with at least 3 genes (see Figure 4c). Only one LUAD immune mone did not correlate with lymphocyte fraction (FDR=0.05). PC-1 of the 31 mones correlates with lymphocyte fraction (r=0.28, p-value= 2.2e-4). Histopathological review of LUAD slides based on PC-1 suggests LUAD tiles with high PC-1 show a strong tumor infiltrating lymphocyte presence and have inflammation, while tiles with low PC-1 are typically not inflamed and show weaker immune infiltration (Figure 4f, Supplementary Figure 11).

2.4.3. *Mones identify immunoglobulin gene expression in highly cellular colon adenocarcinoma tumors.* Supervised correlation analysis can also clarify finer behaviors within WSIs. For example, highly cellular COAD tumors typically show high expression of immunoglobulin (IG) genes. We considered the 52 mones correlated with mone 983 and which differentiate COAD tumor from normal (see section 2.2). The PC-1 dimension of these mones significantly correlates with 87 genes in COAD (FDR<0.05, see Supplementary File 11 for the full gene list), as well with the average expression of these 87 genes (r=0.33, p-value=3.8e-12). Immunoglobulin (IG) genes dominate this gene set, and we define their average expression as a sample's IG score. B-cells express immunoglobulin, and as expected we observed a statistically significant correlation between the log normalized B-cell estimates of [31] and IG score (p-value=2.4e-12). Interestingly, however, B cells comprise only a small fraction of cells in each sample ($0.86\% \pm 1.5\%$), and we did not observe a significant correlation between the B-cell estimates and mone PC-1 (correlation coefficient=-0.009, P-value=0.85). Thus mone analysis suggests that IG expression is not due solely to B-cells. This supports recent studies suggesting colon cancer cells themselves express IG genes (see [32]).

2.5. *External Validation on CPTAC.* We externally validated mone patterns on CPTAC-LUAD and CPTAC-LUSC. Mone 983 correlates with cell density in CPTAC-LUSC (r=0.41, p-value=1.1e-40, Supplementary Figure 13) and CPTAC-LUAD (r=0.21, p-value=1.8e-12, Supplementary Figure 13). Pathologist evaluation confirmed the ability of the PC-1 dimension of TCGA-LUAD immune mones (see section 2.4.2) to separate CPTAC-LUAD slides by immune activity. We also observed that HoverNet frequently mislabels lymphocytes as dead cells (Supplementary Figure 13). Therefore, mone analysis has more robust behavior on CPTAC-LUAD than HoverNet (Supplementary Figure 13).

3. Discussion. Mone analysis provides interpretability of deep neural networks through the linear correlations of deep learning features against phenotypes and gene expression profiles. We have empirically shown that mones efficiently encode strong morphological features that can often be used to replace multi-layer perceptrons with robust and interpretable

linear classification models. Mone-analysis is flexible, can be used in a diverse set of interpretation modalities, and can be applied to features engineered through various training methodologies. Moreover, we have demonstrated that integrative mone-gene correlation analysis can identify specific transcriptional processes from images, and verified these through expert pathological review.

3.1. *Mones provide image-based interpretability.* Linear models based on mones have several empirical and theoretical strengths for image analysis. Individual mones and small mone clusters directly correlate with phenotypes (see section 2.1 and 2.3), enabling a simpler interpretation of CNNs compared with methods that integrate all CNN features together. Most interpretation models assume deep feature representations are complex and non-linear, and therefore provide interpretability primarily through example regions identified by problem-specific classifiers. On the other hand, our results demonstrate CNNs decompose H&E stained images into interpretable features that linearly correlate with phenotypes, and the highly non-linear feature representation assumption can be relaxed for interpreting CNNs trained on H&E slides.

Some prior linear analyses of deep learning features have enabled partial interpretation of CNNs trained on one cancer type (see [12] and [33]), but mones facilitate pan-cancer interpretations. For example, our results demonstrate a pan-cancer mone can encode conserved morphological features across multiple cancer types (see section 2.1), and a conserved morphological feature can be encoded by distinct mones in different cancer types (see section 2.2). Moreover, mones within strongly correlated clusters can be linearly combined to better identify shared encoded morphology (see section 2.2 and [11]).

Mone analysis improves low dimensional visualization of large image sets, such as via t-SNE plots. Pre-trained CNNs are universal feature extractors that encode morphological features predictive of a multitude of labels (see [1], [34], and [35]). Because not all high-dimensional complex relations can be easily embedded in 2D, mone analysis can be used as a first phase of targeted search to identify the mones relevant to classes of interest (see section 2.3), allowing more accurate visualization of class separations in mone space.

Mone analysis guides classifier design by exploring statistical properties of the learned features. For example, linear associations of correlated mone clusters with phenotypes suggests utility of sparse linear models for reliable classification (see sections 2.3 and 2.5). Robustness of linear models using a small number of mones (see sections 2.3 and 2.5) provides empirical evidence for the theoretical results establishing robustness of sparse deep learning models ([36],[37],[38]). Correlation analysis of mones with gene expression values is a powerful approach for interpreting mones. We identified clusters of highly correlated mone-gene sets, demonstrating clear connection of mones to the underlying genetics. Some recent studies have used exhaustive sets of deep learning features to predict expression profiles (see [14, 15, 16]), but our work shows that small mone-gene clusters can be sufficient and provide simpler interpretability. Both supervised and unsupervised analyses identify meaningful clusters (see section 2.4). Supervised analysis using fixed gene sets is particularly interesting, as it enables direct assessment of genes of interest via mones (see section 2.4.2).

3.2. *Interpretation without classifier training.* A notable advantage of the mone approach is that it does not require a trained classifier, which is especially desirable when feature engineering and classification are decoupled (e.g. transfer learning, unsupervised learning). Interpretation models that require a trained classifier are restricted to the morphologies that are predictive of a predetermined set of classes and are utilized by the classifier. Additionally, different classification architectures applied to a fixed learned feature space may use different features and morphologies for predicting a class label. Mone analysis does not require a

classifier to determine if the learned features encode a given phenotype (sections 2.4.2 and 2.4.3). It can be immediately applied to any learned feature space irrespective of training methodology, and can be used in an unsupervised fashion to identify encoded morphologies (section 2.4.1).

Optimization of training classifiers that are robust to stain differences across datasets is an open research question. However, an advantage of mone analysis is that it can be more robust to stain differences than pure classification models, as we observed for cell classification (see section 2.5). It can also be used to analyze how stain differences affect feature representations and a classifier's ability to make reliable predictions.

3.3. *Mone analysis across architectures and data modalities.* Our analyses demonstrate that mones provide an efficient and interpretable CNN embedding of image data, but a caveat is that they have been restricted to Inception v3 mones. Architectures with fully connected layers tend to increase non-linearity in feature representations. Therefore, models that do not utilize multiple sequences of fully connected layers, such as Inception, are more appropriate for linear mone analysis. For example, recent work suggests a small subset of VGG19 features may also be interpreted via their direct association with phenotypes [12]. However, we believe Inception V3 mones are more appropriate for linear association studies because they are the direct inputs to the classification layer. A few other studies have explored correlations between deep autoencoder features and gene (see [33]) or protein expression [39] for other architectures, but the relations between mones across architectures remains a broad and open research topic. We have found that Inception V3 mone 983 in BRCA can be reliably estimated via linear models using ResNet152V3 and DenseNet201 mones ($R^2 > 0.95$). Furthermore, we have been able to convert Xception mones to Inception V3 mones using autoencoders with reasonable accuracy ($R^2 \approx 0.5$). Recent work suggests InceptionV3 and ResNet features are almost equivalent [40].These studies suggest that integrative linear mone-gene correlation analysis can be made effective across a range of deep learning architectures.

While this work focuses solely on H&E WSIs, we believe mone-based interpretation will be valuable for extension to other spatial data types. Immunohistochemistry (IHC) images are primed for rapid progress, as recent work has shown that several IHC markers can be virtualized from H&E (see [41] and [42]). Generalizing mone analysis for other data types such as spatial transcriptomics and multi-channel protein data is also an exciting and open area, though new architectures will need to be explored to handle the high dimensionality of such images. The interpretation of CNNs for these image types is a challenging but important task, and we expect that integrative multi-modal multi-architecture mone analysis will be a potent and informative approach.

4. Methods.

4.1. *Data acquisition and pre-processing.* The 20X H&E WSIs of TCGA are pre-processed, tiled, and passed through an InceptionV3 model pre-trained on image-net data in [1]. The cached 2048 global average pooling layer features of InceptionV3 (called mones in this manuscript) were written to disk and analyzed. For each of these 2048 mones, the median value across all tiles within a slide was computed to yield the slide-level mone value.

4.2. *Differential mone analysis.* Differential mone analysis identifies mones with statistically significant distributional differences across classes. Welch's t-test, Kolmogorov Smirnov (KS) test, Wilcoxon Rank Sum (WRS) test, and optimal Bayesian Filter (OBF) (see [19] for details) were used for statistical analysis. t-test, WRS test, and KS test use the Benjamini-Hochberg procedure [43] for FDR correction. The scipy python package [44] was

used to implement t-test, KS test, and WRS test. The statsmodels [45] implementation of the Benjamini-Hochberg procedure was used.

Minimal risk OBF (see [19] for details) identifies mones with posterior probabilities larger than $1-\alpha$, where α is the FDR rate. FDR-OBF (see [46] for details) outputs the feature set that bounds the sample conditioned FDR by α. OBF can report the FDR of any arbitrary feature set. Unless otherwise stated FDR-OBF is used. OBF uses Jeffrey's prior, assumes the prior probability of a mone having distributional differences is 50% (to model no preference on the identity of a mone, i.e., with or without distributional differences across classes), and sets the normalization constant of the prior to 0.1. Mones with distributional differences across classes are hereafter called markers, and mones without distributional differences are called non-markers. The posterior probabilities of OBF can be used to estimate the first two moments (mean and standard deviation) of the number of markers (see [46] for details). Assuming mone identities (marker or non-marker) are independent across cancers, the posterior probabilities can be multiplied to calculate the probability of a joint event.

OBF intrinsically computes the ratio between sample variance and weighted geometric mean of class-conditioned variances, hereafter denoted by $a(m)$ for mone m. Similar to many ANOVA-based analyses this ratio measures distributional differences and is closely related to Bhattacharyya distance (see [47]). It converges to 1 for non-marker mones and converges to larger values for markers. Larger $a(m)$ values denote larger distributional differences. Assuming balanced samples of sizes 200 and 100 we compute the $a(m)$ values resulting in a posterior of 0.95 as thresholds to distinguish moderate [a(m)=1.088] and strong [a(m)=1.159] mones separating tumor from normal slides (see Supplementary Figure 3).

Structured multi-class OBF (see [48] for details) considers the four possible relations (known as structures by OBF) between frozen normal, frozen tumor, and FFPE tumor slides: (A) a mone does not differentiate between slides (prior probability=0.5), (B) a mone has one distribution for frozen slides (both tumor and adjacent normal slides) and another distribution for FFPE slides (prior probability=0.5/3), (C) a mone has one distribution for tumor slides (both frozen and FFPE) and another distribution for frozen normal slides (prior probability=0.5/3), and (D) a mone has one distribution for FFPE tumor slides and frozen adjacent normal slides and another distribution for frozen tumor slides. Mones with structure B for which frozen tumor and FFPE tumors lie on both sides of frozen adjacent normal slides (based on mean values) are considered ineffective due to FFPE/frozen differences.

4.3. *Mone correlation analysis.* For each cancer type, we calculated correlations between mones over all samples of the given cancer type. This analysis was done for each cancer type. We use the Ledoit-Wolf shrinkage [49] implementation of the scikit-learn python package [50] for computing covariance matrices. We then compute the correlation matrix from the covariance matrix. We apply the Fisher transform to correlation coefficients and approximate the null with its asymptotic Gaussian distribution. Benjamini-Hochberg [43] procedure is used for FDR correction. We use seaborn package[51] with default values to generate clustergrams. Correlation matrices are averaged to compute the pooled correlation matrix. Statistically significant mone-mone correlations are referred to as "correlated mone pairs."

Differential mone correlations denote the difference between the correlation coefficient of tumor and normal slides, i.e., for each cancer, differential matrix is computed by subtracting the correlation coefficient matrix of normal slides form the correlation coefficient matrix of tumor slides. Differential mone correlation analysis uses the asymptotic Gaussian distribution of the difference between Fisher transformed correlation coefficients to compute the p-values. Statistically significant differential mone correlations are referred to as "differentially correlated mone pairs."

4.4. *Linear classification models.* We implement MLDA and LR-LASSO using the scikit-learn python package [50] with default values except we set $C = 100$ and use the "saga" solver [52] in non-binary problems for LR-LASSO. We observed little sensitivity of AUCs to the C values ranging from 1 to 1000, and hence use $C = 100$ throughout. We randomly split data to train and test sets 10 times, which are then used to compute the mean and variance of the AUCs. We use the Scipy package [44] to implement the Wilcoxon signed rank test.

4.5. *Cell segmentation and classification.* Cellpose [53] was used to segment and count number of cells in BRCA, LUAD, and COAD tiles. Given the fixed magnification and tile size the number of cells per tile captures tile level cell density. Median number of cells per tile was used as slide level cell density index. HoverNet [54] was used to segment, count, and classify nuclei within COAD tumor slides. HoverNet was executed using the pre-trained PanNuke model ([55]), such that nuclei were classified into one of five types: neoplastic epithelial, non-neoplastic epithelial, connective (including fibroblasts and endothelial), inflammatory (including leukocytes, lymphocytes, and macrophages), and dead nuclei. Median number of cells nuclei across tiles were used as cell density. Median number of predicted inflammatory nuclei across tiles were used to characterize presence of immune cells.

4.5.1. *Integrative mone-gene analysis.* Gene expression data were downloaded from the GDC portal [56]. We only used slide-gene expression pairs where both the slide and the expression profile were from the same vial. Log normalized FPKMs were used. Genes with zero counts in more than half the mone-gene pairs or expression standard deviation below 0.25 were removed. Given a set of mone-gene pairs, we stack the mone and gene vectors and compute the covariance matrix using the Ledoit-Wolf shrinkage method [49] implemented in the scikit-learn python package [50]. Correlation values are computed given the covariance matrix similar to mone correlation analyses. Statistical significance tests are performed similar to mone correlation analyses.

4.5.2. *Immune profiling and analysis.* Leukocyte fractions of TCGA samples were obtained from [31]. All T-cell and B-cell categories were summed to obtain T-cell and B-cell proportions, respectively. The fractions of T-cell and B-cells were summed to obtain lymphocyte fractions. Log normalization of fractions were used throughout. Correlation analysis of immune scores with mones and IG score was performed similar to mone correlation analysis. B-cell percentages above 3% were removed for computing the B-cell correlations as they were deemed outliers.

Acknowledgements. The authors would like to thank Jill Rubinstein for her feedback. A. F. would like to thank The Jackson Laboratory for the JAX scholar award. J.H.C. acknowledges support from grants R01CA230031 and P30CA034196.

Conflict of interest. The authors declare no conflict of interest.

REFERENCES

[1] Javad Noorbakhsh, Saman Farahmand, Ali Foroughi pour, Sandeep Namburi, Dennis Caruana, David Rimm, Mohammad Soltanieh-ha, Kourosh Zarringhalam, and Jeffrey H. Chuang.
Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images.
nature communications.
1, 2, 3, 5, 7, 8, 16

[2] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley.
Deep learning for healthcare: review, opportunities and challenges.
Briefings in bioinformatics, 19(6):1236–1246, 2018.
1

[3] Shidan Wang, Donghan M Yang, Ruichen Rong, Xiaowei Zhan, Junya Fujimoto, Hongyu Liu, John Minna, Ignacio Ivan Wistuba, Yang Xie, and Guanghua Xiao.
Artificial intelligence in lung cancer pathology image analysis.
Cancers, 11(11):1673, 2019.
1

[4] Samuel Dodge and Lina Karam.
Understanding how image quality affects deep neural networks.
In *2016 eighth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2016.
2

[5] Tejal Nair, Ali Foroughi pour, and Jeffrey H. Chuang.
The effect of blurring on lung cancer subtype classification accuracy of convolutional neural networks.
In *IEEE conference on bioinformatics and biomedicine*, pages 2987–2989. IEEE, 2020.
2

[6] Frederick M Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I Olopade, Jakob N Kather, et al.
The impact of digital histopathology batch effect on deep learning model accuracy and bias.
bioRxiv, 2020.
2

[7] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans.
innvestigate neural networks!
J. Mach. Learn. Res., 20(93):1–8, 2019.
2, 16

[8] Amirata Ghorbani, Abubakar Abid, and James Zou.
Interpretation of neural networks is fragile.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.
2

[9] Ming Y Lu, Melissa Zhao, Maha Shady, Jana Lipkova, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood.
Deep learning-based computational pathology predicts origins for cancers of unknown primary.
arXiv preprint arXiv:2006.13932, 2020.
2

[10] Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller, and Alexander Binder.
Resolving challenges in deep learning-based analyses of histopathological images using explanation methods.
Scientific reports, 10(1):1–12, 2020.
2

[11] Ellery Wulczyn, David F Steiner, Melissa Moran, Markus Plass, Robert Reihs, Fraser Tan, Isabelle Flament-Auvigne, Trissia Brown, Peter Regitnig, Po-Hsuan Cameron Chen, et al.
Interpretable survival prediction for colorectal cancer using deep learning.
NPJ digital medicine, 4(1):1–13, 2021.
2, 7

[12] Kevin Faust, Adil Roohi, Alberto J Leon, Emeline Leroux, Anglin Dent, Andrew J Evans, Trevor J Pugh, Sangeetha N Kalimuthu, Ugljesa Djuric, and Phedias Diamandis.
Unsupervised resolution of histomorphologic heterogeneity in renal cell carcinoma using a brain tumor–educated neural network.

JCO Clinical Cancer Informatics, 4:811–821, 2020.
2, 7, 8

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei.
Imagenet: A large-scale hierarchical image database.
In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
2

[14] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, et al.
A deep learning model to predict rna-seq expression of tumours from whole slide images.
Nature communications, 11(1):1–15, 2020.
2, 7

[15] Alona Levy-Jurgenson, Xavier Tekpli, Vessela N Kristensen, and Zohar Yakhini.
Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer.
Scientific reports, 10(1):1–11, 2020.
2, 7

[16] Liviu Badea and Emil Stănescu.
Identifying transcriptomic correlates of histology using deep learning.
PLOS ONE, 15(11):1–30, 11 2020.
2, 7

[17] Rishi R Rawat, Itzel Ortega, Preeyam Roy, Fei Sha, Darryl Shibata, Daniel Ruderman, and David B Agus.
Deep learned tissue "fingerprints" classify breast cancers by er/pr/her2 status from h&e images.
Scientific reports, 10(1):1–13, 2020.
2

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna.
Rethinking the inception architecture for computer vision.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
3

[19] Ali Foroughi pour and Lori A. Dalton.
Theory of optimal bayesian feature filtering.
Bayesian Analysis, 15(4):1169–1197, 2020.
3, 8, 9

[20] Robert A Ambros and Ronald Trost.
Cellularity in breast carcinoma.
American journal of clinical pathology, 93(1):98–100, 1990.
4

[21] Kanji Tanaka, Daigo Yamamoto, Masanori Yamada, and Homa Okugawa.
Influence of cellularity in human breast carcinoma.
The Breast, 13(4):334–340, 2004.
4

[22] A Serrablo, P Paliogiannis, F Pulighe, S Saudi-Moro Moro, V Borrego-Estella, F Attene, F Scognamillo, and C Hörndler.
Impact of novel histopathological factors on the outcomes of liver surgery for colorectal cancer metastases.
European Journal of Surgical Oncology (EJSO), 42(9):1268–1277, 2016.
4

[23] Frederikke Petrine Fliedner, Trine Bjørnbo Engel, Henrik H El-Ali, Anders Elias Hansen, and Andreas Kjaer.
Diffusion weighted magnetic resonance imaging (dw-mri) as a non-invasive, tissue cellularity marker to monitor cancer treatment response.
BMC cancer, 20(1):1–9, 2020.
4

[24] Jeonghyuk Park, Yul Ri Chung, Seo Taek Kong, Yeong Won Kim, Hyunho Park, Kyungdoc Kim, Dong-Il Kim, and Kyu-Hwan Jung.
Aggregation of cohorts for histopathological diagnosis with deep morphological analysis.
Scientific reports, 11(1):1–11, 2021.
5

[25] Timon PH Buys, Raj Chari, Eric HL Lee, May Zhang, Calum MacAulay, Stephen Lam, Wan L Lam, and Victor Ling.
Genetic changes in the evolution of multidrug resistance for cultured human ovarian cancer cells.

Genes, Chromosomes and Cancer, 46(12):1069–1079, 2007.
5

[26] Wei Zhang, Yi Liu, Na Sun, Dan Wang, Jerome Boyd-Kirkup, Xiaoyang Dou, and Jing-Dong Jackie Han.
Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer.
Cell reports, 4(3):542–553, 2013.
5

[27] Su Li, Hua Li, Ying Xu, and Xiaomei Lv.
Identification of candidate biomarkers for epithelial ovarian cancer metastasis using microarray data.
Oncology letters, 14(4):3967–3974, 2017.
5

[28] Karolina Sterzyńska, Andrzej Klejewski, Karolina Wojtowicz, Monika Świerczewska, Marta Nowacka, Dominika Kaźmierczak, Małgorzata Andrzejewska, Damian Rusek, Maciej Brązert, Jacek Brązert, et al.
Mutual expression of aldh1a1, lox, and collagens in ovarian cancer cell lines as combined cscs-and ecm-related models of drug resistance development.
International journal of molecular sciences, 20(1):54, 2019.
5

[29] Michael S Rooney, Sachet A Shukla, Catherine J Wu, Gad Getz, and Nir Hacohen.
Molecular and genetic properties of tumors associated with local immune cytolytic activity.
Cell, 160(1-2):48–61, 2015.
5

[30] Luigi Cari, Francesca De Rosa, Maria Grazia Petrillo, Graziella Migliorati, Giuseppe Nocentini, and Carlo Riccardi.
Identification of 15 t cell restricted genes evaluates t cell infiltration of human healthy tissues and cancers and shows prognostic and predictive potential.
International journal of molecular sciences, 20(20):5242, 2019.
5

[31] Vésteinn Thorsson, David L Gibbs, Scott D Brown, Denise Wolf, Dante S Bortone, Tai-Hsien Ou Yang, Eduard Porta-Pardo, Galen F Gao, Christopher L Plaisier, James A Eddy, et al.
The immune landscape of cancer.
Immunity, 48(4):812–830, 2018.
5, 6, 10

[32] Zi-Han Geng, Chun-Xiang Ye, Yan Huang, Hong-Peng Jiang, Ying-Jiang Ye, Shan Wang, Yuan Zhou, Zhan-Long Shen, and Xiao-Yan Qiu.
Human colorectal cancer cells frequently express igg and display unique ig repertoire.
World journal of gastrointestinal oncology, 11(3):195, 2019.
6

[33] Jordan T Ash, Gregory Darnell, Daniel Munro, and Barbara E Engelhardt.
Joint analysis of expression levels and histological images identifies genes associated with tissue morphology.
Nature communications, 12(1):1–12, 2021.
7, 8

[34] 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.
7

[35] Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai AJ Sommer, Peter Bankhead, et al.
Pan-cancer image-based detection of clinically actionable genetic alterations.
Nature Cancer, 1(8):789–799, 2020.
7

[36] Subutai Ahmad and Luiz Scheinkman.
How can we be so dense? the robustness of highly sparse representations.
CoRR, vol. abs, page 11, 2019.
7

[37] Adnan Siraj Rakin, Zhezhi He, Li Yang, Yanzhi Wang, Liqiang Wang, and Deliang Fan.
Robust sparse regularization: Simultaneously optimizing neural network robustness and compactness.

arXiv preprint arXiv:1905.13074, 2019.
7

[38] Yan Sun, Wenjun Xiong, and Faming Liang.
Sparse deep learning: A new framework immune to local traps and miscalibration.
Advances in Neural Information Processing Systems, 34, 2021.
7

[39] Luke Ternes, Mark Dane, Marilyne Labrie, Gordon Mills, Joe Gray, Laura Heiser, and Young Hwan Chang.
Me-vae: Multi-encoder variational autoencoder for controlling multiple transformational features in single cell image analysis.
bioRxiv, 2021.
8

[40] David McNeely-White, J Ross Beveridge, and Bruce A Draper.
Inception and resnet features are (almost) equivalent.
Cognitive Systems Research, 59:312–318, 2020.
8

[41] Christopher R Jackson, Aravindhan Sriharan, and Louis J Vaickus.
A machine learning algorithm for simulating immunohistochemistry: development of sox10 virtual ihc and evaluation on primarily melanocytic neoplasms.
Modern Pathology, pages 1–11, 2020.
8

[42] Zhaoyang Xu, Carlos Fernández Moro, Béla Bozóky, and Qianni Zhang.
Gan-based virtual re-staining: a promising solution for whole slide image analysis.
arXiv preprint arXiv:1901.04059, 2019.
8

[43] Yoav Benjamini and Yosef Hochberg.
Controlling the false discovery rate: a practical and powerful approach to multiple testing.
Journal of the Royal statistical society: series B (Methodological), 57(1):289–300, 1995.
8, 9

[44] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stefan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, lhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antonio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors.
Scipy 1.0: Fundamental algorithms for scientific computing in python.
Nature Methods, 2020.
8, 10

[45] Skipper Seabold and Josef Perktold.
statsmodels: Econometric and statistical modeling with python.
In *9th Python in Science Conference*, 2010.
9

[46] A. Foroughi Pour and L. A. Dalton.
Bayesian error analysis for feature selection in biomarker discovery.
IEEE Access, 7:127544–127563, 2019.
9

[47] Ali Foroughi pour and Lori A Dalton.
Optimal bayesian filtering for biomarker discovery: Performance and robustness.
IEEE/ACM transactions on computational biology and bioinformatics, 17(1):250–263, 2018.
9

[48] Ali Foroughi pour and Lori A Dalton.
Biomarker discovery via optimal bayesian feature filtering for structured multiclass data.
In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 331–340, 2018.
9

[49] Olivier Ledoit and Michael Wolf.
A well-conditioned estimator for large-dimensional covariance matrices.
Journal of multivariate analysis, 88(2):365–411, 2004.
9, 10

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
Scikit-learn: Machine learning in Python.
Journal of Machine Learning Research, 12:2825–2830, 2011.
9, 10

[51] Michael Waskom and the seaborn development team.
mwaskom/seaborn, September 2020.
9

[52] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien.
Saga: A fast incremental gradient method with support for non-strongly convex composite objectives.
arXiv preprint arXiv:1407.0202, 2014.
10

[53] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu.
Cellpose: a generalist algorithm for cellular segmentation.
Nature Methods, 18(1):100–106, 2021.
10, 17

[54] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot.
Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images.
Medical Image Analysis, 58:101563, 2019.
10

[55] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot.
Pannuke dataset extension, insights and baselines.
arXiv preprint arXiv:2003.10778, 2020.
10

[56] 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.
10

[57] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu.
Attention mechanisms in computer vision: A survey.
arXiv preprint arXiv:2111.07624, 2021.
16

Figures.

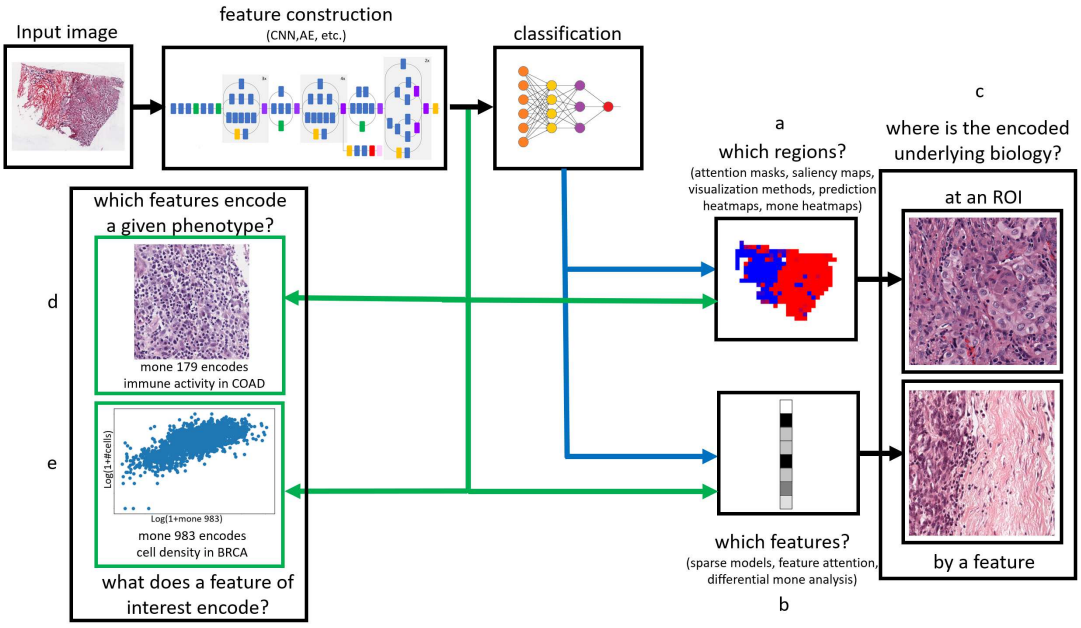

Figure 1: **An overview of interpretation methods in deep learning.** Blue arrows denote methods that require a trained classifier, and green arrows denote methods that do not require a trained classifier. (a) Several methods identify regions which drive the network's prediction. These masks can be generated by the network, e.g. spatial self-attention [57], or as a post-process via visualization methods [7] such as GradCam, or prediction heatmaps [1]. Heatmaps of individual mones and mone-based classifiers can be used to detect predictive regions. (b) Channel attention [57] and sparse models, including sparse mone-based classifiers, identify subsets of features that are predictive of class labels. Differential mone analysis identifies discriminative features without training of a classifier. (c) Methods in (a) and (b) are typically used to select example image regions that have high attention, affect predictions the most, or affect the value of a feature. Mone analysis can be used to (d) identify features that encode a given phenotype of interest and (e) identify the morphology a feature of interest encodes without training a classification model.

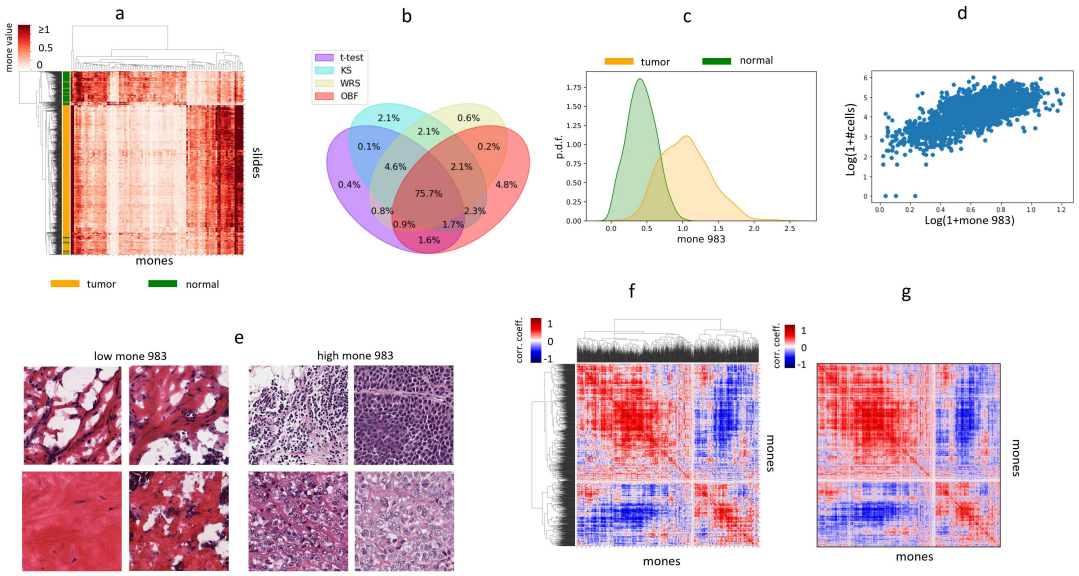

Figure 2: **Individual mones and mone pairs encode and distinguish phenotypes.** (a) Clustermap of BRCA slides using the top100 mones differentiating the slides. 100 mones are sufficient to separate frozen normal (green) from frozen tumor (orange) slides. (b) Venn diagram of statistically significant mones differentiating tumor from adjacent normal frozen slides, comparing different statistical tests. Venn diagrams were calculated for each cancer type, and the observed plot shows the average across all cancer types. On average the statistical tests agree on 75% of mones differentiating between tumor and normal slides. (c) Probability density function of mone 983 among frozen tumor (orange) and adjacent normal (green) BRCA slides. (d) Log-normalized scatter plot of slide level mone 983 and cellpose (see [53]) estimates of cellularity across BRCA frozen slides. (e) Example tiles from slides with extreme mone 983 values (high and low). (f) Cluster map of the mone-mone correlation matrix of LUAD tumor slides, demonstrating that many mones pairs are highly correlated. (g) Mone-mone correlation matrix of LUSC slides, with mones ordered identically to the Figure 1f cluster map.

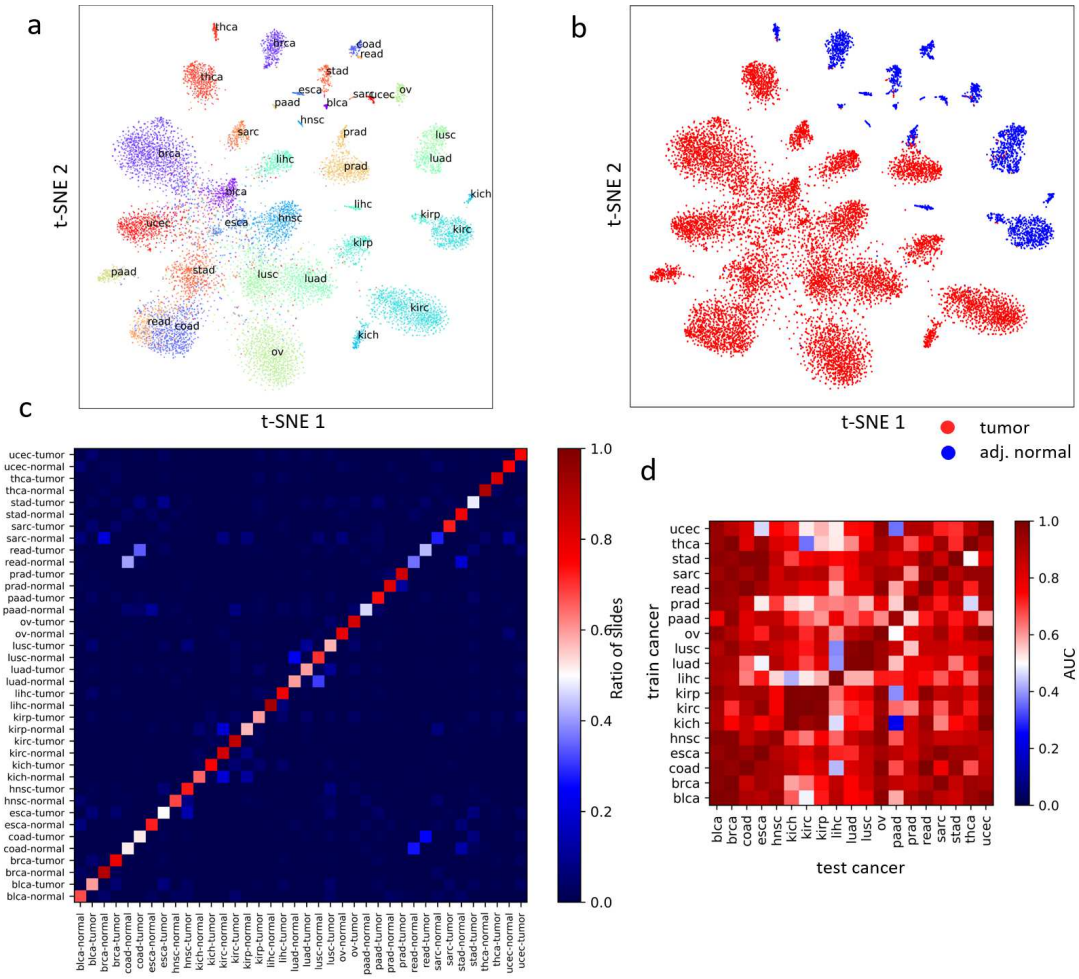

Figure 3: **The joint distribution of mones reliably separates tumor and normal slides and the underlying cancer.** 2D t-SNE plots of the mone-based MLDA feature space distinguishing 38 classes (19 cancers, tumor/normal status) based on (a) cancer type and (b) tumor/normal status. (c) Normalized confusion matrix of the 38-class mone-based logistic regression classifier. The color depicts the ratio of slides with a given true class predicted as any of the possible classes. The large diagonal values suggest the classifier has high accuracy. (d) The cross-classification AUCs of mone-based logistic regression tumor/normal classifier trained on each cancer and applied to all cancers.

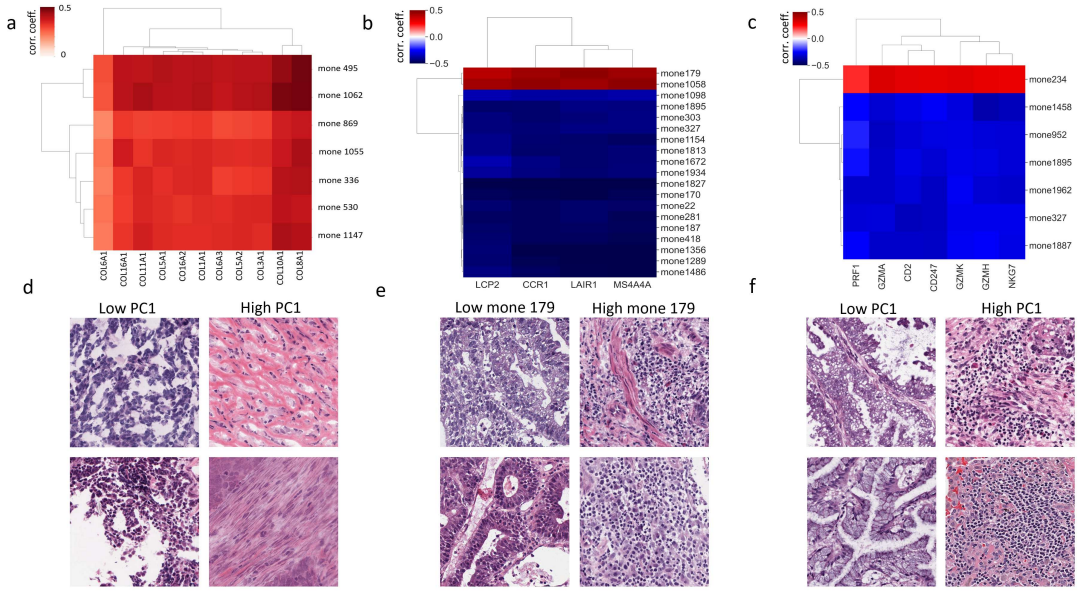

Figure 4: **Mone-gene correlation analysis identifies highly correlated mone-gene clusters.** Correlation matrix of (a) a cluster of highly correlated mones and collagen genes in OV , and a cluster of highly correlated mones and immune-related genes in (b) COAD and (c) LUAD. See Supplementary Figure 9 for adjusted p-values. Example tiles from slides with (d) high and low PC-1 in OV, (e) high and low mone 179 in COAD, and (f) high and low PC-1 in LUAD. Histopathology review identifies that mone-predicted (d) OV tiles with high PC-1 are rich in collagen, and (e) COAD with high mone 179 and (f) luad tiles with high PC-1 have a strong lymphocyte presence. See Supplementary Figures 10-13 for additional examples at both high and low mone values.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.pdf
- SupplementaryFiles.zip