

Measuring the Stability of Learned Features

Kris Sankaran

Department of Statistics
University of Wisconsin - Madison
ksankaran@wisc.edu

February 23, 2021

Abstract

Many modern datasets don't fit neatly into $n \times p$ matrices, but most techniques for measuring statistical stability expect rectangular data. We study methods for stability assessment on non-rectangular data, using statistical learning algorithms to extract rectangular latent features. We design controlled simulations to characterize the power and practicality of competing approaches. This motivates new strategies for visualizing feature stability. Our stability curves supplement the direct analysis, providing information about the reliability of inferences based on learned features. Finally, we illustrate our approach using a spatial proteomics dataset, where machine learning tools can augment the scientist's workflow, but where guarantees of statistical reproducibility are still central. Our raw data, packaged code, and experimental outputs are publicly available.

How should we perform statistical inference on nontabular data? One idea, discussed in [Bühlmann, 2019], is to first transform the data into tabular form, taking advantage of modern feature learning algorithms. The point is important enough to quote at length,

I would like to re-emphasize the importance of new sources of information. Indeed, images, videos and audios are typically cheap devices to record data. [The authors] do not mention recent progress with autoencoders [Hinton and Salakhutdinov, 2006, Vincent et al., 2010]: when using such techniques, one would again end up with numeric features which can then be used for further downstream analysis using techniques from high-dimensional statistics or statistical machine learning [Hastie et al., 2015, Bühlmann and Van De Geer, 2011].

Here, we explore this proposal in depth, illustrating how methods from high-dimensional statistics can be used to study the inferential properties of machine-generated features. Specifically, we present algorithms and visualizations that can be used to characterize the statistical stability of features learned from nontabular

data sources. In the process, we also uncover novel challenges particular to this new setting.

To ground the discussion, consider the following examples,

- Spatial Omics: In addition to measuring gene or protein expression at the cell level, it has become possible to study how expression varies spatially across a tissue [Lundberg and Borner, 2019]. A variety of spatial features are thought to influence biological processes. For example, for some types of cancer, it is thought that the infiltration of tumors by immune cells expressing particular proteins can influence disease prognosis [Keren et al., 2018].
- Ecological and public health: Satellite imagery are now routinely integrated into ecological and public health projects, since they can capture important environmental features and are readily available globally [Rolf et al., 2020, Bondi et al., 2020]. These methods have been found to be effective proxies for otherwise resource-intensive collection methods, like on-the-ground surveys, opening up the possibility of more universal and easily updated monitoring.

In both cases, machine learning methods are key for extracting useful features from novel sources of data. However, unlike many common machine learning applications, the learned features here are subject to critical examination, either to inform biological mechanisms or to ensure vulnerable populations are not put at risk.

This question of how to perform inference on learned features is not a new one — principal components can be bootstrapped Diaconis and Efron [1983], excess error can be estimated from selected features Gong [1986], and confidence regions are available for exploratory projection pursuit Elgueo and Holmes-Junca [1988]. More recently, studies have investigated the use of black-box predictions as features in downstream analysis [Wang et al., 2020b].

However, for both the proteomics and satellite imagery examples presented above, these methods can't be directly applied, because the raw data are not tabular. Instead, features are typically automatically learned using deep learning. What is new in this setting?

- Modern learned features are “distributed” McClelland et al. [1986]. This means that any pattern observed by the algorithm will be encoded by a large number of more elementary features, not any single feature specialized to that pattern. A deep learning algorithm is able to recognize a highway in a satellite image because it can merge evidence from across neurons (i.e., an elementary learned feature) that activate when particular color, shape, and edge features are present. This approach turns out to be much more effective than curating specialized features for many computational tasks, but it poses a challenge for human inspection.
- From one run to the next, the learned features can change. This is unlike in principal components analysis, say, where the learned components are ordered in a natural way. If the deep learning features were more specialized, it might be possible to recognize the same feature across two runs, and then match them.

However, the features are distributed, so it isn't easy to say that any given neuron from the first run matches any other neuron(s) in the second.

- It's impractical to bootstrap methods that take hours to run, even if they could be done in parallel. Moreover, it's unclear what information should be compared across bootstraps — the model parameters, the learned features, or summary statistics about the features.
- Some form of sample-splitting must take place, to ensure that features are not evaluated using the same data that was used to learn them. However, it's unclear how the splitting should be carried out. How much data should be used for feature learning, and how much for inference?

This work discusses these questions, proposing relevant definitions and algorithms, examining their behavior through simulation, and illustrating their use on a spatial proteomics dataset. Our basic takeaways are,

- While learned features are not interpretable when viewed in isolation, their associated feature subspaces often are. A feature learning algorithm may require a large number of elementary features in order to develop effective distributed representations, but the effective dimensionality of these representations is often small. These algorithms learn many, but highly correlated, features.
- Given enough training data, learned feature subspaces are often stable from one run to the next, and this can be quantified using a Procrustes analysis. Unsupervised feature learning algorithms are typically more stable than supervised ones.
- For problems where unsupervised feature learning is effective, a fast approximation to full deep model training, called the random convolutional features (RCF) model, can suffice for a feature stability analysis.
- Inference is less data-greedy than feature learning, in the sense that when few samples are reserved for inference, stable features can still be identified. This is no longer the case when few samples are reserved for feature learning.

Section 1 provides a description of our problem setting. Section 2 summarizes the key technical tools used in this study. We present generic algorithms for measuring feature subspace stability in Section 3, and we study its properties through a simulation in Section 4. We conclude in Section 5 with an application to a spatial proteomics dataset.

1 Problem Setup

Our goal is to characterize the stability of features that were algorithmically derived from n samples $x_i \in \mathcal{X}$. For example, each x_i might be a spatial proteomics measurement or a satellite image. They could also be more general data types — x_i might be the audio recording for one of n speakers, or the graph derived from one of n molecules.

Optionally, a vector $\mathbf{y} \in \mathbb{R}^n$ of responses associated with each observation will be available. We denote the full set of available data by \mathcal{D} .

Definition 1.1. A *feature learner* is a parameterized mapping $T(\cdot; \theta) : \mathcal{X} \rightarrow \mathbb{R}^L$ which takes data from the raw input domain \mathcal{X} and represents it using a vector in \mathbb{R}^L .

For example, in the proteomics and satellite applications, we expect the learner to transform a set of raw pixel intensities into a vector of features reflecting biological or environmental properties of the imaged sample. The parameter θ is estimated from data, typically through an optimization problem,

$$\hat{\theta} := \arg \min_{\theta \in \Theta} \mathcal{L}(\mathcal{D}, T(\cdot; \theta))$$

for some loss function \mathcal{L} . In an unsupervised feature learner, candidates $\theta \in \Theta$ are functions of x_1, \dots, x_n alone. For a supervised feature learner, the class includes functions of both x_1, \dots, x_n and \mathbf{y} . To simplify notation, we will write $z_i = T(x_i; \hat{\theta}) \in \mathbb{R}^L$ to denote the learned features associated with the i^{th} observation.

A first attempt might try to assign a stability score to each of the L coordinates of z . An investigator would then have confidence that, if data were sampled again from the same population, and if features were extracted by the same black box, then the features with high stability scores would reappear in the new analysis. The essential challenge is that the learned features are not the same from one run to the next; the l^{th} learned feature from run 1 need not have any relationship with the l^{th} feature from run 2. Parallel problems are well-known in clustering and factor analysis. Cluster labels can be permuted without changing the quality of the clustering, leading to the label-switching problem. Likewise, without any criteria for sorting or postprocessing factors, latent factors from a factor analysis are unidentifiable.

To address this issue, we distinguish between two notions of stability, which we call feature subspace and selection stability, respectively. The idea of subspace stability is that, even if there is no direct correspondence between learned features across runs, they may all reflect the same underlying latent features. Different runs of the feature learning algorithm return different bases for nearly the same subspace. To make this more precise, suppose that the b^{th} run of the feature learning algorithm produces,

$$\mathbf{Z}_b = \begin{pmatrix} z_1^b \\ \vdots \\ z_n^b \end{pmatrix} \in \mathbb{R}^{n \times L}$$

and define an alignment function \mathcal{A} which takes learned to aligned features,

$$\mathcal{A} : \mathbf{Z}_1, \dots, \mathbf{Z}_B \rightarrow \mathbf{M}, (\mathbf{Z}_1, \dots, \mathbf{Z}_B).$$

We think of $\mathbf{M} \in \mathbb{R}^{n \times K}$ as the average of all B representations and $\mathbf{Z}_b \in \mathbb{R}^{n \times K}$ as the version of the b^{th} learned representation after they have been put into a common coordinate system. \mathcal{A} is just a Procrustes analysis. With this notation, we can now define subspace stability.

Definition 1.2. The *subspace stability* of B learned representations $\mathbf{Z}_1, \dots, \mathbf{Z}_B$ with respect to an alignment function \mathcal{A} is the average distance between the aligned features and \mathbf{M} ,

$$FSS_{\mathcal{A}}(\mathbf{Z}_1, \dots, \mathbf{Z}_B) = \frac{1}{B} \sum_{b=1}^B \|\mathbf{Z}_b - \mathbf{M}\|_2^2$$

$$\text{where } \mathbf{M}, (\mathbf{Z}_1, \dots, \mathbf{Z}_B) = \mathcal{A}(\mathbf{Z}_1, \dots, \mathbf{Z}_B)$$

By selection stability, we mean that a given aligned feature is repeatedly discovered by a model selection procedure. At this point, the features can be thought of as fixed; we are back on more familiar inferential ground. That is, let \mathcal{S} be a selection function, which takes \mathbf{Z}_b and a response \mathbf{y} and returns a subset $S_b \subseteq \{1, \dots, K\}$ of features important for predicting \mathbf{y} .

Definition 1.3. The *selection stability* of the k^{th} aligned feature with respect to the selection \mathcal{S} is the fraction

$$SS_{\mathcal{S}}^k(\mathbf{Z}_1, \dots, \mathbf{Z}_B) = \frac{1}{B} \sum_{b=1}^B \mathbb{I}(k \in \mathcal{S}(\mathbf{Z}_b, \mathbf{y})).$$

2 Background

We review techniques that are used in our algorithmic proposals and computational experiments.

2.1 Stability

Many statistical quantities are based on the idea that meaningful conclusions should be stable to perturbations [Yu \[2013\]](#). We will make use of stability selection, a technique for evaluating significance in high-dimensional linear models [Meinshausen and Bühlmann \[2010\]](#).

For a dataset $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{y} \in \mathbb{R}^n$, stability selection proceeds as follows. First, B subsamples of size $\lfloor \frac{n}{2} \rfloor$ are drawn from the data, and for each a lasso regression is run over a grid of regularization parameters λ . Each of these regressions results in p coefficient trajectories $\hat{\beta}_j(\lambda)$, and important features are expected to become nonzero earlier on the regularization path (that is, even with large regularization λ). For a given λ and feature j , let $\hat{\Pi}_j(\lambda)$ measure the fraction of subsamples for which $|\hat{\beta}_j(\lambda)| > 0$. The paths $(\hat{\Pi}_j(\lambda))_{j=1}^p$ describe the importance of each of the p regression features. For a given stringency threshold π_{thr} , the procedure selects features $\hat{S} = \{j : \max_{\lambda} \hat{\Pi}_j(\lambda) \geq \pi_{thr}\}$.

Let $q_{\Lambda} = \mathbb{E} |\hat{S}_{\Lambda}|$ denote the expected number of selected features. It can be shown that, for $\pi_{thr} \geq \frac{1}{2}$, and assuming a sufficiently well-behaved \hat{S} , the expected number of falsely selected features is bounded by $\frac{1}{2\pi_{thr}-1} \frac{q_{\Lambda}^2}{p}$. The term $\frac{q_{\Lambda}}{p}$ is like the typical fraction of selected features; the coefficient $\frac{q_{\Lambda}}{2\pi_{thr}-1}$ describes the fraction of those that are expected to be false positives.

2.2 Feature Learning

We consider three particular feature learning algorithms. The first is called the Variational Autoencoder (VAE). Like principal components analysis, this algorithm learns an L -dimensional representation of a dataset by optimizing a reconstruction objective. Formally, the algorithm first posits a generative model $p(z)p_\xi(x|z)$ of the data; $p(z)$ is a prior on latent features and $p_\xi(x|z)$ is a likelihood parameterized by ξ . The algorithm finds a pair ξ, φ maximizing the lower bound,

$$\log p_\xi(x) \geq \mathbb{E}_{q_\varphi} [\log p_\xi(x | z)] - D_{KL}(q_\varphi(z | x) \| p(z))$$

where $q_\varphi(z|x) = N(\mu_\varphi(x), \sigma_\varphi^2(x))$ maps raw data examples to distributions in a latent space of more meaningful features. This optimization problem is nonconvex, and is typically solved through a variant of stochastic gradient descent. There are many particular implementations of the VAE; our experiments are based on [Van Den Oord et al., 2017].

Second, we investigate features learned in a supervised way through a Convolutional Neural Network (CNN). For regression, a CNN optimizes the empirical estimate of the risk $\mathbf{E}\|y - f_{W_{1:J}}(x)^T \beta\|_2^2$ over $W_{1:J}$ and β . $f_{W_{1:J}}$ transforms the raw input into the “last layer”’s features, and is defined recursively according to

$$\begin{aligned} f_{W_{1:j}}^j(x) &= \sigma\left(W_j f_{W_{1:(j-1)}}^{j-1}(x)\right) \\ f^0(x) &= x \end{aligned}$$

for $\sigma(x)$ defined as $\sigma(x) := x\mathbb{I}(x \geq 0)$ and matrices W restricted to the set of convolution operators. Like in the VAE, this optimization is nonconvex and is typically found through first-order optimization methods. Our particular implementation is the CBR architecture described in [Raghu et al., 2017].

Third, we consider a random convolutional features (RCF) model [Rahimi and Recht, 2008]. A random sample of L training examples $x_{i_1}, \dots, x_{i_L} \in \mathbb{R}^{w \times h \times c}$ are selected; the x_i ’s are assumed to be c -channel images with dimension $w \times h$. For each sample, a random $s \times s$ patch, which we call $w_p \in \mathbb{R}^{s \times s \times c}$, is extracted. For any c -channel image x , the l^{th} feature z_l is found by convolving x with w_l and spatially averaging over all activations.

To train an RCF, the training data are first featurized into $\mathbf{Z} \in \mathbb{R}^{n \times L}$ using random image patches, as described above. Then, a ridge regression model is trained, solving

$$\hat{\beta} := \arg \min_{\beta} \|y - \mathbf{Z}\beta\|_2^2 + \lambda\|\beta\|_2$$

For a new example x^* , the same image patches w_1, \dots, w_L are used to create a featurization z^* , and predictions are made with $z^{*T}\hat{\beta}$. Unlike either the VAE or CNN, this model does not require gradient based training, and it can serve as a fast, and often effective, baseline.

2.3 Procrustes Analysis

Procrustes Analysis gives a way of aligning multiple tables. Given two centered tables \mathbf{X} and \mathbf{Y} , the Procrustes problem finds the rotation matrix \mathbf{R} solving the optimization,

$$\min_{\mathbf{R} \in \mathcal{O}(p,p)} \|\mathbf{X} - \mathbf{Y}\mathbf{R}\|_F^2,$$

where $\mathcal{O}(p,p)$ denotes the space of $p \times p$ orthonormal matrices. Using the associated Lagrangian,

$$\max_{\mathbf{R}, \mathbf{\Lambda} \succeq 0} \|\mathbf{X} - \mathbf{Y}\mathbf{R}\|_F^2 - \mathbf{\Lambda} (\mathbf{R}^T \mathbf{R} - \mathbf{I})$$

the solution can be shown to be $\hat{\mathbf{R}} = \mathbf{U}^T \mathbf{V}$ for \mathbf{U} and \mathbf{V} obtained by the SVD of $\mathbf{X}^T \mathbf{Y}$. For B matrices $\mathbf{X}_1, \dots, \mathbf{X}_B$, the generalized Procrustes problem finds B rotations $\mathbf{R}_1, \dots, \mathbf{R}_B$ and mean matrix \mathbf{M} solving

$$\min_{\mathbf{R}_1, \dots, \mathbf{R}_B \in \mathcal{O}(p,p), \mathbf{M}} \sum_{b=1}^B \|\mathbf{X}_b \mathbf{R}_b - \mathbf{M}\|_F^2.$$

While there is no closed form solution, the optimization can be solved by cyclically updating each \mathbf{R}_b via standard Procrustes problems and then updating $\mathbf{M} = \frac{1}{B} \sum_{b=1}^B \mathbf{X}_b \mathbf{R}_b$.

2.4 Representation Analysis

Our approach is closely related to Singular Vector Canonical Correlation Analysis (SVCCA), a method used to compare features learned across deep learning models. In SVCCA, the two representations are identified with two matrices, $X \in \mathbb{R}^{n \times l_1}$ and $Y \in \mathbb{R}^{n \times l_2}$. The ij^{th} cell in each matrix corresponds to the activation of neuron j on sample i . A representation is a pattern of activations across a collection of neurons.

To compare representations, SVCCA first computes singular value decompositions $X = U_X D_X V_X^T$ and $Y = U_Y D_Y V_Y^T$. The coordinates of these matrices with respect to the top K singular vector directions are then extracted, $Z_X = U_{X,1:K} D_{X,1:K}$ and $Z_Y = U_{Y,1:K} D_{Y,1:K}$. Finally, a canonical correlation analysis is performed on these coordinate matrices. That is, the top CCA directions u_1, v_1 are found by optimizing

$$\begin{aligned} & \text{maximize } u^T \hat{\Sigma}_{Z_X}^{-\frac{1}{2}} \hat{\Sigma}_{Z_X Z_Y} \hat{\Sigma}_{Z_Y}^{-\frac{1}{2}} v \\ & \text{subject to } \|u\|_2 = \|v\| = 1 \end{aligned}$$

and subsequent directions are found by solving the same problem after orthogonalizing Z_X and Z_Y to previously found directions. The value of the objective for each of the K directions is denoted ρ_k , and the overall similarity between the two representations is taken to be the average of these canonical correlations: $\frac{1}{K} \sum_{k=1}^K \rho_k$.

Note that, while in principle, it would be possible to perform a CCA on the activations X and Y directly, for representations with many neurons, the dimensionality reduction step can simplify computation, because it avoids inverting a high-dimensional covariance matrix.

2.5 Sparse Components Analysis

As in SVCCA, we reduce the dimensionality of learned features before comparing them. We use both PCA and a variant called Sparse Components Analysis (SCA) [Chen and Rohe, 2020]. For a given data matrix X and a number of components K , and sparsity parameter γ , SCA solves the optimization

$$\begin{aligned} & \text{maximize}_{Z, B, Y} \|X - ZBY^T\|_F \\ & \text{subject to } Z \in \mathcal{O}(n, K) \\ & \quad Y \in \mathcal{O}(p, k) \\ & \quad \|Y\|_1 \leq \gamma. \end{aligned}$$

The matrix Y provides the K SCA loadings, and the product ZB provide the coordinates of each sample with respect to that basis. Note that if B is forced to be diagonal, then this optimization reduces to sparse PCA. This optimization does not directly provide an ordering of the loadings. However, the proportion of variance explained by each dimension can still be computed, and this can be used to re-order the dimensions in a similar way to PCA.

3 Algorithms

Algorithms 1 through 3 give our approach to measuring feature stability. We first motivate the general setup, before explaining specific choices used in our experiments.

The three algorithms do the following,

1. Train a feature learning algorithm on B perturbed versions of the dataset \mathcal{D} . This yields B sets of learned features \mathbf{Z}_b .
2. Use an alignment strategy \mathcal{A} to represent the learned features in a shared coordinate system. These aligned features are called $\underline{\mathbf{Z}}_b$.
3. Using the aligned features, evaluate the importance of each feature dimension using a selection mechanism \mathcal{S} .

The only subtlety is that the feature learning and selection steps are performed on different subsets of \mathcal{D} , indexed by I and I^C , respectively. This is needed to maintain validity of inference – if the same samples were used for selection and learning, features would appear more important than they are.

In our experiments, we use the bootstrap to perturb the data reserved for feature learning. That is, if I_b^* resamples I with replacement, then $\mathcal{D}[I_b^*]$ gives a draw from $\mathcal{P}(\mathcal{D})$. This lets us obtain B sets of learned features by optimizing

$$\hat{\theta}_b := \arg \min_{\theta \in \Theta} \mathcal{L}(\mathcal{D}[I_b^*], T(\cdot; \theta))$$

and setting $\mathbf{Z}_b = T(\mathcal{D}, \hat{\theta}_b)$. This step is summarized by Algorithm 1.

Algorithm 1: Feature Learning

Result: Sets of learned features $(\mathbf{Z}_b)_{b=1}^B$ for \mathcal{D} . Indices $I \subset [n]$ used for feature learning.

Inputs: Dataset \mathcal{D} and perturbation process \mathcal{P} . Candidate feature learners $\{T(\cdot; \theta)\}_{\theta \in \Theta}$ and criterion \mathcal{L} .

1. Randomly split samples into disjoint subsets $I, I^C \subset [n]$.

2. Generate B perturbed datasets,

for $b = 1, \dots, B$ **do**

| $\mathcal{D}_b \sim \mathcal{P}(\mathcal{D}[I])$

end

3. Train B feature learners,

for $b = 1, \dots, B$ **do**

| $\hat{\theta}_b = \arg \min_{\theta \in \Theta} \mathcal{L}(\mathcal{D}_b, T(\cdot, \theta))$

| $\mathbf{Z}_b = T(\mathcal{D}, \hat{\theta}_b)$

end

For the alignment strategy \mathcal{A} , we first reduce the dimensionality of each \mathbf{Z}_b to K dimensions, call this $\tilde{\mathbf{Z}}_b$. Then, we solve a generalized Procrustes problem, finding \mathbf{R}_b 's so that the $\mathbf{Z}_b := \tilde{\mathbf{Z}}_b \mathbf{R}_b$ have low Frobenius distance to \mathbf{M} . For the dimensionality reduction step, we apply either PCA or SCA after centering and scaling. Given this \mathcal{A} , we can compute a feature subspace stability score using Algorithm 2.

Algorithm 2: Feature Subspace Stability

Result: Aligned features $(\mathbf{Z}_1, \dots, \mathbf{Z}_B)$. Subspace stability score $FSS_{\mathcal{A}}$.

Inputs: Learned features $(\mathbf{Z}_1, \dots, \mathbf{Z}_B)$. Alignment strategy \mathcal{A} .

1. Align features,

$$\mathbf{M}, (\mathbf{Z}_1, \dots, \mathbf{Z}_B) = \mathcal{A}(\mathbf{Z}_1, \dots, \mathbf{Z}_B)$$

2. Compute feature subspace stability,

$$FSS_{\mathcal{A}}(\mathbf{Z}_1, \dots, \mathbf{Z}_B) = \frac{1}{B} \sum_{b=1}^B \|\mathbf{Z}_b - \mathbf{M}\|_2^2.$$

For the selection mechanism, we use stability selection. This means that our selection mechanism \mathcal{S} is parameterized by lasso regularization strength λ and selection stringency π_{thr} . In our experiments, we display the full selection curves $\hat{\Pi}_k^b(\lambda)$ for each set of aligned features. From these curves, we can identify important features $S_b(\lambda, \pi_{thr})$ for any choice of λ or π_{thr} . However, for clarity, we suppress this depen-

dence on λ and π_{thr} in Algorithm 3.

Algorithm 3: Selection Stability

Result: Selection stability scores for all aligned dimensions $SS_{\mathcal{S}}^1, \dots, SS_{\mathcal{S}}^K$.

Inputs: Reserved indices I^C . Selection mechanism \mathcal{S} . Aligned features

$(\mathbf{Z}_1, \dots, \mathbf{Z}_B)$, where each $\mathbf{Z}_b \in \mathbb{R}^{n \times K}$.

1. Define selection sets,

for $b = 1, \dots, B$ **do**

| $S_b = \mathcal{S}(\mathbf{Z}_b [I^C])$

end

2. Compute selection scores,

for $k = 1, \dots, K$ **do**

| $SS_{\mathcal{S}}^k = \frac{1}{B} \sum_{b=1}^B \mathbb{I}(k \in S_b)$

end

4 Simulation Experiment

We use a simulation experiment to understand properties of the proposed algorithms. Our guiding questions are,

- (G1) How different are the results obtained via supervised vs. unsupervised feature learning algorithms $T(\cdot; \theta)$?
- (G2) When we vary the relative sizes of I and I^C , we expect a trade-off between feature learning and inferential quality. Can we characterize this trade-off?
- (G3) How does the dimensionality reduction approach used in \mathcal{A} affect downstream inferences?
- (G4) Is it ever possible to substitute the RCF feature learner for either of the more time-consuming VAE or CNN models?

To answer these questions, we evaluate our proposal using a full factorial design with three factors,

1. Relative sizes of I and I^C : We use $|I| \in \{0.15n, 0.5n, 0.9n\}$.
2. Dimensionality reduction procedure: We use both PCA and SCA.
3. Algorithm used: We train CNN, VAE and RCF models.

We use $B = 20$ perturbations in each case. Hence, 3 splits \times 3 models \times 20 perturbations = 180 models are trained, from which 180×2 reductions = 360 dimensionality-reduced features are obtained.

4.1 Simulated Data

The key ingredient in this simulation is the construction of a dataset where all “true” features are directly controlled. To motivate the simulation, imagine studying a set of tumor pathology slides, with the hope of discovering features that are predictive of disease prognosis. Each pathology slide gives a view into a cellular ecosystem — there are several types of cells, with different sizes, density, and affinities to one another.

In the simulation, we first generate the latent properties of each slide. Prognosis is defined as a linear function of these properties. Then, images are generated that also reflect the latent properties. The essential challenge is that the investigator only has access to the images and patient prognoses, not the true properties behind each image. A good feature extractor $T(x; \hat{\theta})$ should recover important cell ecosystem properties from the images alone. Example images for varying values of y are given in Figure 1. In total, our simulation generates 10,000 such RGB images, each of dimension $64 \times 64 \times 3$.

We now give details. The locations of cells are governed by an intensity function drawn from a two-dimensional marked Log Cox Matern Process (LCMP) Diggle et al. [2013]. Recall that a Matern process is a Gaussian process with covariance function,

$$C_{\nu, \alpha}(\|x - y\|) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{\|x - y\|}{\alpha} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{\|x - y\|}{\alpha} \right), \quad (1)$$

where α acts like a bandwidth parameter and ν controls the roughness of the simulated process.

Suppose we have R types of cells. Then, our LCMP should have R classes. This can be constructed as follows. First, a nonnegative process $\Lambda(x)$ is simulated along the image grid, $\Lambda(x) \sim \exp(\mathcal{N}(0, \mathbf{C}_{\nu_A, \alpha_A}))$, where $\mathbf{C}_{\nu_A, \alpha_A}$ is the covariance matrix induced by the C_{ν_A, α_A} in equation 1. This is a baseline intensity that determines the location of cells, regardless of cell type. Then, R further processes are simulated, $B_r(x) \sim \exp(\beta_r + \mathcal{N}(0, \mathbf{C}_{\nu_B, \alpha_B}))$. These processes will reflect the relative frequencies of the R classes at any location x ; the intercept β_r makes a class either more or less frequent across all positions x .

Given these intensity functions, we can simulate N cell locations by drawing from an inhomogeneous Poisson process with intensity $\Lambda(x)$. For a cell at location x , we assign it cell type r with probability $\frac{B_r^\tau(x)}{\sum_{r'=1}^R B_{r'}^\tau(x)}$. Here, we have introduced a temperature parameter τ which controls the degree of mixedness between cell types at a given location.

To complete the procedure for simulating images, we add two last source of variation — the number of cells and the cell size. The number of cells per image is drawn uniformly from 50 to 1000. The cells from class R are drawn with a random radius drawn from Gamma($5, \lambda_r$). A summary of all parameters used to generate each image is given in Table 1. Each parameter is drawn uniformly within its range, which has been chosen to provide sufficient variation in image appearance. These parameters are the “true” latent features associated with the simulated images; they give the most concise description of the variation observed across the images.

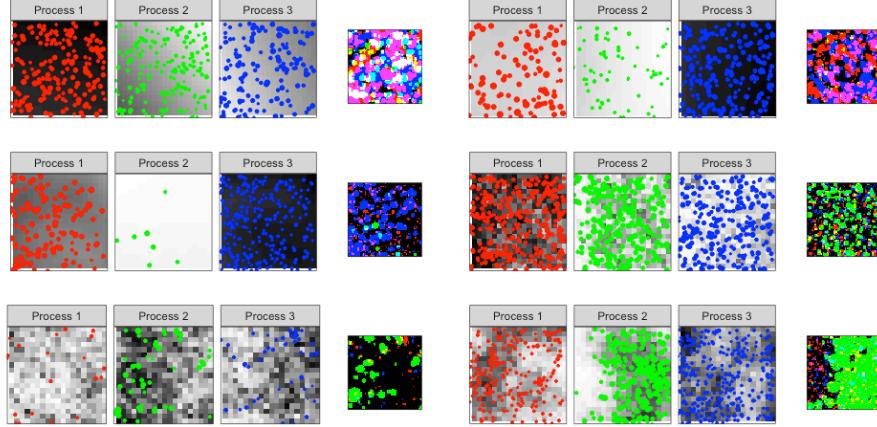


Figure 1: Example simulated images, for low (top), average (middle), and high (bottom) values of y_i . For each sample, three relative intensity functions $B_k(x)$ are generated, shown as a greyscale heatmaps. Samples drawn from each process are overlaid as circles. The final images, displayed at the right of each group, combines points from the three processes, removing the original generating intensity function, which is not available to the feature learner. Small y_i values are associated with smoother, less structured intensity functions.

Feature	Description	Influence	Range
N_i	The total number of cells.	0.5	[50, 1000]
$\nu_{i,\Lambda}$	The roughness of the overall intensity process.	-0.5	[0, 8]
$\alpha_{i,\Lambda}$	The bandwidth of the overall intensity process.	-0.5	[0, 8]
β_{ir}	The intercept controlling the frequency of class r .	1 for $r = 1, -1$ otherwise	[-0.15, 0.15]
ν_{iB}	The roughness of the relative intensity processes.	-0.5	[0, 3]
α_{iB}	The bandwidth of relative intensity processes.	-0.5	[0, 3]
τ_i	The temperature used in cell type assignment.	0.5	[0, 3]
λ_{ir}	The shape parameter controlling the sizes of each cell type.	1 for $r = 1, 0$ otherwise	[100, 500]

Table 1: Our simulation mechanism is governed by the above parameters. Parameters N_i and λ_{ir} control the number and sizes of imaginary cells. $\nu_{i,\Lambda}, \alpha_{i,\Lambda}, \beta_{ik}, \nu_{iB}, \alpha_{iB}$, and τ_i control the overall and relative intensities of the marked LCMP from which the cell positions are drawn. Example draws are displayed in Figure 1.

These features are the latent properties used to generate the prognosis for each patient i . Specifically, we generate $y_i = \sum_k \text{Influence}_k \times \left[\frac{\text{Feature}_{ik} - \bar{\text{Feature}}_k}{SD(\text{Feature}_k)} \right]$ for the values Influence given in Table 1. Note that there is no additional noise: if the Feature_{ik} 's were known for each sample, then the y_i 's could be predicted perfectly. Therefore, the simulation gauges the capacity of the feature learners to recover these known latent features.

4.2 Results

We now summarize findings of our factorial experiment. All simulated data, trained models, and aligned features have been posted publicly; links can be found in the appendix.

4.2.1 Distributed features

To compare the features learned by supervised and unsupervised approaches (G1), we first directly visualize example learned features. Figure 2 shows the activations of learned features across 2000 images for two perturbed versions of the training data when $|I| = 0.9n$. For the three algorithms, the learned features correspond to,

- CNN: Activations from the final hidden layer of neurons, used directly as input for the regression. There are a total of 512 nonnegative features¹.
- VAE: Spatially-pooled activations from the middle, encoding layer of the variational autoencoder. There are a total of 64 real-valued features.
- RCF: The spatially-pooled activations corresponding to each of 1048 random convolutional features.

Our first observation is that, across algorithms, there is no simple correspondence between learned and source features (i.e., parameters of the underlying simulation). For example, it is not the case that one set of features represents the number of cells N , and a different set maps to the roughness ν_A . Rather, there appear to be clusters of learned features, and each cluster corresponds to a pattern of correlations across multiple source features. For example, in Run 1 of the CNN, a cluster of learned features are strongly negatively correlated with y_i , λ_{i1} , and τ and positively correlated with N_i . We also find large subsets of features, across all models, that are only weakly correlated with any source feature.

Next, note that certain source features are “easier” to represent than others, in the sense that more of the learned features are strongly correlated with them. Many features are correlated with N_i , the total number of cells in the image, and λ_{i1} , the size of the cells from Process 1. Depending on the model, the bandwidth α_{ir} , roughness ν_{ir} , and prevalence β_{ik} parameters are either only weakly or not at all correlated with the learned features. Interestingly, the convolutional network learns to represent β_1 well, but not β_2 or β_3 – this is consistent with the fact that only β_1 influences the

¹They are nonnegative because they follow an ReLU layer.

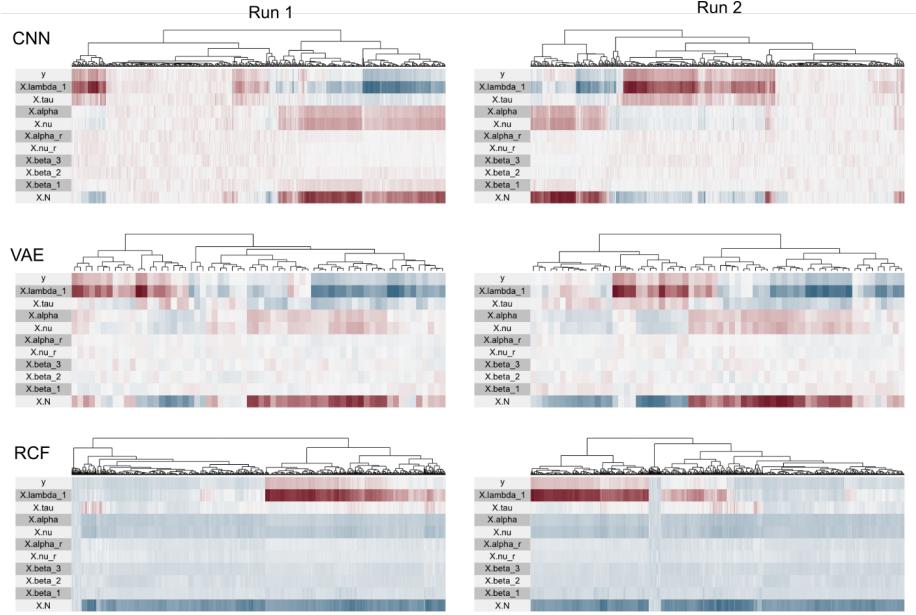


Figure 2: Each feature learning algorithm learns a distributed representation of the true latent features in the simulation. The two columns give results from two runs. Within each heatmap, rows correspond to the parameters in Table 1. Columns are activations of learned features; they have been reordered by a hierarchical clustering using the package [Barter and Yu, 2018]. The color of a cell gives the correlation between the associated true and learned features. Blue and Burgundy encode positive and negative correlation, respectively. All learned features are associated with multiple true features at once.

response y_i . Even when features learn to detect variation in α_{ir} and ν_{ir} , they cannot disambiguate between these two parameters.

Finally, consider differences between feature learning algorithms. The CNN and VAE features tend to be more clustered, with strong correlation across several source features. In contrast, the RCF features show more gradual shifts in correlation strength. They also show relatively little variation in correlation strength across features other than λ_{i1} and N_i .

Note that the features do not naturally map onto one another from one run to the next. This is not obvious from Figure 2, but a zoomed in version in Supplementary Figure 12 showing only the first 15 features per run makes this clear.

4.2.2 Feature Learning vs. Inference

To consider the trade-offs between feature learning and inferential quality (G2) and dimensionality reduction strategy (G3), Figure 3 displays the top canonical correlations between learned features Z_b and the original source features, across bootstrap

replicates and algorithms. Note that we calculate these scores separately for training, development, and testing splits. The training and development splits are subsets of I . Training samples were used in the optimization for each feature learner; development samples were used to choose hyperparameters of the optimizer. They are shown separate from the test samples I^C to make it possible to recognize potential overfitting in the feature learning process.

Even after the initial dimensionality reduction, the CCA canonical correlations decay quickly. The dimensionality reduction method used does not have much effect at this stage. In general, the fraction of data belonging to I also matters little; however, there is an exception in the CNN features. Here, the features learned when $|I|$ is only 15% of the data have noticeably lower canonical correlation in the second dimension. Note also that, from the point of view of these canonical correlations, the RCF and VAE have comparable feature learning behaviors.

The feature learning algorithms do not seem to overfit the simulation data. If they did, then the canonical correlations on the training and development data would be larger than those on the test data. That said, there are noticeable differences between the data splits, and this effect will be visible in later figures as well. We hypothesize that the differences are due to the alignment process. During the Procrustes rotation, one matrix $\mathbf{Z}_b[I^C]$ may get “lucky” and learn an alignment with axes that better reflect variation in the source features. In this way, even though the test data may not have been used to learn features, they may have higher correlation with the source features than the original $\mathbf{Z}_b[I]$.

To shed further light on G2 and G3, Figure 4 shows the median number of features selected by stability selection $|S(\lambda, \pi_{thr})|$ across training sample sizes. The median is taken across all bootstrap iterations. We fix a regularization strength λ where small perturbations in π_{thr} lead to large changes in the number of selected features. We have run stability selection for 250 replicates, all restricted to either training, development, or test data, as indicated by the color of each line.

All the curves decrease because increasing the stringency π_{thr} leads to fewer selected features. We find that the features learned by the CNN are more frequently selected. This is expected, considering that the CNN features are learned in a supervised fashion. More surprising is the fact that the RCF features are more frequently selected than the VAE features, suggesting that the simple baseline might already provide features useful for interpretation, giving an affirmative answer to G4.

Finally, it appears that $|S(\lambda, \pi_{thr})|$ is largest when using 50% of the data for feature learning. Though the features learned using 90% of the data may be higher quality, it is not possible to select them as frequently, because stability selection will have low power when it can only subsample from the 10% of the data reserved for inference. This phenomenon appears for features learned by both supervised and unsupervised methods. For this reason, in the remainder of this study, we will focus on results obtained using a 50% split between learning and inference, i.e., $|I| = |I^C|$, though results using different splits are available in the appendix.

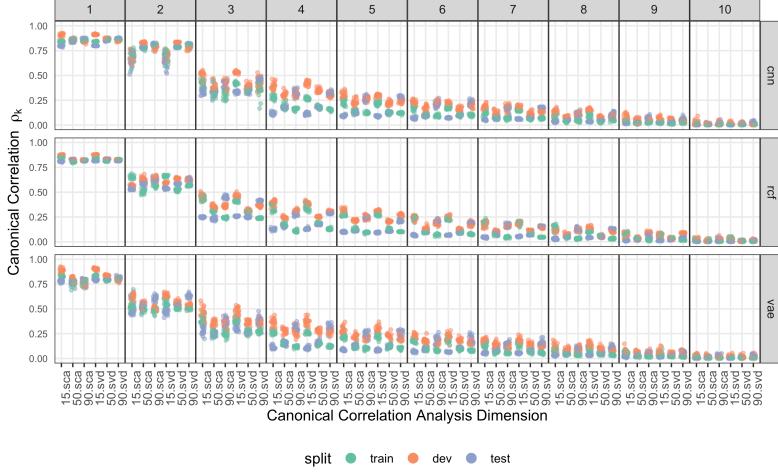


Figure 3: The top canonical correlations between the sources features and the aligned Z_b . Horizontal rows correspond to CNN, RCF, and VAE algorithms. Columns give the CCA dimension. Within each display, the canonical correlations across all bootstraps is shown. The x -axis within each cell gives the amount of data used for feature learning and the dimensionality reduction method used. Canonical correlations are computed separately for samples within training, development, and test splits, indicated by each point’s color.

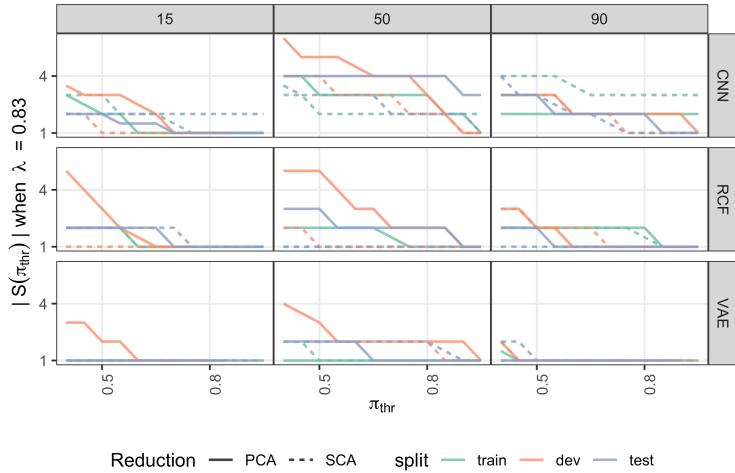


Figure 4: The median number of features selected by stability selection across bootstrap runs Z_b , viewed as a function of π_{thr} . Rows and columns correspond to models and fraction of the data used for training, respectively. Stability selection is run within training, development, and test splits (color) and using two dimensionality reduction procedures during alignment (solid vs. dashed lines).

4.2.3 Stability visualization

Figures 5 and 6 summarize the results of Algorithms 2 and 3 applied to several feature learners. Within a single panel, each star glyph corresponds to a single sample. The arms of the glyph link the bootstrap replicates for that given sample: z_i^1, \dots, z_i^{20} . Each glyph is shaded by the true value of y_i for that sample. Only the first two aligned feature dimensions are shown.

Both the model and reduction strategy used influence the stability of the learned features. Based on the relative sizes of the glyphs, the CNN features are least stable, those from the RCF are most stable, and the VAE features are intermediate. Separate regions of the learned feature space appear more stable than others. Features learned with 15% of the data are highly unstable and show little association with the response. For larger $|I|$'s, the learned features are more stable. Supplementary Figure 14 shows that in this case, there are also stronger associations with \mathbf{y} .

Figure 6 displays stability selection paths $\Pi_k^b(\lambda)$ for combinations of feature learning and dimensionality reduction procedures. Like in Figure 3, we find that the largest variation is between the data splits, but that overfitting does not appear to be a problem. Indeed, for several features, the development and test splits have higher selection probabilities.

Revisiting the effect of dimensionality reduction approach on selection (G3), we find the first major differences between dimensionality reduction strategies. The selection paths are not monotone for SCA, and they also vary substantially across runs. This is most likely a result of the fact that the matrix B in the decomposition $X = ZBY^T$ need not be diagonal, which correlates the resulting coordinates. Further, when using the PCA, the top features are also the most selected ones; this is not the always the case with SCA.

Next, compare the unsupervised, supervised, and accelerated methods (G1, G4), from the selection perspective. Between algorithms, we find that the selection paths for RCF features form a tighter band across all bootstrap samples, another mark in its favor. For all algorithms, this band seems to widen for later feature dimensions; the selection probabilities also generally rise more gradually. For Feature 3 onwards, the selection paths for the VAE rise more gradually than the corresponding paths for either the CNN or RCF dimensions, suggesting a lack of association with \mathbf{y} .

5 Data Analysis

In this section, we conduct a feature stability analysis on the spatial proteomics dataset² reported in the study [Keren et al., 2018], where the authors found a relationship between the spatial organization of Triple Negative Breast Cancer (TNBC) tissues and disease progression. In a classical proteomics study, the expression levels for a set of proteins is measured for a collection of cells, but the cell locations are unknown. In contrast, these data provide for each patient (1) an image delineating cell boundaries and (2) the protein expression levels associated with each cell in the images.

²The data are publicly available. See also the appendix for preprocessed versions.

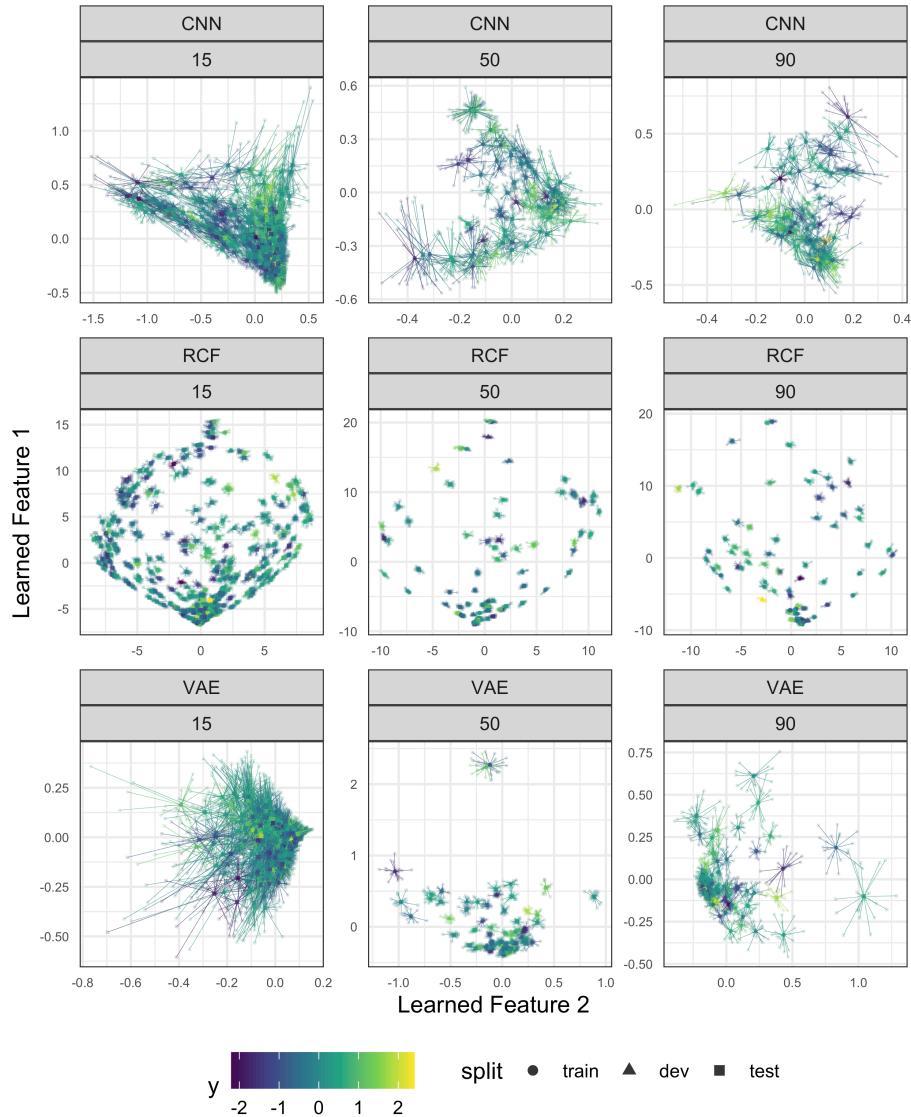


Figure 5: Example aligned features Z_b derived from several models with varying feature learning sample sizes $|I|$. For clarity, only a subset of samples is shown. For a version of this figure showing all samples, and collapsing each star to a point, refer to Supplementary Figure 14.

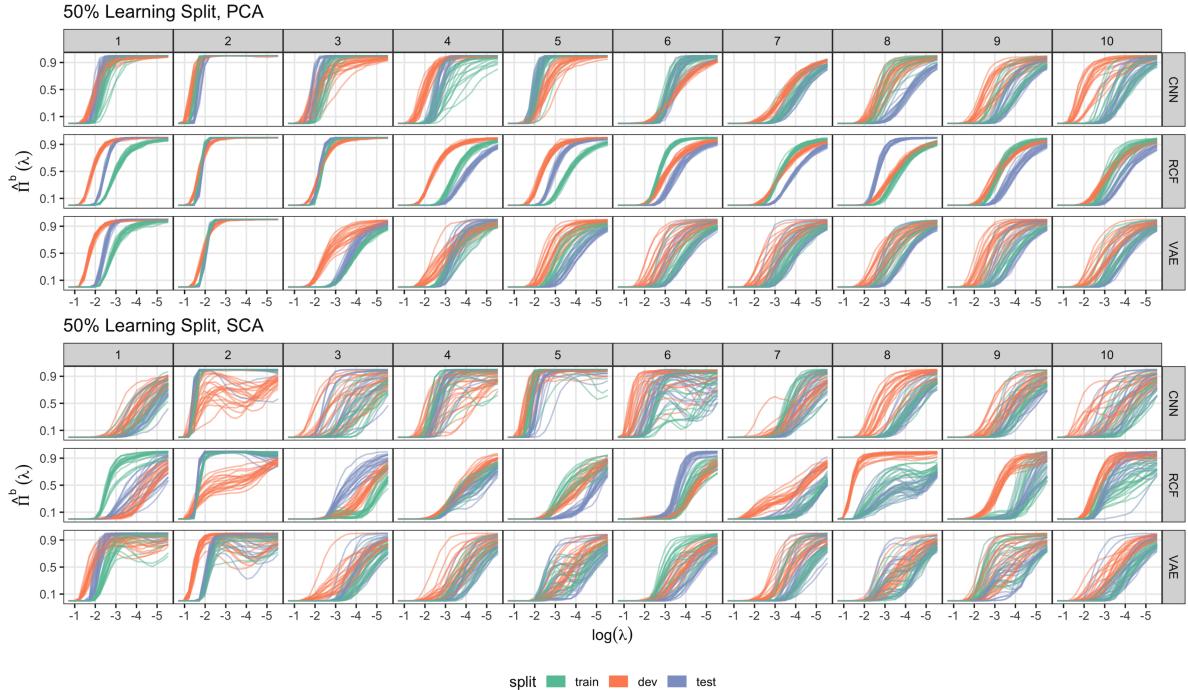


Figure 6: Stability selection paths $\hat{\Pi}_k^b(\lambda)$ for 10 learned features, across models and alignment strategies. Top and bottom groups use PCA and SCA-reduced features, respectively, and each row corresponds to a feature learning algorithm. Each column corresponds to one learned feature dimension. Each curve gives $\hat{\Pi}_k^b(\lambda)$ for one bootstrap replicate b restricted to samples from either the training, development, or test set. The x -axis is the regularization strength $\log \lambda$ and the y -axis is the probability of selecting that feature at the given regularization strength.

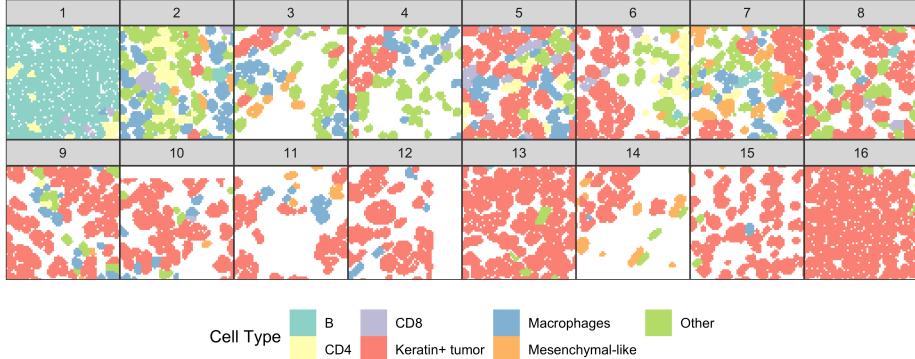


Figure 7: Example patches from the TNBC data. Panels are ordered by y_i , the (log) fraction of cells they contain that belong to the tumor. The goal of the prediction algorithm in this example is to correctly place patches from new patients along this gradient, after having observed many patches and their corresponding y_i 's.

We will work only with the spatial cell delineations, not the protein expression levels. This allows us to study the mechanics of feature learning within the images without having to worry about linking the expression and image data, which is in itself a complex integration problem. Though this means we lose some scientific depth, we gain substantial implementation simplicity, and the analysis serves as a clear illustration. Our complete data are 41 2048 × 2048-dimensional images, each taken from a separate patient. We associate each pixel with one of 7 categories of tumor and immune cell types.

To setup a prediction problem, we first split each image into $512 \times 512 \times 7$ patches. These patches are our x_i . Patches from 32 of the patients are reserved form feature learning. Four among these 32 are used as a development split, to tune parameters of the feature learning algorithms. As a response variable, we use $y_i = \log\left(\frac{\#\{\text{Tumor cells in } x_i\}}{\#\{\text{Immune cells in } x_i\}}\right)$. Example cell patches are shown in Figure 7.

As a baseline, we compare against a ridge regression with pixelwise composition features. Specifically, we train a model with y as a response and the average number of pixels belonging to each of the cell-type categories as a (length 7) feature vector. This helps us to determine whether the model has learned more interesting features useful for counting cells, like cell size and boundaries, rather than simply averaging across pixel values. Indeed, Figure 8 makes clear that, with the exception of the RCF-SCA combination, all feature learning - dimensionality reduction combinations outperform this manual baseline.

Stability curves associated with the learned features from the CNN, RCF, and VAE models are shown in Figure 9. Interestingly, across all algorithms, the top aligned features are not necessarily those with the highest selection probabilities. For example, CNN Feature 4 has higher selections than Feature 3, RCF Feature 6 has higher selection than Feature 4, and VAE Feature 5 is more frequently selected than Feature 3. In contrast to the simulation, the CNN and VAE do not have substantial differences in

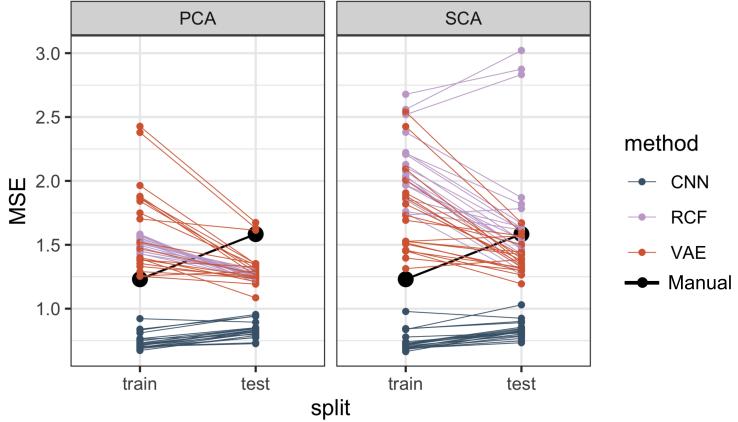


Figure 8: Relative performance of feature learning strategies on the TNBC data. Linked points come from the same bootstrap replicate. The manual features refers to the regression model that simply uses the raw pixel counts for each of the cell types. Predictions from the development split are omitted; this split has few patches, and the estimated MSE’s have high variance.

their selection curves. Further, the RCF features seem to have high selection probabilities across more dimensions than either the CNN or VAE. This suggests that the most salient features in x_i are also relevant for predicting y_i , and that supervision is not as critical in this problem as it was in the simulation.

Example aligned coordinates are given in Figure 10. Consistent with the conclusion, we find that the association with the response is clearly visible with respect to the learned feature dimensions, even when using unsupervised algorithms. The changes in the sizes of glyphs across regions of the learned feature space are especially pronounced in this application. For example, in the VAE, the representations of samples with higher tumor-to-immune ratio y_i are much more stable than those with low ratio.

There is no consensus on how to best interpret automatically learned features [Doshi-Velez and Kim, 2017]. Nonetheless, we present one simple approach in Figure 11, overlaying example patches onto aligned coordinates. For example, in the CNN, the first dimension distinguishes between the relative number of tumor and immune cells, while the second dimension reflects the density of cells. In the RCF, the second dimension captures the diversity of cell types, with more uniform samples on the left and more heterogeneous ones on the right. The second dimension of the VAE seems related to both cell density and number of cell types. The analogous display from an SCA reduction is given in Supplementary Figure 24.

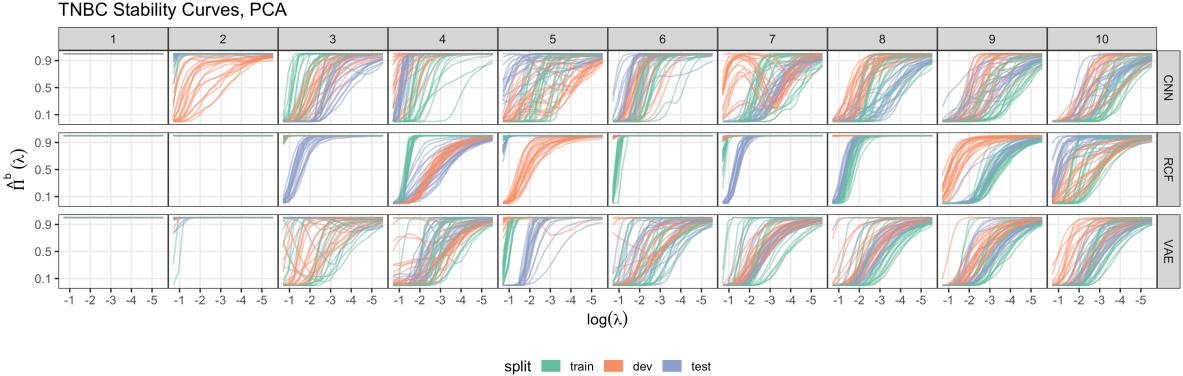


Figure 9: Selection curves for features learned from the TNBC data. 50% of the data were used for feature learning and PCA was used for dimensionality reduction. Each curve is read as in Figure 6. The analogous paths using SCA alignment are given in Supplementary Figure 25.

6 Discussion

This study has investigated the stability of machine-generated, rather than hand-crafted, features. A better understanding of stability in this modern regime has consequences for how these methods can be used in real-world applications, especially those intended for scientific workflows.

Our results raise several questions for further study. It is natural to ask to what extent similar behaviors are exhibited across other data domains, model types, or training regimes. For example, it would not be unusual to represent the cell data in our case study using a marked graph linking neighboring cells. Do the features learned by a graph autoencoder have similar stability properties? In other domains, we may ask whether our methods can be adapted to text or audio data.

Further, there are questions that may guide us towards better instantiations of Algorithms 1 through 3. While we have relied on the bootstrap, our notation encompasses more general perturbations \mathcal{P} . For example, how might learned features change when discarding nonrandomly chosen subsets of training data? Perhaps learning features based on different temporal or spatial subsets could reveal a drift in the important features over time or space. Alternatively, we could imagine perturbing the model training procedure, using different hyperparameters. It would be of interest to trace out the dependence of the learned features on the mechanics of the feature learner.

Similarly, using a Procrustes rotation for \mathcal{A} and stability selection for \mathcal{S} provides a reasonable point of departure, but more sophisticated approaches are possible. For example, dimensionality reduction could be optimized to support alignment; this could be accomplished using multiple canonical correlation analysis. Combinations of features could also be approximately matched across feature learners, using a form of optimal transport, identifying sets of features that all activate on similar input sam-

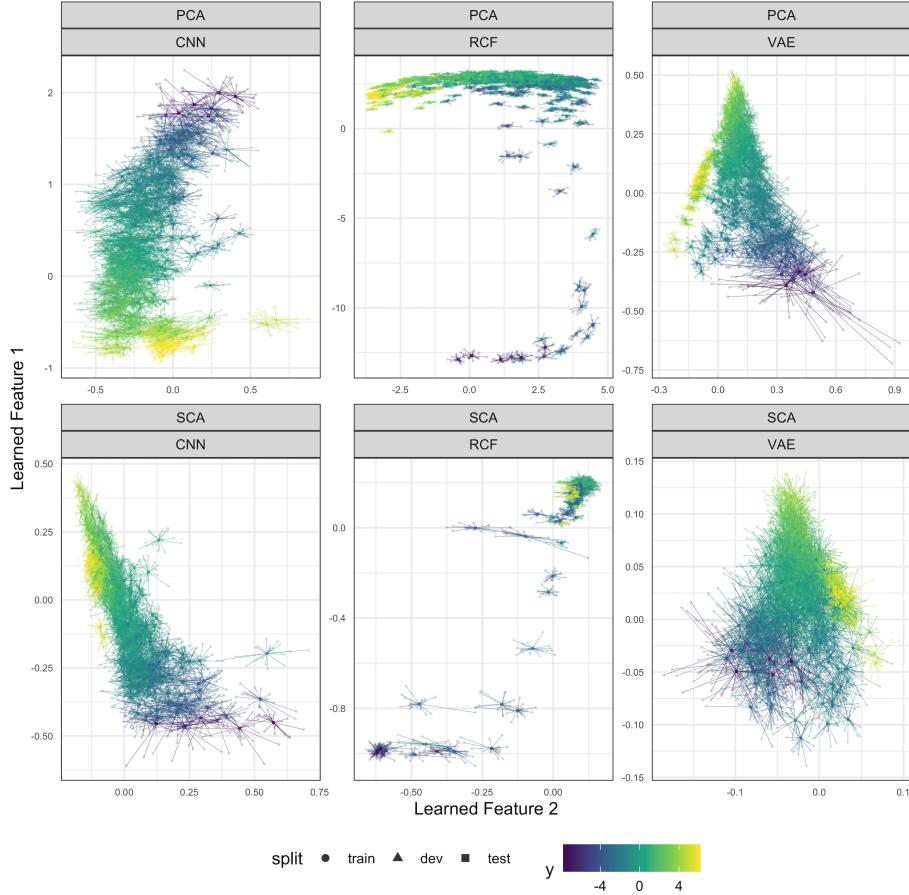


Figure 10: Representation of samples according to some of the important learned feature dimensions. Glyphs are read as in Figure 5. To interpret these features, example patches are arranged according to these coordinates in Figure 11 and Supplementary Figure 24. Only a subsample of points are shown; the full dataset with stars collapsed to points is shown in Figure 21.

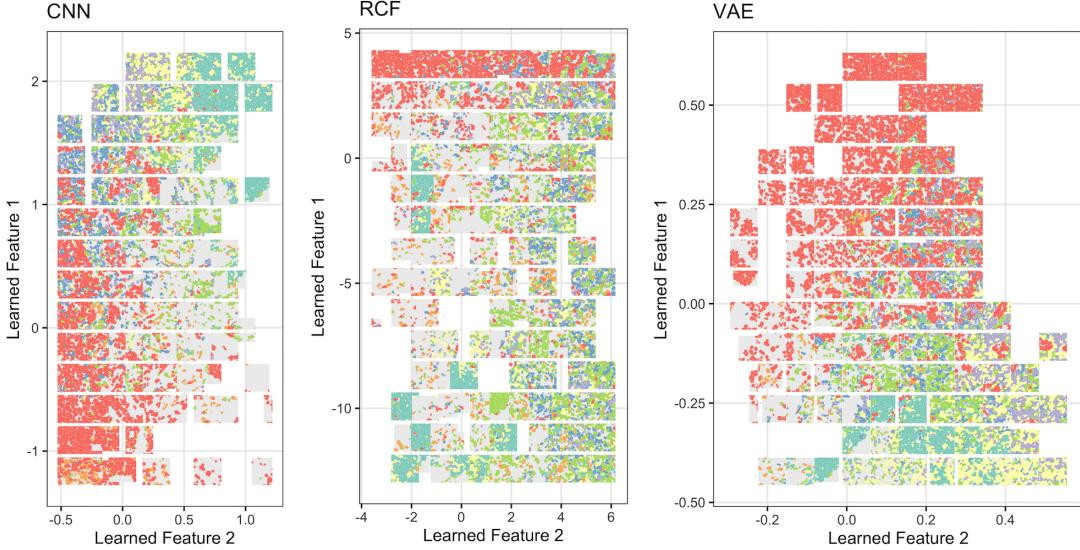


Figure 11: A version of the PCA panels of Figure 10 that overlays representative samples across the learned feature space. Cells are color coded as in Figure 7. Note that the overall shape of the region in which images are displayed mirrors the shape of the cloud of points in Figure 10.

ples. This has been applied in a federated learning context, but not in the study of stability [Wang et al., 2020a].

It is also possible to propose alignments based on more refined measures of correlation [Josse and Holmes, 2016, Azadkia and Chatterjee, 2019]. We have found that using an even split between feature learning and inference gives reasonable results, but our results are admittedly coarse. Though we have concentrated on stability selection, the procedure \mathcal{S} could be any selective inference procedure. Finally, our approaches to summarizing and displaying the resulting representations may be improved through a more careful application of interactive visualization principles.

More broadly, this study is situated in the body of work seeking to bridge the rift between the two cultures [Breiman, 2001, Efron, 2020]. Across the selective inference, conformal prediction, and interpretability literatures, there is a growing understanding that the statistics and machine learning communities could benefit by learning to speak one another’s languages [Angelopoulos et al., 2020, Ren and Candès, 2020]. Nonetheless, while the data sources and feature learning context we consider is novel, our basic motivation is an old statistical idea that still rings true – quoting from [Mosteller and Tukey, 1977],

One hallmark of the statistically conscious investigator is a firm belief that, however the survey, experiment, or observational program actually turned out, it could have turned out some somewhat differently.

In order to accomplish this study, we have adapted tools from the representational analysis, dimensionality reduction, and high-dimensional inference communities. These tools give a window into the workings of modern feature extraction techniques, helping us view commonplace algorithms in a new way. In a sense, our intent is to do more than summarize data – it is to generate new data. In the same way that a standard error around a mean is a new piece of information that supports more nuanced reasoning, we hope to see the development of techniques that generate data characterizing the behavior of modern feature learning algorithms.

References

- Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.
- Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *arXiv preprint arXiv:1910.12327*, 2019.
- Rebecca L Barter and Bin Yu. Superheat: An r package for creating beautiful and extendable heatmaps for visualizing complex data. *Journal of Computational and Graphical Statistics*, 27(4):910–922, 2018.
- Elizabeth Bondi, Andrew Perrault, Fei Fang, Benjamin L Rice, Christopher D Golden, and Milind Tambe. Mapping for public health: Initial plan for using satellite imagery for micronutrient deficiency prediction. 2020.
- Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.
- Peter Bühlmann. Comments on: Data science, big data and statistics. *TEST*, 28(2):330–333, 2019.
- Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
- Fan Chen and Karl Rohe. A new basis for sparse PCA. *arXiv preprint arXiv:2007.00596*, 2020.
- Persi Diaconis and Bradley Efron. Computer-intensive methods in statistics. *Scientific American*, 248(5):116–131, 1983.
- Peter J Diggle, Paula Moraga, Barry Rowlingson, and Benjamin M Taylor. Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.
- Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

- Bradley Efron. Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115(530):636–655, 2020.
- Eric Elguero and Susan Holmes-Junca. Confidence regions for projection pursuit density estimates. In *Compstat*, pages 59–63. Springer, 1988.
- Gail Gong. Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *Journal of the American Statistical Association*, 81(393):108–113, 1986.
- Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.
- Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
- Julie Josse and Susan Holmes. Measuring multivariate association and beyond. *Statistics surveys*, 10:132, 2016.
- Leeat Keren, Marc Bosse, Diana Marquez, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo-Ryun Yang, Allison Kurian, David Van Valen, Robert West, et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell*, 174(6):1373–1387, 2018.
- Emma Lundberg and Georg HH Borner. Spatial proteomics: a powerful discovery tool for cell biology. *Nature Reviews Molecular Cell Biology*, 20(5):285–302, 2019.
- James L McClelland, David E Rumelhart, PDP Research Group, et al. Parallel distributed processing. *Explorations in the Microstructure of Cognition*, 2:216–271, 1986.
- Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.
- Frederick Mosteller and John Wilder Tukey. *Data analysis and regression: a second course in statistics*. 1977.
- Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in neural information processing systems*, pages 6076–6085, 2017.
- Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21:1313–1320, 2008.
- Zhimei Ren and Emmanuel Candès. Knockoffs with side information. *arXiv preprint arXiv:2001.07835*, 2020.
- Esther Rolf, Jonathan Proctor, Tamara Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *NBER Working Paper*, 2020.

Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020a.

Siruo Wang, Tyler H McCormick, and Jeffrey T Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020b.

Bin Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.

7 Reproducibility

Instructions to reproduce our simulations and data analysis example are available in a **README** on the study’s github page. Training split creation, the regression baseline, and feature learning can be reproduced in the ipython notebooks

- `tnbc_splits.ipynb`: Define training and test splits for the TNBC dataset.
- `tnbc_baseline.ipynb`: Train a ridge regression baseliens on the TNBC dataset.
- `model_training.ipynb`: Train either a CNN, RCF, or VAE feature learner.

MIBI-ToF data preparation, generation of simulation data, feature stability analysis, and visualization of results are done within the rmarkdown documents,

- `generate.Rmd`: Simulate the LCGP cells dataset.
- `stability.Rmd`: Perform feature alignment and stability selection.
- `visualize_features.Rmd`: Visualize the aligned and selected features output by `stability.Rmd`.
- `summary_plots.Rmd`: Compute the CCA and selection summaries reported in Figures 3 and 4.

To support code reusability between experiments, two helper packages were prepared,

- `stability`

- `inference`

This packages can be installed by calling,

```
git clone https://github.com/krisrs1128/learnedinference.git
Rscript -e 'devtools::install('learnedinference/inference')'
pip3 install learnedinference/stability
```

from the terminal.

We have prepared a `docker image` with all necessary software pre-installed. For example, to rerun our stability analysis, the following commands may be used,

```
docker run -it krisrs1128/li:latest bash
git clone https://github.com/krisrs1128/learnedinference.git
source learnedinference/.env
# download relevant data
Rscript rmarkdown -e 'rmarkdown::render('learnedinference/inference/vignet
```

Finally, we have released raw data and intermediate results from our analysis,

- `sim.data.tar.gz`: Our toy simulation dataset.
- `tnbc_data.tar.gz`: The preprocessed MIBI-ToF data, with all patient's data split into 64×64 patches and with the associated splits and response value stored in a metadata file.
- `simulation_outputs.tar.gz`: All the models trained in our simulation experiments.
- `tnbc_outputs.tar.gz`: All models trained in our illustration on the TNBC dataset.
- `simulation_figure_data.tar.gz`: The data written by 'stability.Rmd' which was used to generate figures for our simulation.
- `tnbc_figure_data.tar.gz`: The data written by 'stability.Rmd' which was used to generate figures for our data illustration.
- `tnbc_raw.tar.gz`: The original MIBI-ToF tiffs, before splitting into patches.

8 Supplementary Figures

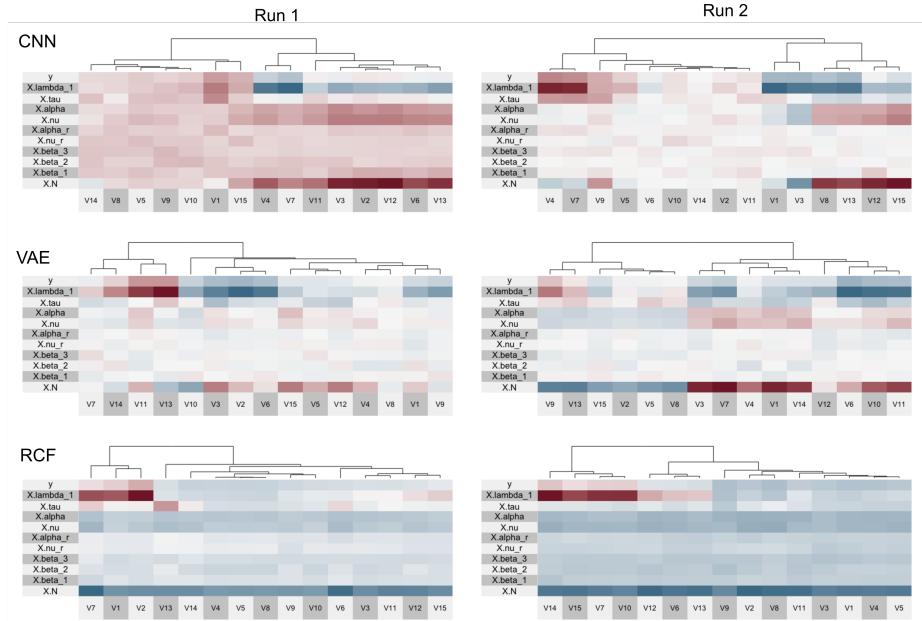


Figure 12: A zoomed version of Figure 2, showing that the features cannot be directly mapped onto one another, from one run to the next.

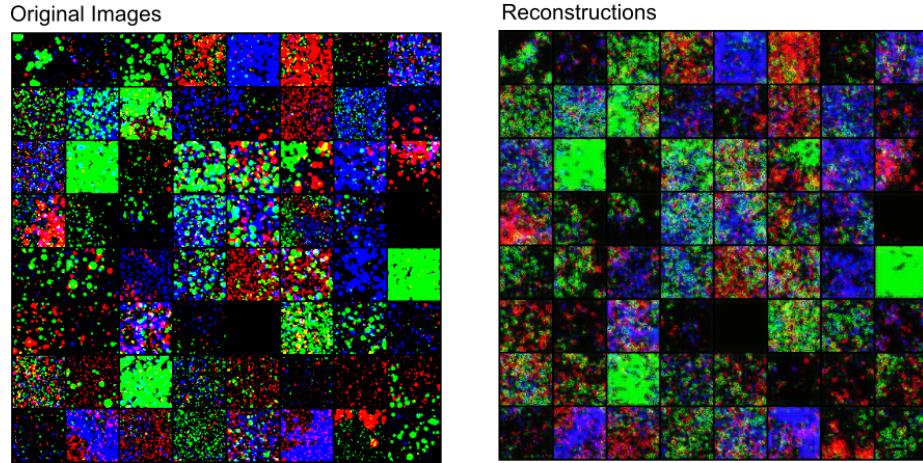


Figure 13: Example reconstructions from the VAE model applied to the simulation data. Original image patches are shown on the left, corresponding reconstructions are given on the right. Though fine-grained details are missed by our version of the VAE model, most of the key global features of each patch seem accurately reflected.

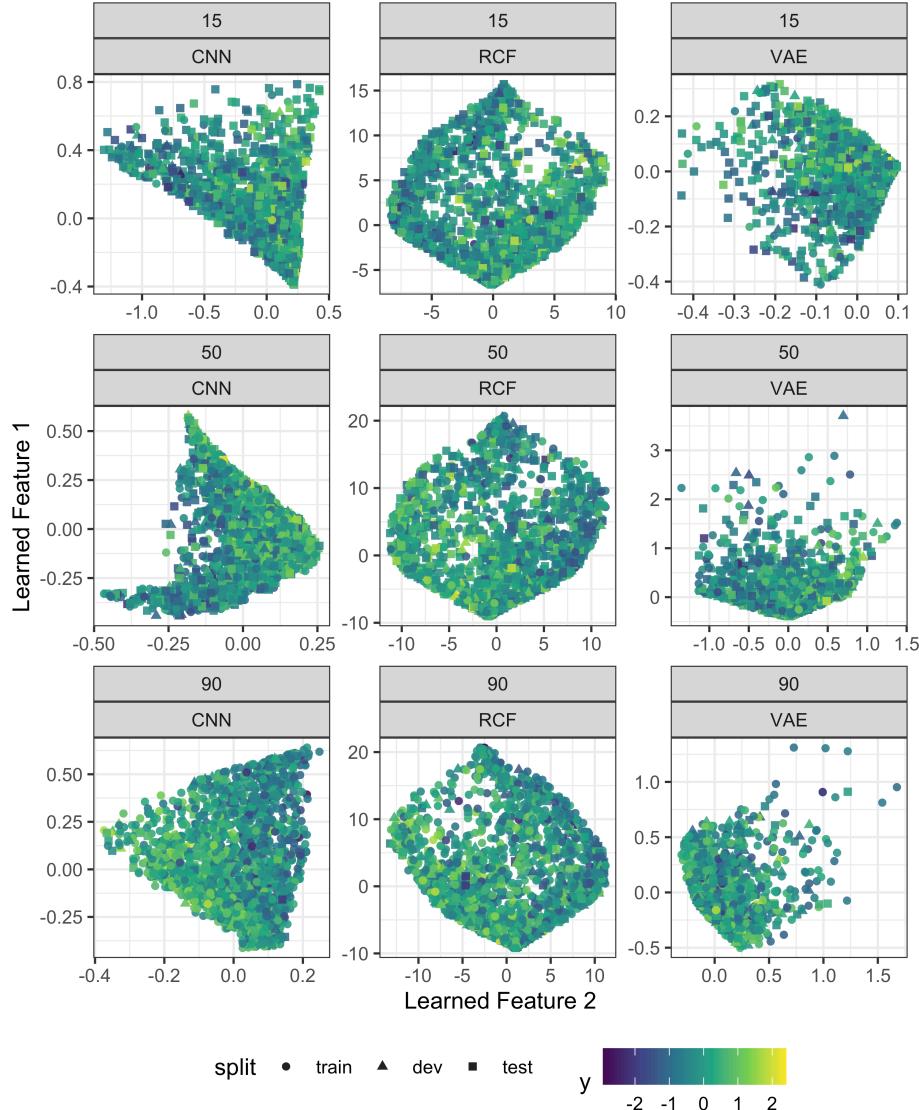


Figure 14: The analogous figure to Figure 5, but keeping only the center of each bootstrap alignment, and showing the embeddings for all images. This view makes the association with the response y clearer, especially for larger $|I|$.

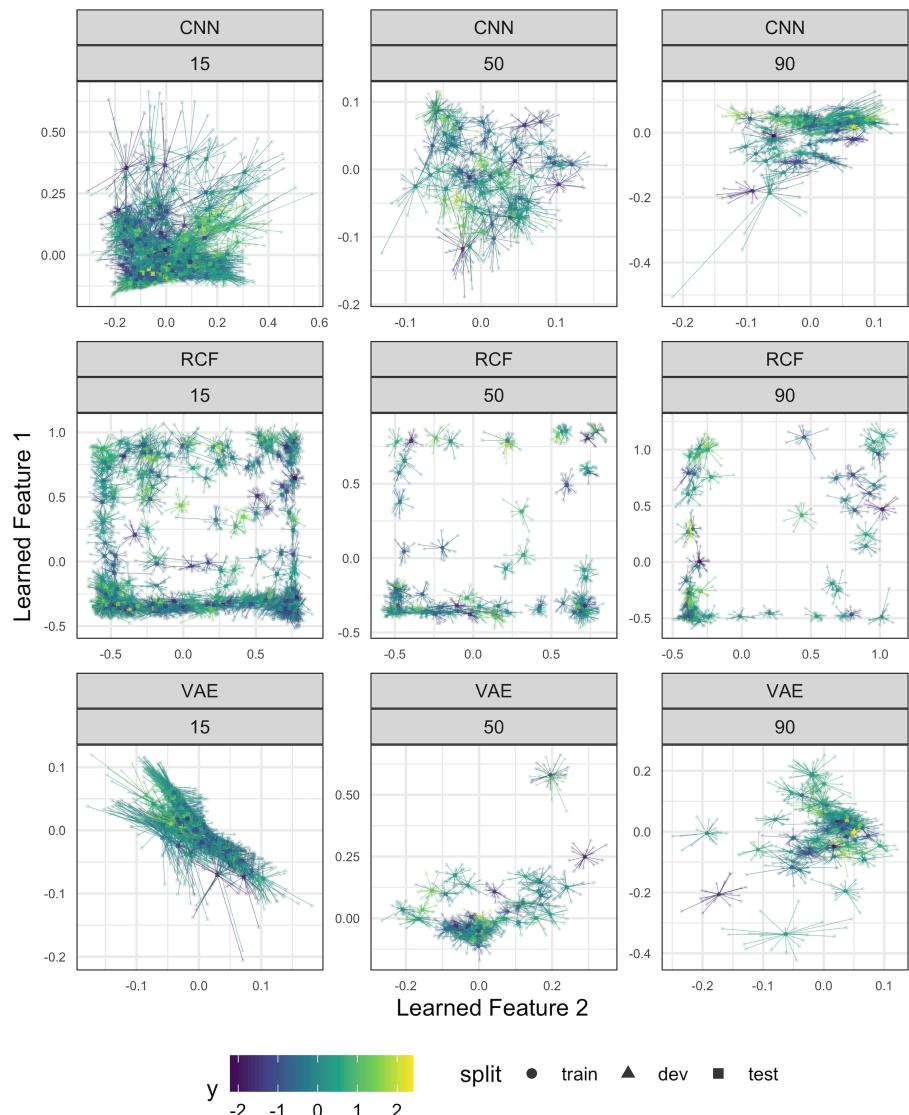


Figure 15: The analogous figure to Figure 5, but using SCA dimensionality reduction.

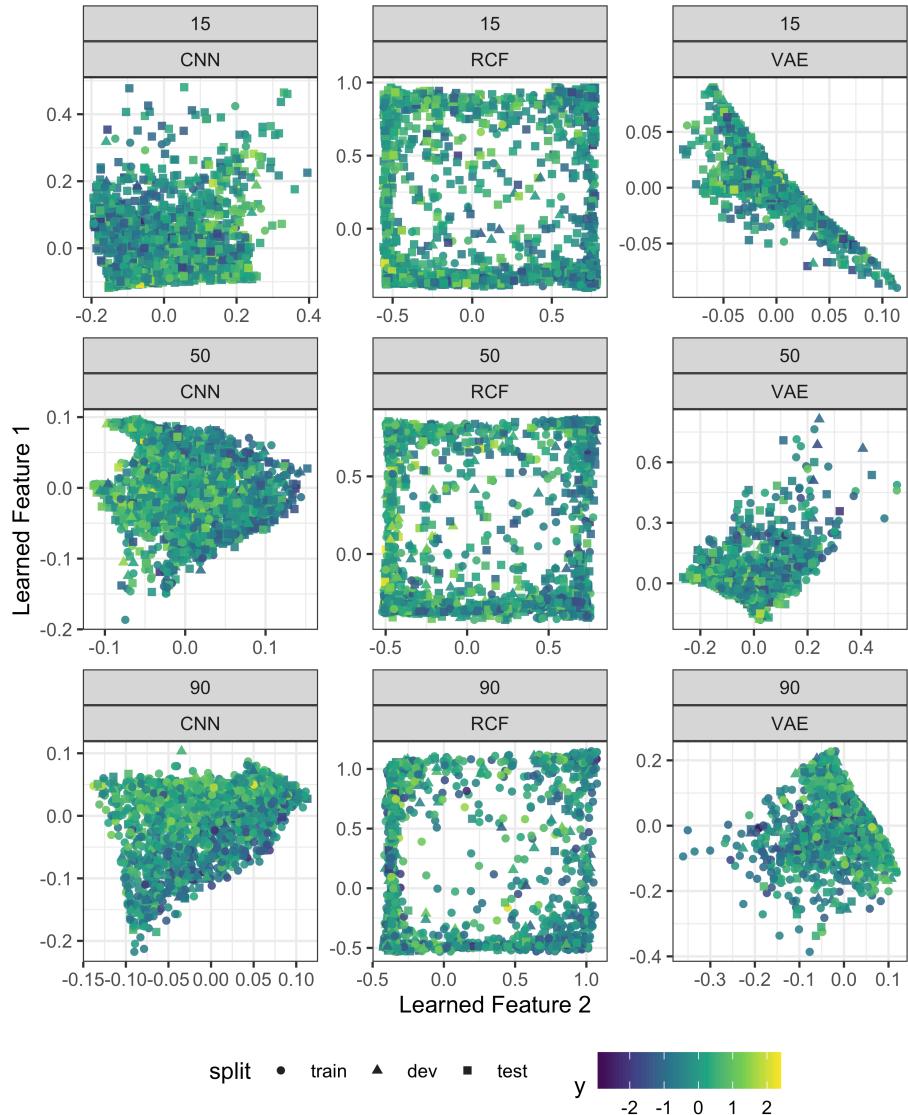


Figure 16: The coordinates of all samples in the first two dimensions after aligning the SCA-reduced features. A subset, along with the bootstrap glyphs, is shown in Figure 15.

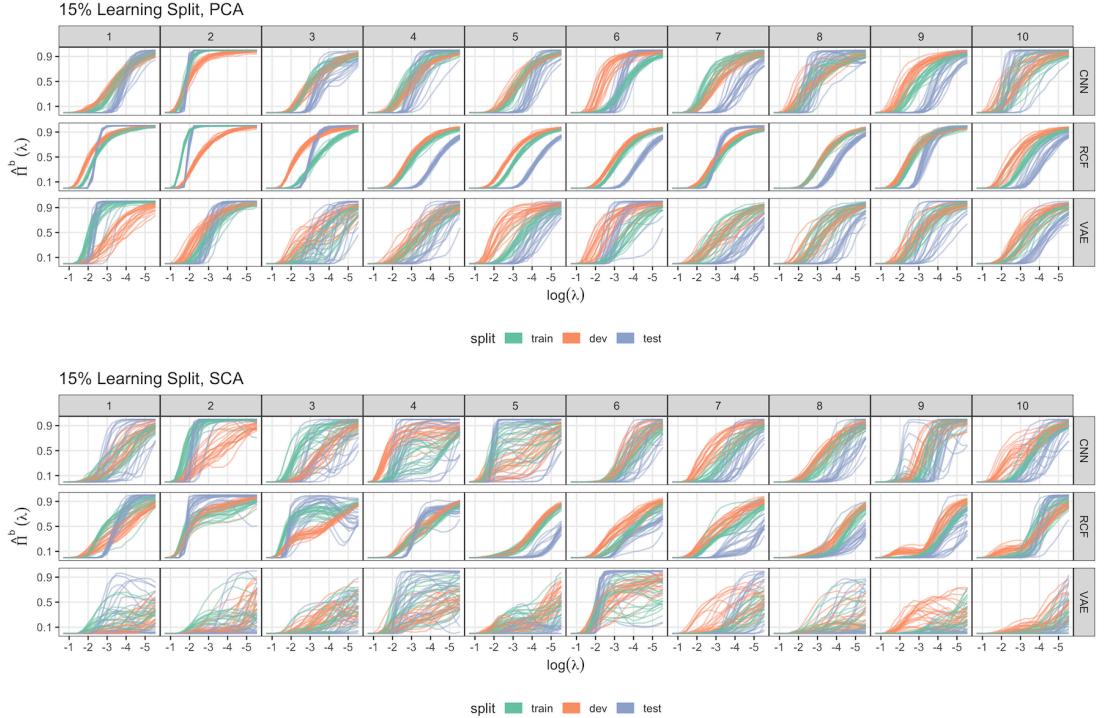


Figure 17: Selection paths for all models when using 15% of the data for feature learning. Most features, especially those from the CNN, have lower selection probability than the analogous features at other split sizes. Other key characteristics of the selection curves, like the effect of SVD vs. SCA reduction, and the stability of RCF features, remain the same.

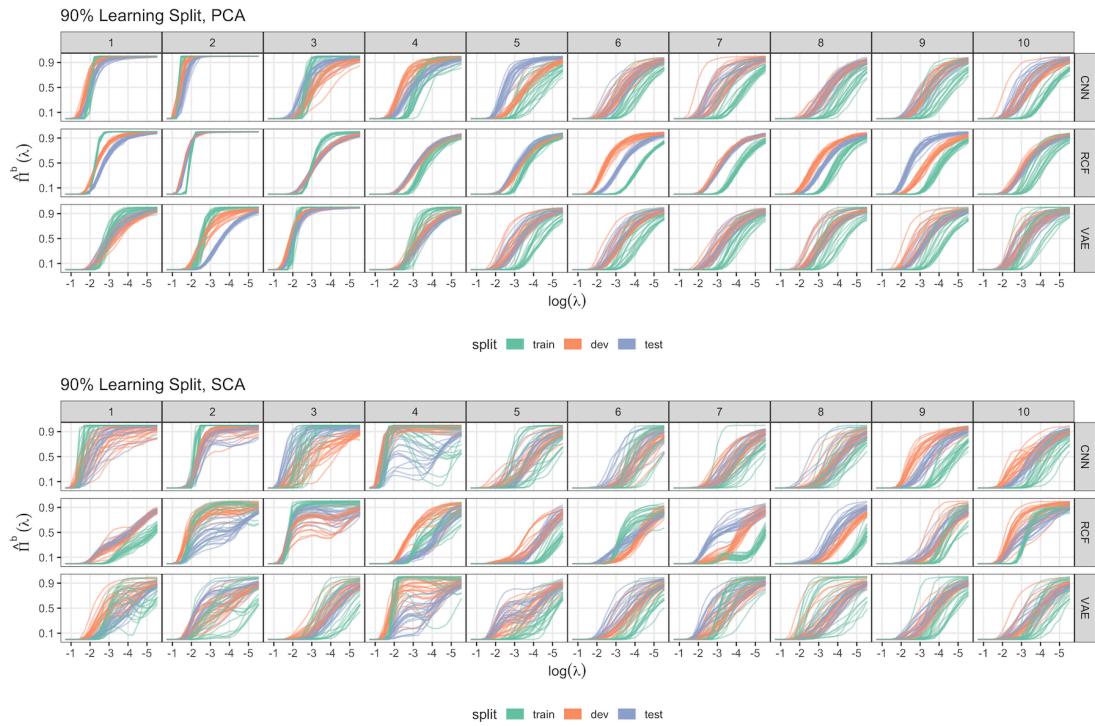


Figure 18: Selection paths for all models when using 90% of the data for feature learning. The top learned features across several methods are strongly related to the response.

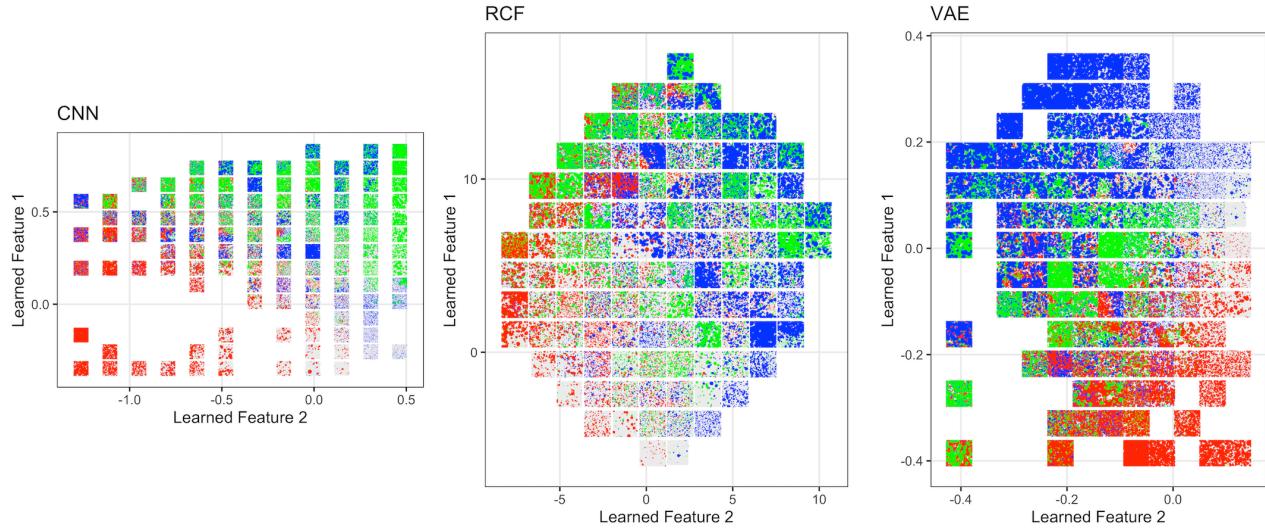


Figure 19: Patches overlaid on the learned features in Figure 14, to guide interpretation of regions of the learned feature space. The figure is read in the same way as Figure 11.

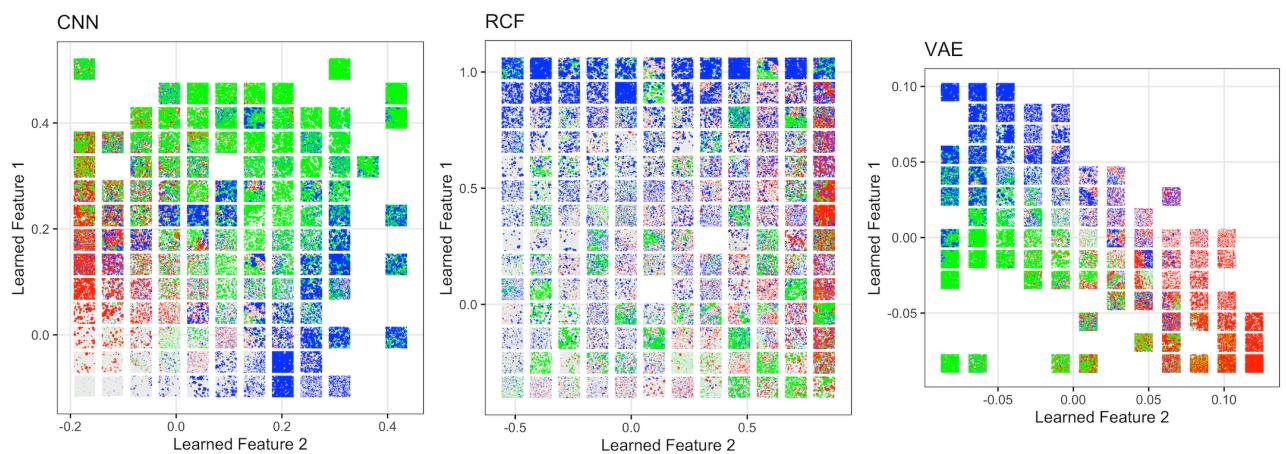


Figure 20: The analogous figure to 14, but using SCA dimensionality reduction instead.

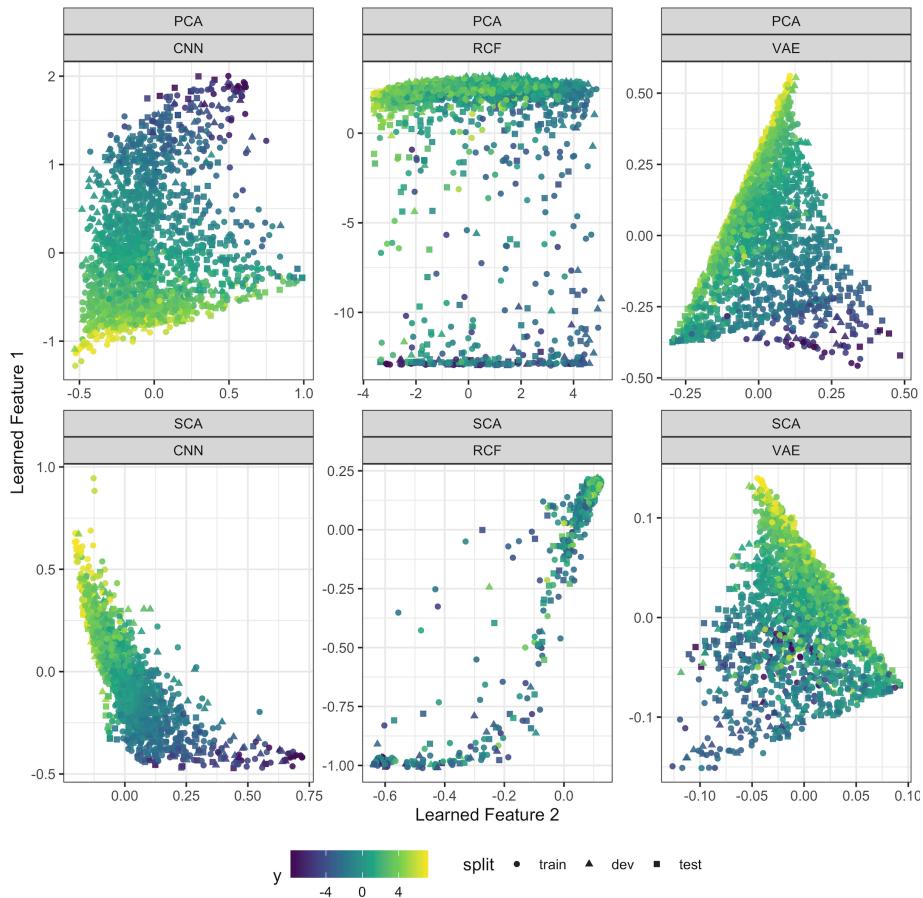


Figure 21: The version of Figure 10 showing all samples, and collapsing all glyphs to a point.

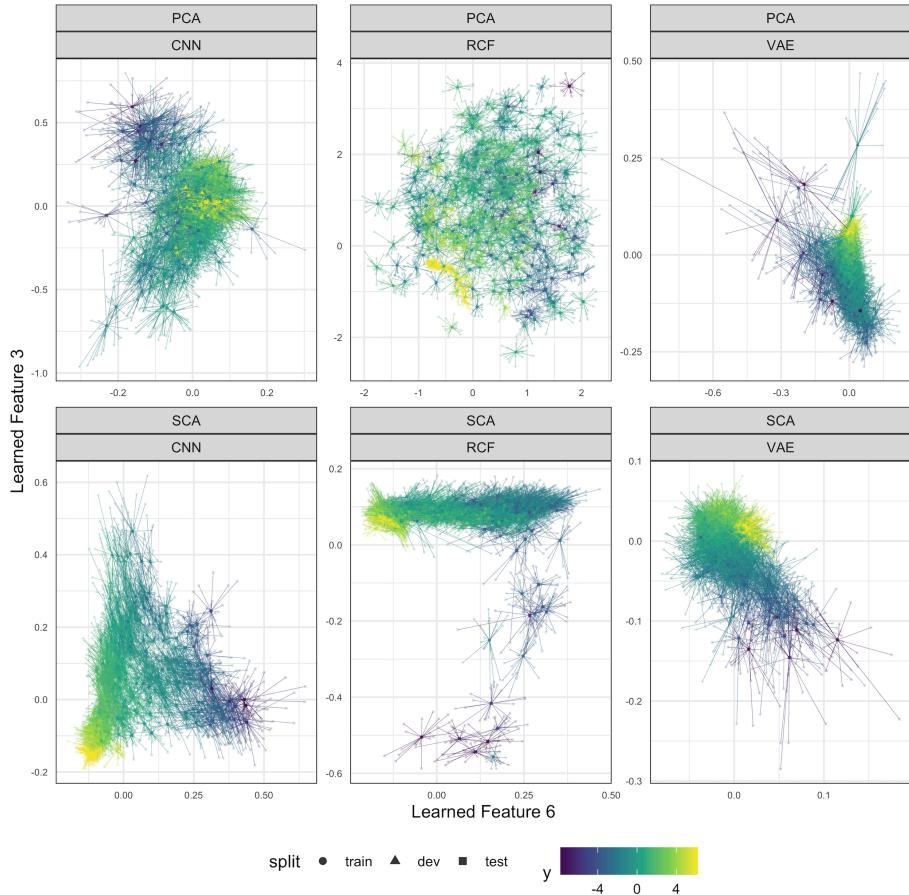


Figure 22: Examples of features that are found to be less important in the TNBC dataset application, according to the selection curves in Figure 9. Note that the relationship between the displayed locations and the measured y is harder to capture using a linear boundary, compared to those in Figure 10.

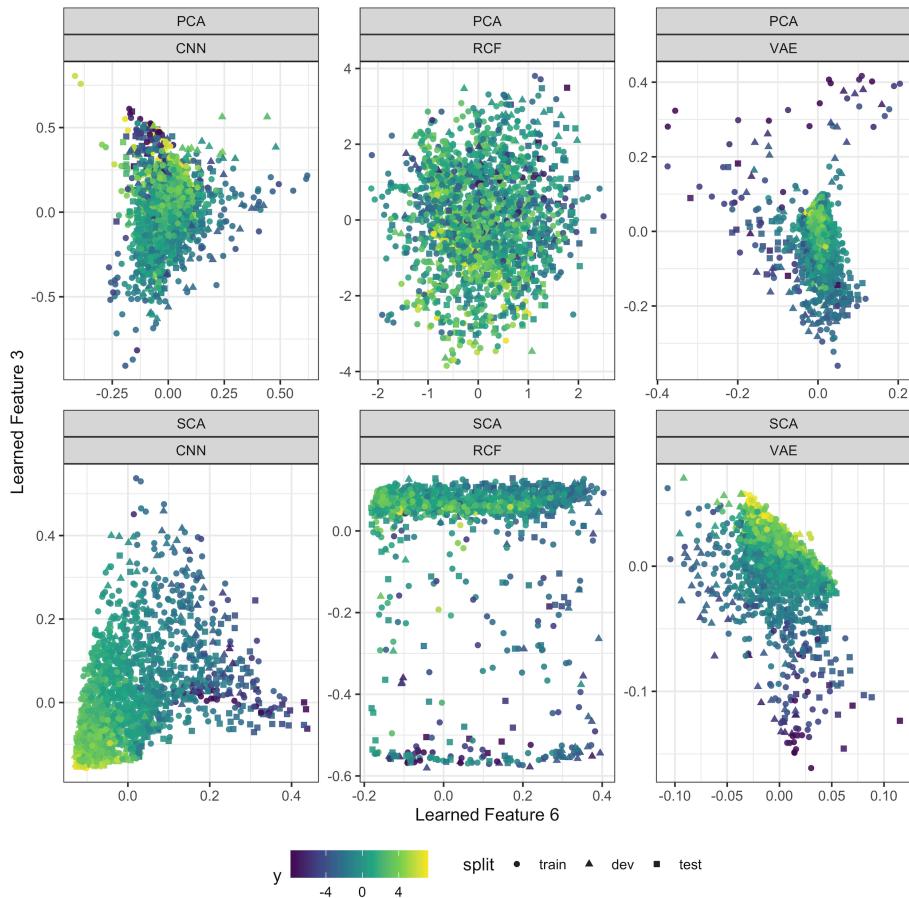


Figure 23: The version of Figure 22 showing all samples, and collapsing all glyphs to point.

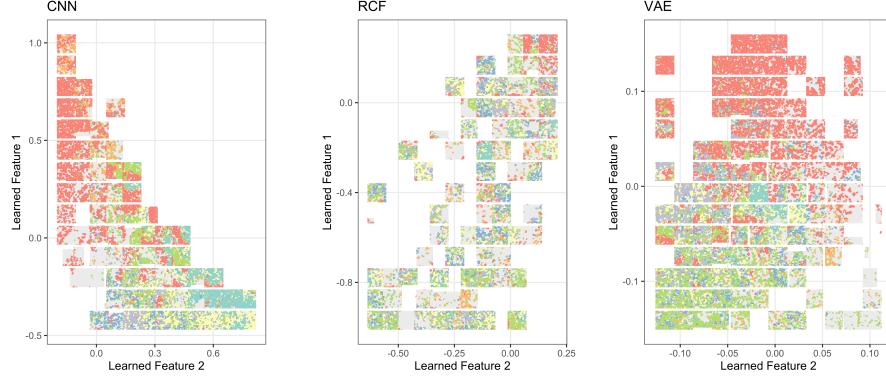


Figure 24: TNBC patches overlaid on the SCA-reduced aligned features, analogous to the PCA-reduced view in Figure 11.

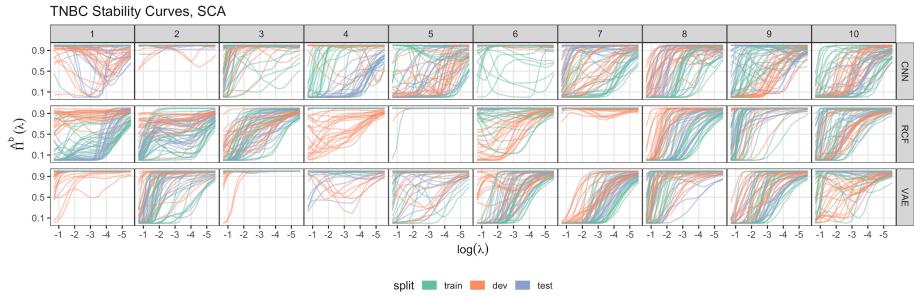


Figure 25: Selection curves for the TNBC dataset application when using SCA for dimensionality reduction.