Title:

Sequential compression of gene expression across dimensionalities and methods reveals no single best method or dimensionality

Authors:

Gregory P. Way[1,2,3] (0000-0002-0503-9348), Michael Zietz[2] (0000-0003-0539-630X), Vincent Rubinetti[2] (0000-0002-4655-3773), Daniel S. Himmelstein[2] (0000-0002- 3012-7446) and Casey S. Greene[2*] (0000-0001-8713-9213)

Affiliations:

[1]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

[2]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA 19104, USA.

[3]Imaging Platform, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.

Corresponding Author:

Casey S. Greene

10-131 SCTR 34th and Civic Center Blvd,

Philadelphia, PA 19104

Office: 215-573-2991

Fax: 215-573-9135

Keywords:

Machine Learning, Dimensionality Reduction, Latent Space, Gene Expression, Autoencoders, Compression, Neural Network Interpretation

Abstract:

Background

Unsupervised compression algorithms applied to gene expression data extract latent, or hidden, signals representing technical and biological sources of variation. However, these algorithms require a user to select a biologically-appropriate latent dimensionality. In practice, most researchers select a single algorithm and latent dimensionality. We sought to determine the extent by which using multiple dimensionalities across ensemble compression models improves biological representations.

Results

We compressed gene expression data from three large datasets consisting of adult normal tissue, adult cancer tissue, and pediatric cancer tissue. We compressed these data into many latent dimensionalities ranging from 2 to 200. We observed various tradeoffs across latent dimensionalities and compression models. For example, we observed high model stability between principal components analysis (PCA), independent components analysis (ICA), and non-negative matrix factorization (NMF). We identified more unique biological signatures in ensembles of denoising autoencoder (DAE) and variational autoencoder (VAE) models in intermediate latent dimensionalities. However, we captured the most pathway-associated features using all compressed features across algorithms and dimensionalities. Optimized at different latent dimensionalities, compression models detect generalizable gene expression signatures representing sex, neuroblastoma MYCN amplification, and cell types. In two

supervised machine learning tasks, compressed features optimized predictions at different latent dimensionalities.

Conclusions

There is no single best latent dimensionality or compression algorithm for analyzing gene expression data. Instead, using feature ensembles from different compression models across latent space dimensionalities optimizes biological representations.

Background:

Dimensionality reduction algorithms compress input data into feature representations that capture sources of variation. Applied to gene expression data, compression algorithms can identify latent biological representations and technical processes. These biological representations reveal important information about the samples and can help to generate hypotheses that are difficult or impossible to observe in the original genomic space. For example, applying PCA to a large cancer transcriptomic compendium determined the influence of copy number alterations on gene expression measurements [1]. Applying ICA to transcriptome data aggregated gene modules representing core pathways and hidden transcriptional programs [2,3]. Training NMF models using bulk gene expression data estimated cell type proportion [4,5]. DAEs have revealed latent signals characterizing oxygen exposure and transcription factor targets [6,7], and VAEs have identified biologically relevant latent features that discriminate cancer subtypes and drug response [8,9]. Additional latent variable approaches have been used to detect and remove technical artifacts, including batch effects [10,11]. Here, we focus on using compression to identify biological representations by analyzing

processed data with batch effect already mitigated. Nevertheless, a major challenge to all compression applications is the fundamental requirement that a researcher must determine the number of latent dimensions (*k*) to compress input data.

Instead, it is possible that different biological signatures are best captured at different latent space dimensionalities. To test this, we trained and evaluated various compression models across a wide range of latent space dimensionalities, from $k = 2$ to $k = 200$. We train PCA, ICA, NMF, DAE, and VAE models using RNAseq gene expression data. We selected these methods because they are either widely established in practice (PCA, NMF) or use neural networks that are rapidly growing in popularity (DAE, VAE). We included ICA to compare against PCA: it has been used in practice [7] and also allows us to compare multiple rotations of the low-dimensional space. We applied these methods to three different datasets. We applied these methods to three different datasets: The Cancer Genome Atlas (TCGA) PanCanAtlas [12], the Genome Tissue Expression Consortium Project (GTEx) [13], and the Therapeutically Applicable Research To Generate Effective Treatments (TARGET) Project [14].

We demonstrate various model tradeoffs in reconstruction cost, stability, and gene set coverage in training and testing sets across datasets, algorithms, and latent dimensionalities. We observe that several distinct gene expression signatures are optimized in various models spanning low, intermediate, and high latent dimensionalities. Our primary finding is that there is no single algorithm or dimensionality that is best for all purposes: Instead, using various latent dimensionalities and algorithms optimizes biological representations. Those who plan to apply these algorithms may consider training multiple models over multiple dimensionalities to avoid missing important biological features.

88 We name this sequential compression approach “BioBombe” after the large mechanical
89 device developed by Alan Turing and other cryptologists in World War II to decode encrypted
90 messages sent by Enigma machines. Using the BioBombe approach, we compress gene
91 expression input data with multiple latent dimensionalities to decipher and enhance biological
92 representations embedded within compressed gene expression features.
93
94 **Results:**
95 *BioBombe implementation*
96 We compressed processed RNAseq data from TCGA, GTEx, and TARGET using PCA, ICA,
97 NMF, DAE, and VAE across 28 different latent dimensionalities (*k*) ranging from *k* = 2 to *k* = 200.
98 We split each dataset into 90% training and 10% test sets balanced by cancer type or tissue
99 type and trained models using only the training data. We used real and permuted data and
100 initialized each model five times per latent dimensionality resulting in a total of 4,200 different
101 compression models (**Additional File 1: Figure S1**). We evaluated hyperparameters for DAE and
102 VAE models across dimensionalities and trained models using optimized parameter settings
103 (**Additional File 2**; **Additional File 1: Figure S2**). See **Fig. 1** for an outline of our approach. We
104 provide full BioBombe results for all compression models across datasets for both real [15–17]
105 and permuted data [18–20] in both training and test sets as publicly available resources (see
106 https://greenelab.github.io/BioBombe/).

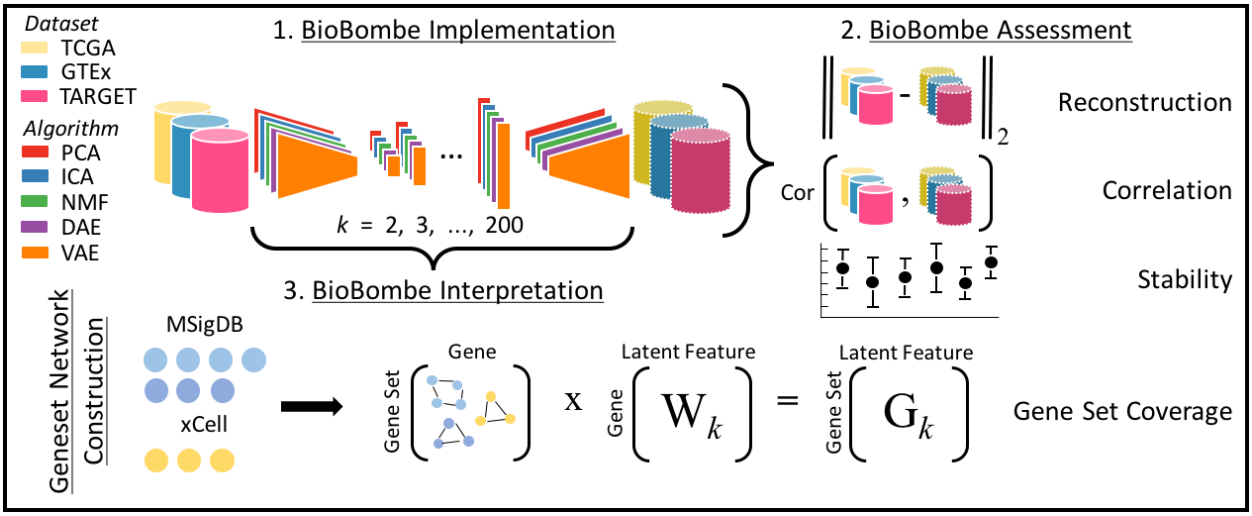

Figure 1: *Overview of the BioBombe approach.* We implemented BioBombe on three datasets using five different algorithms. We sequentially compressed input data into various latent dimensionalities. We calculated various metrics that describe different benefits and trade-offs of the algorithms. Lastly, we implemented a network projection approach to interpret the compressed latent features. We used MSigDB collections and xCell gene sets to interpret compressed features.

Assessing compression algorithm reconstruction

Reconstruction cost, a measurement of the difference between the input and output matrices, is often used to describe the ability of compression models to capture fundamental processes in latent space features that recapitulate the original input data. We tracked the reconstruction cost for the training and testing data partitions for all datasets, algorithms, latent dimensionalities, and random initializations. As expected, we observed lower reconstruction costs in models trained with real data and with higher latent dimensionalities (**Additional File 1: Figure S3**). Because PCA and ICA are rotations of one another, we used their identical scores as a positive control. All the compression algorithms had similar reconstruction costs, with the highest variability existing in low latent dimensionalities (**Additional File 1: Figure S3**).

126

127 *Evaluating model stability and similarity within and across latent dimensionalities*

128 We applied singular vector canonical correlation analysis (SVCCA) to algorithm weight
129 matrices to assess model stability within algorithm initializations, and to determine model
130 similarity between algorithms [21]. Briefly, SVCCA calculates similarity between two
131 compression algorithm weight matrices by learning appropriate linear transformations and
132 iteratively matching the highest correlating features. Training with TCGA data, we observed
133 highly stable models within algorithms and within all latent dimensionalities for PCA, ICA, NMF
134 (along the matrix diagonal in **Fig 2a**). VAE models were also largely stable, with some decay in
135 higher latent dimensionalities. However, DAE models were unstable, particularly at low latent
136 dimensionalities (**Fig 2a**). We also compared similarity across algorithms. Because PCA and ICA
137 are rotations of one another, we used the high stability as a positive control for SVCCA
138 estimates. NMF was also highly similar to PCA and ICA, particularly at low latent
139 dimensionalities (**Fig. 2a**). VAE models were more similar to PCA, ICA, and NMF than DAE
140 models, particularly at low latent dimensionalities, and the instability patterns within DAE
141 models also led to large differences across algorithms (**Fig. 2a**). We observed similar patterns in
142 GTEx and TARGET data, despite TARGET containing only about 700 samples (**Additional File 1:**
143 **Figure S4**).

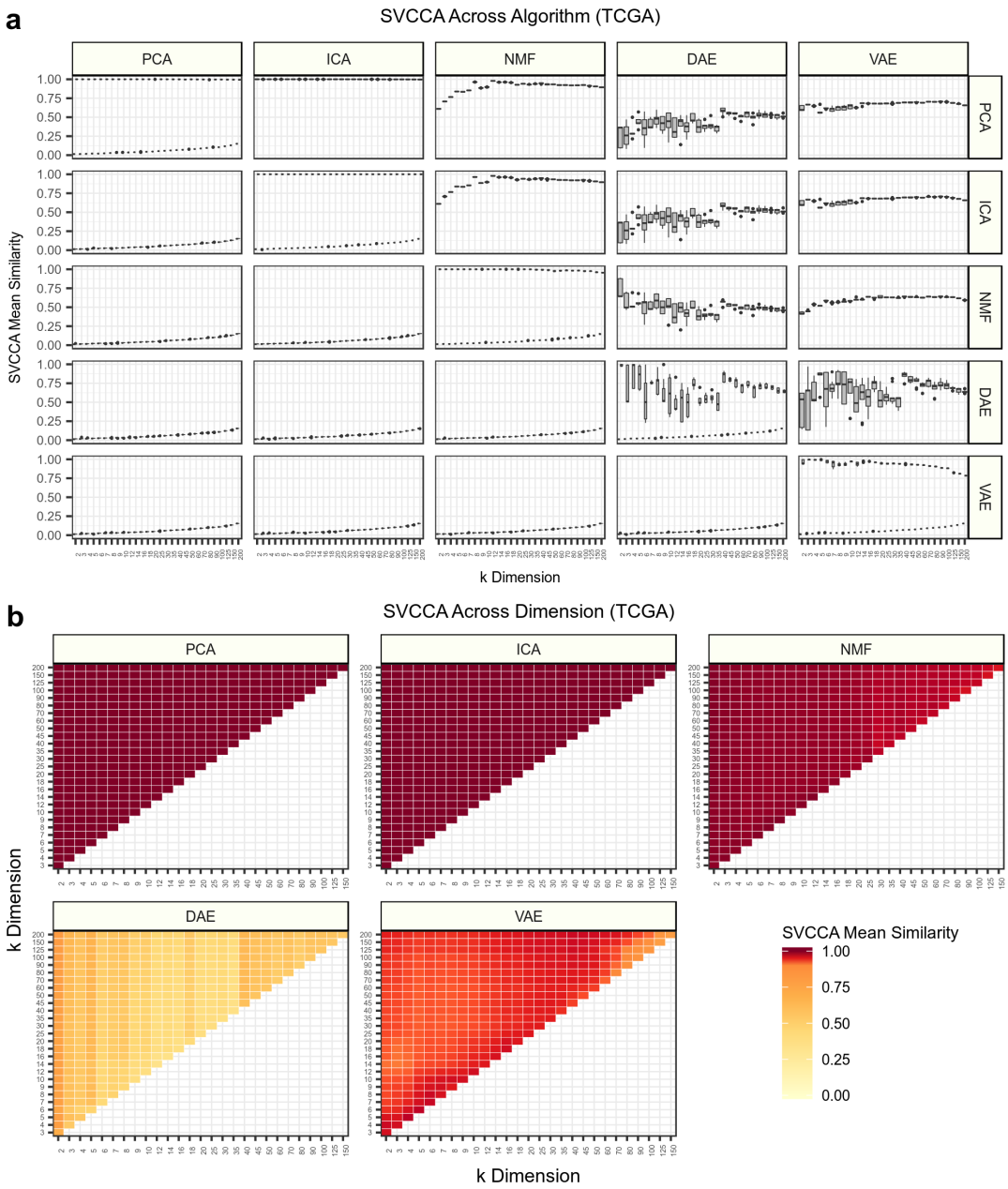

Figure 2: *Assessing algorithm and dimensionality stability with singular vector canonical correlation analysis (SVCCA).* **(a)** SVCCA applied to the weight matrices learned by each compression algorithm in gene expression data from The Cancer Genome Atlas (TCGA). The mean of all canonical correlations comparing independent iterations is shown. The distribution of mean similarity represents a comparison of all pairwise iterations within and across algorithms. The upper triangle represents SVCCA applied to real gene expression data, while the lower triangle represents permuted expression data. Both real and permuted data are plotted along the diagonal. **(b)** Mean correlations of all iterations within algorithms but across *k* dimensionalities. SVCCA will identify min(i, j) canonical vectors for latent dimensionalities k_i and k_j. The mean of all pairwise correlations is shown for all combinations of *k* dimensionalities.

155 We also used SVCCA to compare the similarity of weight matrices across latent
156 dimensionalities. Both PCA and ICA found highly similar solutions (**Fig. 2b**). This is expected
157 since the solutions are deterministic and are arranged with decreasing amounts of variance.
158 NMF also identified highly similar solutions in low dimensionalities, but solutions were less
159 similar in higher dimensionalities. DAE solutions were the least similar, with intermediate
160 dimensionalities showing the lowest mean similarity. VAE models displayed relatively high
161 model similarity, but there were regions of modest model stability in intermediate and high
162 dimensionalities (**Fig. 2b**). We observed similar patterns in GTEx and TARGET data (**Additional**
163 **File 1: Figure S5**).
164
165 *BioBombe features capture various gene expression representations across different latent*
166 *dimensionalities*
167 We tested the ability of features derived using the BioBombe approach to isolate
168 various biological representations. First, we tested the ability to differentiate sample sex; which
169 has been previously observed to be captured in latent space features [8,22,23]. We performed
170 a two-tailed t-test comparing male and female samples in the GTEx test set across all
171 initializations, algorithms, and latent dimensionalities. We optimally identified this phenotype
172 in higher latent dimensionalities, particularly in NMF and VAE models (**Fig. 3a**). The top feature
173 separating GTEx males and females was NMF feature 111 in k = 200 (t = 44.5, p = 7.3 x 10^{-176})
174 (**Fig 3b**). We examined the genes that contribute with high weight to this feature and found
175 only three genes had substantial influence. These three genes all had high positive weights and
176 were encoded on the Y chromosome. We performed the same approach using BioBombe

177 features to identify sex features in TCGA test data (**Fig. 3c**). The top latent dimensionality
178 identified was not consistent across algorithms. The top feature distinguishing TCGA males and
179 females was ICA feature 53 in the k = 90 model (t = 4.9, p = 2.0 x 10^{-6}) (**Fig. 3d**). The separation
180 was not as strong using the more complex TCGA data, but the top 10 gene weights were all
181 encoded on the X chromosome.
182 We also tested the ability of features derived from the BioBombe approach to
183 distinguish MYCN amplification in neuroblastoma (NBL) tumors. MYCN amplification is a
184 biomarker associated with poor prognosis in NBL patients [24]. We performed a two-tailed t-
185 test comparing MYCN amplified vs. MYCN not amplified NBL tumors for each of the latent
186 features derived from the full set of TARGET samples. Each algorithm discovered optimal signal
187 at various latent dimensionalities, but the top scoring features were generally identified in VAE
188 and NMF models with large latent space dimensionalities (**Fig. 3e**). Although there were some
189 potentially mischaracterized samples, feature 12 in VAE k = 50 robustly separated MYCN
190 amplification status in NBL tumors (t = -18.5, p = 6.6 x 10^{-38}) (**Fig. 3f**). This feature also
191 distinguished MYCN amplification status in NBL cell lines [25] that were previously not used in
192 training the compression model or for feature selection (t = -3.2, p = 4.2 x 10^{-3}) (**Fig. 3g**).
193
194
195
196
197
198

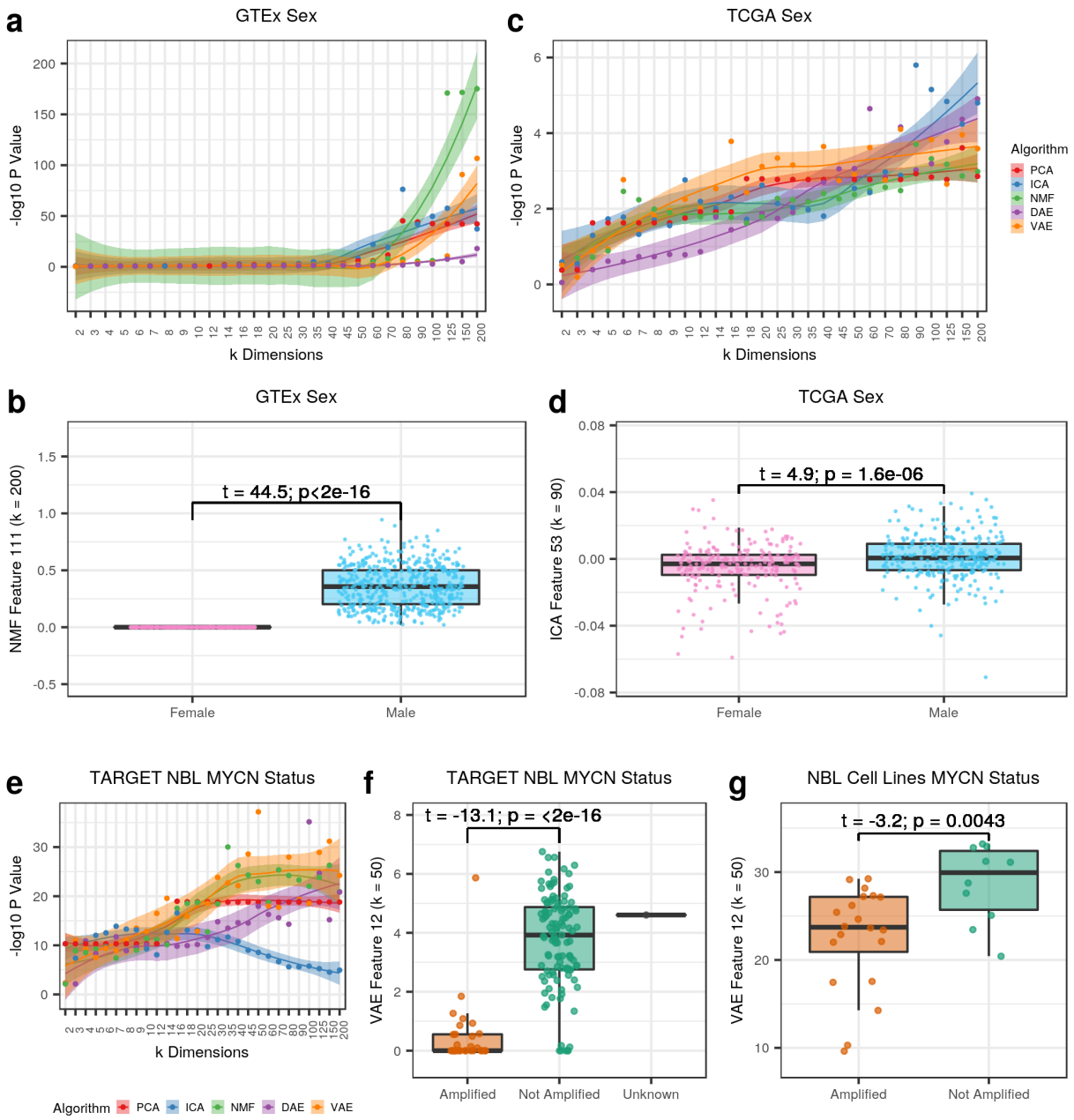

Figure 3: *Using BioBombe as a signature discovery tool.* Detecting GTEx sample sex across **(a)** various latent dimensionalities and algorithms, and **(b)** the latent feature with the highest enrichment. Detecting TCGA patient sex across **(c)** various latent dimensionalities, and **(d)** the latent feature with the highest enrichment. Detecting TARGET MYCN amplification in neuroblastoma (NBL) tumors **(e)** across various latent dimensionalities, and **(f)** the latent feature with the highest enrichment. **(g)** Applying the MYCN signature to an external dataset of NBL cell lines implicates MYCN amplified cell lines.

209 *Assessing gene set coverage of compression models*

210 We used gene sets from Molecular Signatures Database (MSigDB) and xCell [26–28] to
211 interpret biological signals activated in compressed features across all latent dimensionalities,
212 algorithms, and initializations. We applied a network projection approach to all compression
213 algorithm weight matrices to determine gene set coverage. Briefly, we projected all
214 compressed features onto a gene set network and assigned gene sets with the highest
215 enrichment that passed an adjusted statistical significance threshold to each compressed
216 feature (see methods for more details). We tracked coverage of three MSigDB gene set
217 collections representing transcription factor (TF) targets, cancer modules, and Reactome
218 pathways across latent dimensionalities in TCGA data (**Fig. 4**). In all cases, we observed higher
219 gene set coverage in models with larger latent dimensionalities. Considering individual models,
220 we observed high coverage in PCA, ICA, and NMF. In particular, ICA outperformed all other
221 algorithms (**Fig. 4a**). However, while these methods showed the highest coverage, the features
222 identified had relatively low enrichment scores compared to AE models (**Additional File 1:**
223 **Figure S6**).

224 Aggregating all five random initializations into ensemble models, we observed
225 substantial coverage increases, especially for AEs (**Fig. 4b**). VAE models had high coverage for all
226 gene sets in intermediate dimensions, while DAE improved in higher dimensions. However, at
227 the highest dimensions, ICA demonstrated the highest coverage. NMF consistently had the
228 highest enrichment scores, but the lowest coverage (**Fig. 4b**). When considering all models
229 combined (forming an ensemble of algorithm ensembles) within latent dimensionalities, we

observed substantially increased coverage of all gene sets. However, most of the unique gene sets were contributed by the AE models (**Fig. 4c**).

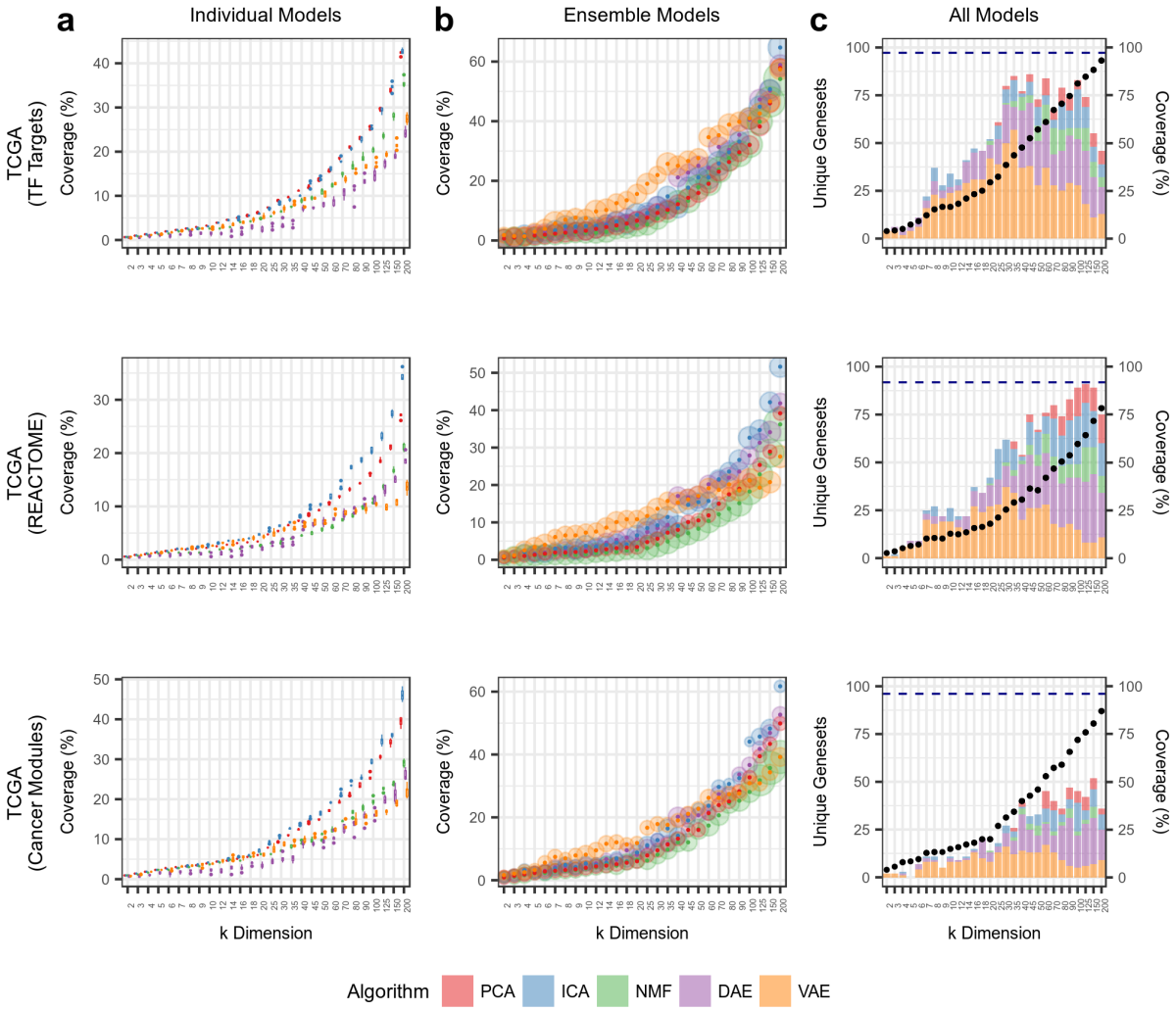

Figure 4: *Assessing gene set coverage of specific gene set collections.* Tracking results in TCGA data for three gene set collections representing transcription factor (TF) targets (C3TFT), Reactome pathways (C2CPREACTOME), and cancer modules (C4CM). **(a)** Tracking coverage in individual models, which represents the distribution of scores across five algorithm iterations. **(b)** Tracking coverage in ensemble models, which represents coverage after combining all five iterations into a single model. The size of the point represents relative enrichment strength. (**c**) Tracking coverage in all models combined within *k* dimensionalities. The number of algorithm-specific unique gene sets identified is shown as bar charts. Coverage for all models combined across all *k* dimensionalities is shown as a dotted navy blue line.

243 Lastly, when we aggregated all BioBombe features across all algorithms and all latent
244 dimensionalities together into a single model, we observed the highest gene set coverage (**Fig.**
245 **4c**). These patterns were consistent across other gene set collections and datasets (**Additional**
246 **File 1: Figure S7**). In general, while models compressed with larger latent space dimensionalities
247 had higher gene set coverage, many individual gene sets were captured with the highest
248 enrichment in models with low and intermediate dimensionalities (**Additional File 1: Figure S8**).
249 These results did not reveal a best method or dimensionality: Various biological representations
250 are best discovered by using various compression algorithms with various latent space
251 dimensionalities.
252
253 *Observing the latent dimensionality of specific tissue and cell type signatures*
254 We measured the Pearson correlation between all samples' gene expression input and
255 reconstructed output. As expected, we observed increased mean correlation and decreased
256 variance as the latent dimensionalities increased in TCGA data (**Fig. 5a**). We also observed
257 similar patterns in GTEx and TARGET data (**Additional File 1: Figure S9**). Across all datasets, in
258 randomly permuted data, we observed correlations near zero (**Additional File 1: Figure S9**). The
259 correlation with real data was not consistent across all algorithms as PCA, ICA, and NMF
260 generally outperformed the AE models.
261 We tracked correlation differences across latent dimensionalities to determine the
262 dimensionality at which specific sample types are initially detected. Most cancer types,
263 including breast invasive carcinoma (BRCA) and colon adenocarcinoma (COAD), displayed
264 relatively gradual increases in sample correlation as the latent dimensionality increased (**Fig.**

265 **5b**). However, in other cancer types, such as low grade glioma (LGG), pheochromocytoma and
266 paraganglioma (PCPG), and acute myeloid leukemia (LAML), we observed large correlation
267 gains with a single increase in latent dimensionality (**Fig. 5c**). We also observed similar
268 performance spikes in GTEx data for several tissues including liver, pancreas, and blood (**Fig.**
269 **5d**). This sudden and rapid increase in correlation in specific tissues occurred at different latent
270 dimensionalities for different algorithms, but was consistent across algorithm initializations.
271 To determine if this rapid increase was a result of models learning specific biological
272 representations or if this observation represented a technical artifact, we more closely
273 examined the sharp increase in GTEx blood tissue correlation between latent space
274 dimensionalities 2 and 3 in VAE models (See **Fig. 5d**). We hypothesized that a difference in
275 reconstruction for a specific tissue at such a low dimensionality could be driven by a change in
276 the cell types captured by the model. We applied network projection of xCell gene sets to all
277 compressed features in both VAE models. xCell gene sets represent computationally derived
278 cell type signatures [27]. The top features identified for the VAE k = 2 model included skeletal
279 muscle, keratinocyte, and neuronal gene sets (**Fig. 6a**). Skeletal muscle was the most significant
280 gene set identified likely because it the tissue with the most samples in GTEx. Similar gene sets
281 were enriched in the k = 3 model, but we also observed enrichment for a specific neutrophil
282 gene set ("Neutrophils_HPCA_2") (**Fig. 6a**). Neutrophils represent 50% of all blood cell types,
283 which may explain the increased correlation in blood tissue observed in VAE k = 3 models. The
284 features implicated using the network projection approach were similar to an
285 overrepresentation analysis using high weight genes in both tails of the VAE k = 3 feature
286 (**Additional File 1: Figure S10**).

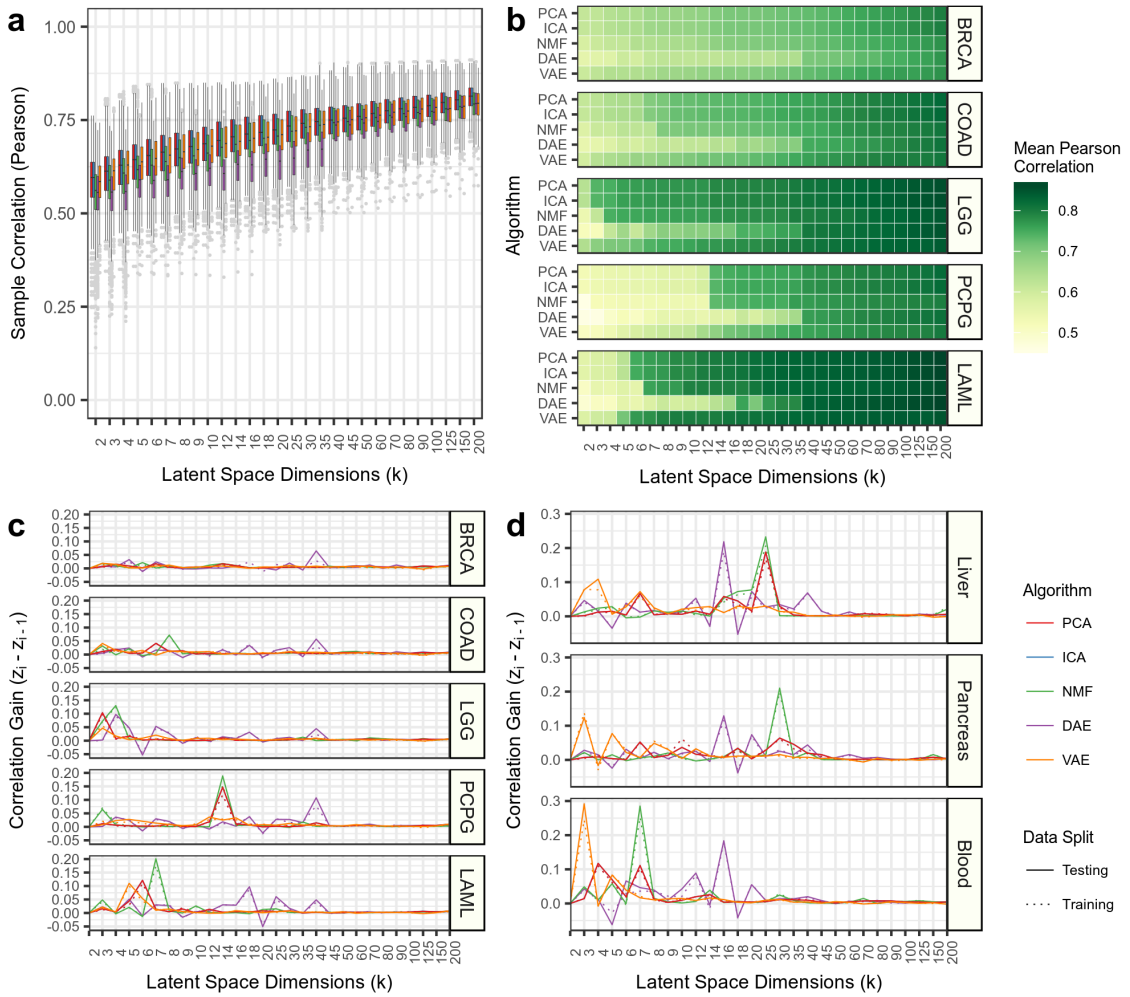

Figure 5: *Different latent dimensionalities implicate different tissue types.* **(a)** Sample Pearson correlation for all data in the testing data partition for The Cancer Genome Atlas (TCGA). The different algorithms follow the legend provided in panel d. **(b)** Mean Pearson correlation for select cancer types in the testing data partition. Pearson correlation gain between sequential latent dimensionalities for **(c)** select cancer types in TCGA and **(d)** select tissue-types in GTEx.

We also calculated the mean absolute value z scores for xCell gene sets in all compression features for both VAE models with *k* = 2 and *k* = 3 dimensionalities (**Fig. 6b**). Again, we observed skeletal muscle, keratinocytes, and neuronal gene sets to be enriched in both models. However, we also observed a cluster of monocyte gene sets (including "Monocytes_FANTOM_2") with enrichment in *k* = 3, but low enrichment in *k* = 2 (**Fig. 6b**).

299 Monocytes are also important cell types found in blood, and it is probable these signatures also
300 contributed to the increased correlation for the reconstructed blood samples in VAE *k* = 3
301 models. We provide the full list of xCell gene set genes for the neutrophil and monocyte gene
302 sets that intersected with the GTEx data in **Additional File 3**.
303 We scanned all other algorithms and latent dimensionalities to identify other
304 compression features with high enrichment scores in the "Neutrophils_HPCA_2" (**Fig. 6c**) and
305 "Monocytes_FANTOM_2" gene sets (**Fig. 6d**). We observed stronger enrichment of the
306 "Neutrophil_HPCA_2" gene set in AE models compared to PCA, ICA, and NMF, especially at
307 lower latent dimensionalities. We observed the highest score for the "Neutrophil_HPCA_2"
308 gene set at *k* = 14 in VAE models (**Fig. 6c**). The top VAE feature at *k* = 14 correlated strongly with
309 the VAE feature learned at *k* = 3 (**Additional File 1: Figure S10**). Conversely, PCA, ICA, and NMF
310 identified the "Monocytes_FANTOM_2" signature with higher enrichment than the AE models
311 (**Fig. 6d**). We observed a performance spike at *k* = 7 for both PCA and NMF models, but the
312 highest enrichment for "Monocytes_FANTOM_2" occurred at *k* = 200 in NMF models.
313
314 *Validating GTEx neutrophil and monocyte signatures in external datasets*
315 We downloaded a processed gene expression dataset (GSE103706) that applied two
316 treatments to induce neutrophil differentiation in two leukemia cell lines [29]. We hypothesized
317 that projecting the dataset on the "Neutrophil_HPCA_2" signature would reveal differential
318 scores in the treated cell lines. We observed large differences in sample activations of treated
319 vs untreated cell lines in the top Neutrophil signature (VAE *k* = 14) (**Fig. 6e**). We also tested the
320 "Monocytes_FANTOM_2" signature on a different publicly available dataset (GSE24759)

321 measuring gene expression of isolated cell types undergoing hematopoiesis [30]. We observed
322 increased scores for isolated monocyte cell population (MONO2) and relatively low scores for
323 several other cell types for top VAE features (**Fig. 6f**).

324 We applied the top signatures for the neutrophil and monocyte gene sets to each
325 external dataset (see **Fig. 6c, d**). We observed variable enrichment patterns across different
326 algorithms and latent dimensionalities (**Additional File 1: Figure S11a**). These separation
327 patterns were associated with network projection scores in NMF models, but were not
328 consistent with other algorithms (**Additional File 1: Figure S11b**). Taken together, in this
329 analysis we determined that 1) adding a single latent dimensionality that captured Neutrophil
330 and Monocyte signatures improved signal detection in GTEx blood, 2) these gene expression
331 signatures are enhanced at different latent dimensionalities and by different algorithms, and 3)
332 these signatures generalized to external datasets that were not encountered during model
333 training.

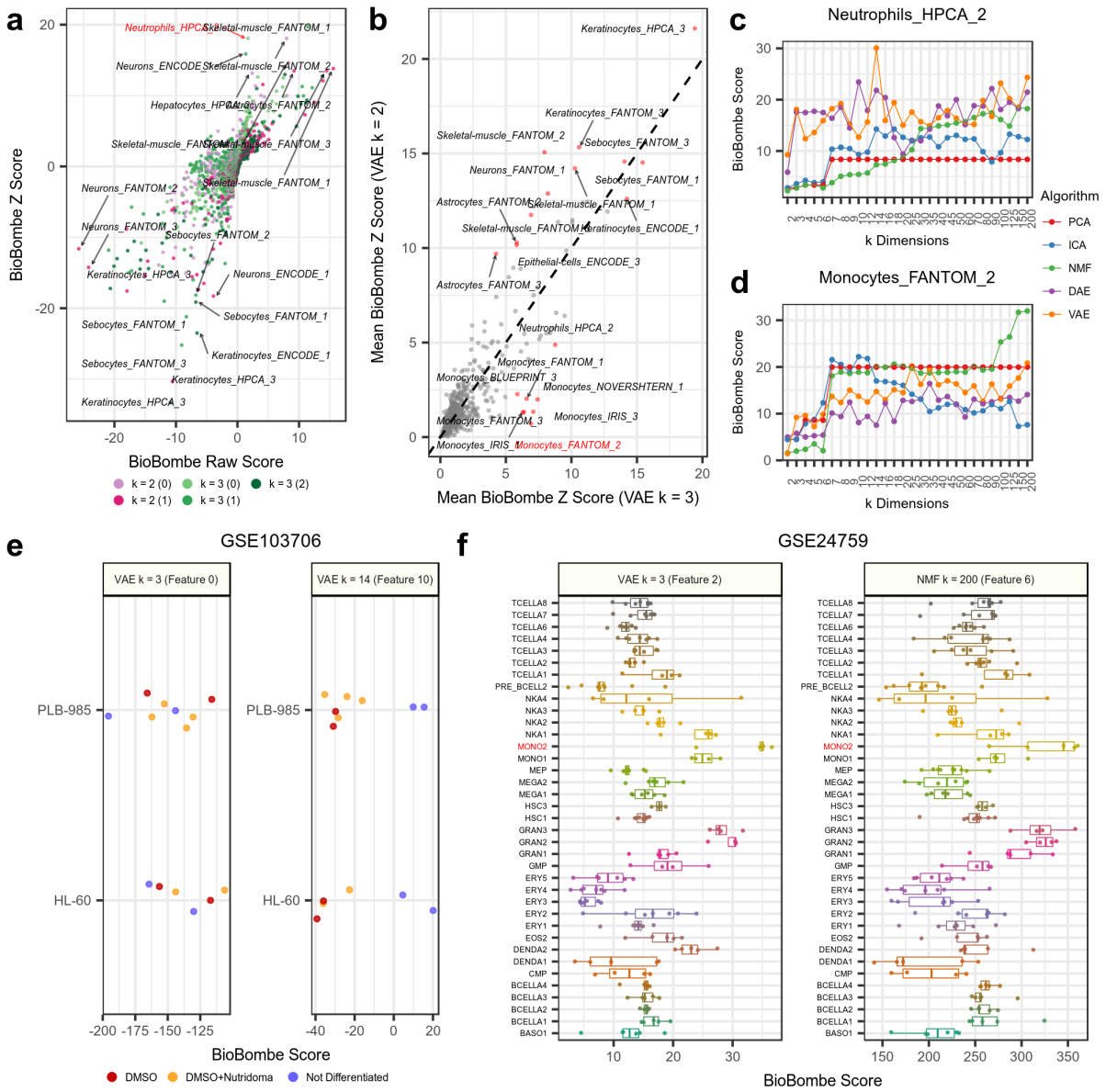

Figure 6: *Interpreting blood cell types in GTEx using xCell gene sets.* **(a)** Comparing BioBombe scores of all compressed latent features for variational autoencoder (VAE) models when bottleneck dimensionalities are set to *k* = 2 and *k* = 3. **(b)** Comparing mean BioBombe Z scores of aggregated latent features across two VAE models with *k* dimensionalities 2 and 3. Tracking the BioBombe Z scores of **(c)** "Neutrophils_HPCA_2" and **(d)** "Monocytes_FANTOM_2" gene sets across dimensionalities and algorithms. Only the top scoring feature per algorithm and dimensionality is shown. **(e)** Projecting the VAE feature *k* = 3 feature and the highest scoring feature (VAE *k* = 14) that best captures a neutrophil signature to an external dataset measuring neutrophil differentiation treatments (GSE103706). **(f)** Projecting the VAE *k* = 3 feature that best captures monocytes and the feature of the top scoring model (NMF *k* = 200) to an external dataset of isolated hematopoietic cell types (GSE24759).

347 *Using BioBombe features in supervised learning applications*

348 We used BioBombe compressed features in two supervised machine learning tasks.
349 First, we trained logistic regression models using compressed BioBombe features from
350 individual model iterations as input to predict each of the 33 different TCGA cancer types.
351 Nearly all cancer types could be predicted with high precision and recall (**Additional File 1:**
352 **Figure S12**). We observed multiple performance spikes at varying latent dimensionalities for
353 different cancer types and algorithms, which typically occurred in small latent dimensionalities
354 (**Fig. 7a**). Next, we input BioBombe features into the supervised classifier to predict samples
355 with alterations in the top 50 most mutated genes in TCGA (**Additional File 1: Figure S13**). We
356 focused on predicting four cancer genes and one negative control; *TP53*, *PTEN*, *PIK3CA*, *KRAS*,
357 and *TTN* (**Fig. 7b**). *TTN* is a particularly large gene and is associated with a high passenger
358 mutation burden and should provide no predictive signal [31]. As expected, we did not observe
359 any signal in predicting *TTN* (**Fig. 7b**). Again, we observed performance increases at varying
360 latent dimensionalities across algorithms. However, predictive signal for mutations occurred at
361 higher latent dimensionalities compared to cancer types (**Fig. 7c**). Compared to features trained
362 within algorithm and within iteration, an ensemble of five VAE models and an ensemble of five
363 models representing one iteration of each algorithm (PCA, ICA, NMF, DAE, and VAE) identified
364 cancer type and mutation status in earlier dimensionalities compared to single model iterations
365 (**Fig 7c**). We also tracked the logistic regression coefficients assigned to each compression
366 feature. DAE models consistently displayed sparse models, and the VAE ensemble and model
367 ensemble also induced high sparsity (**Fig. 7d**).

368 Lastly, we trained logistic regression classifiers using all 30,850 BioBombe features
369 generated across iterations, algorithms, and latent dimensionalities. These models were sparse
370 and high performing; comparable to logistic regression models trained using raw features (**Fig.**
371 **7e**). Of all 30,850 compressed features in this model, only 317 were assigned non-zero weights
372 (1.03%). We applied the network projection approach using Hallmark gene sets to interpret the
373 biological signatures of the top supervised model coefficients. The top positive feature was
374 derived from a VAE trained with *k* = 200. The top hallmarks of this feature included
375 "ESTROGEN_RESPONSE_EARLY", "ESTROGEN_RESPONSE_LATE", and "P53_PATHWAY". The top
376 negative feature was derived from a VAE trained with *k* = 150 and was associated with hallmark
377 gene sets including "BILE_ACID_METABOLISM", "EPITHELIAL_MESENCHYMAL_TRANSITION",
378 and "FATTY_ACID_METABOLISM". **Additional File 4** includes a full list of logistic regression
379 coefficients and hallmark network projection scores. Overall, the features selected by the
380 supervised classifier were distributed across algorithms and latent dimensionalities suggesting
381 that combining signatures across dimensionalities and algorithms provided the best
382 representation of the signal (**Fig. 7f**).
383
384
385
386

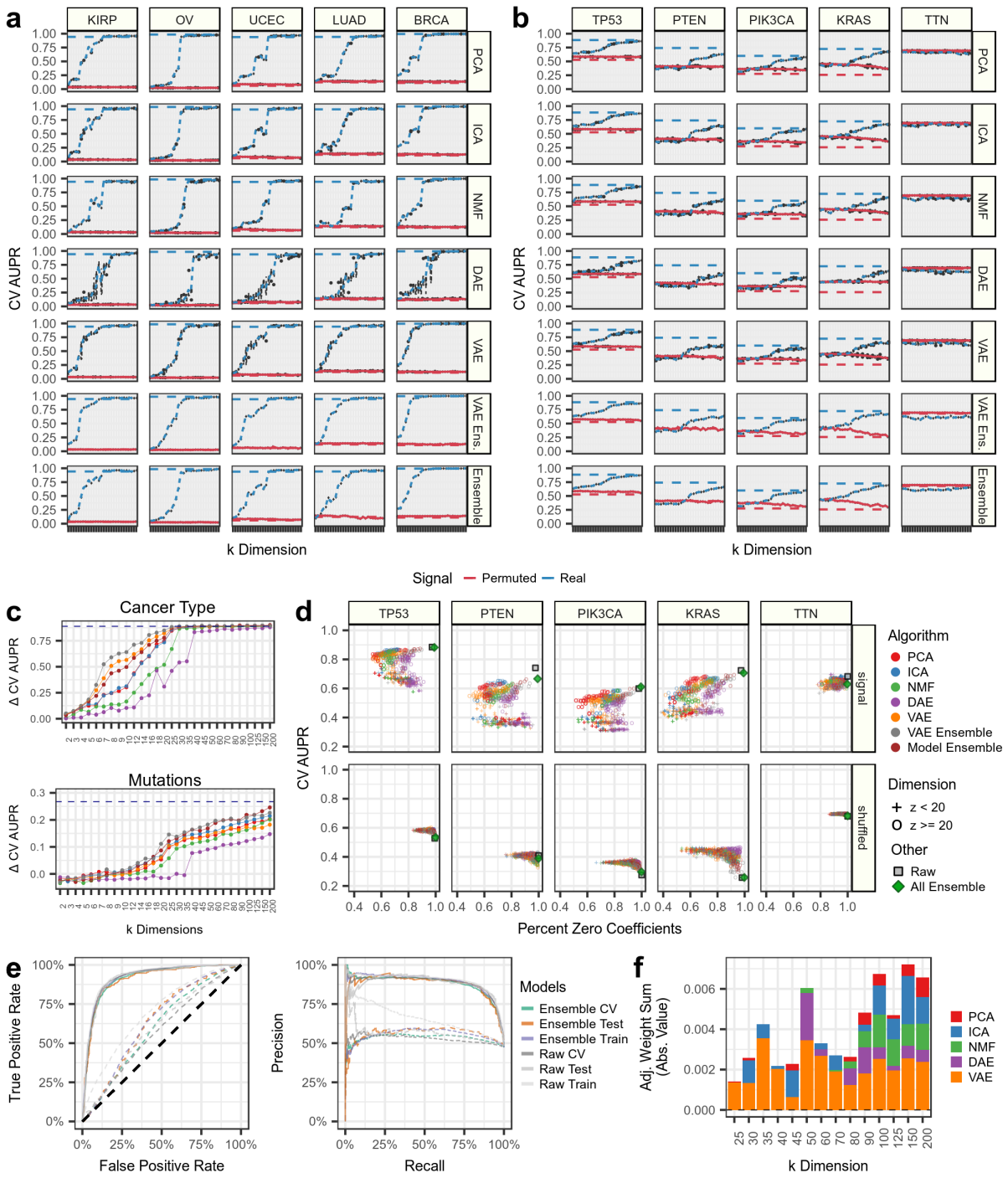

Figure 7: *Using BioBombe sequential compression in The Cancer Genome Atlas (TCGA) as features in supervised machine learning tasks*. Predicting **(a)** cancer-type status and **(b)** gene mutation status for select cancer-types and important cancer genes using five compression algorithms and two ensemble models. The area under the precision recall (AUPR) curve for cross validation (CV) data partitions is shown. The blue lines represent predictions made with permuted data input into each compression algorithm. The dotted lines represent AUPR on

untransformed RNAseq data. The dotted gray line represents a hypothetical random guess. **(c)** Tracking the average change in AUPR between real and permuted data across latent dimensionalities and compression models in predicting (*top*) cancer types and (*bottom*) mutation status. The average includes the five cancer types and mutations tracked in panels a and b. **(d)** Tracking the sparsity and performance of supervised models using BioBombe compressed features in real and permuted data. **(e)** Performance metrics for the all-compression feature ensemble model predicting *TP53* alterations. (*left*) Receiver operating characteristic (ROC) and (*right*) precision recall curves are shown. **(f)** The average absolute value weight per algorithm for the all-compression-feature ensemble model predicting *TP53* alterations. The adjusted scores are acquired by dividing by the number of latent dimensionalities in the given model.

Discussion:

Our primary observation is that compressing complex gene expression data using multiple latent dimensionalities and algorithms improves biological representations. Across multiple latent dimensionalities, we identified optimal features to stratify sample sex, MYCN amplification, blood cell types, cancer types, and mutation status. These features generalized to other data, providing additional evidence for the intrinsic qualities of biological representations embedded in gene expression data [23,32–34]. Furthermore, the complexity of biological features was associated with the number of latent dimensionalities used. We predicted gene mutation using models with high dimensionality, but we detected cancer type with high accuracy using models with low dimensionality. In general, unsupervised learning algorithms applied to gene expression data extract biological and technical signals present in input samples. When applying these algorithms, researchers must determine how many latent dimensions to compress their input data into and different studies can have a variety of goals. For example, compression algorithms used for visualization can stratify sample groups based on the largest sources of variation [35–40]. In visualization settings, selecting a small number of latent dimensions is often best, and there is no need to compress data across multiple latent

422 dimensionalities. However, if the analysis goal includes learning biological signatures to identify
423 more subtle patterns in input samples, then there is not a single optimal latent dimensionality
424 nor optimal algorithm. For example, though ICA and PCA represent rotations of each other, we
425 found that the methods varied in their ability to capture specific biological signals into single
426 features, which highlights the challenge of picking only a single algorithm. While compressing
427 data into a single latent dimensionality will capture many biological signals, the "correct"
428 dimensionality is not always clear, and several biological signatures may be better revealed by
429 alternative latent dimensionalities.

430 If optimizing a single model, a researcher can use one or many criteria to select an
431 appropriate latent dimensionality. Measurements such as Akaike information criterion (AIC),
432 Bayesian information criterion (BIC), stability, and cross validation (CV) can be applied to a
433 series of latent dimensionalities [41,42]. Other algorithms, like Dirichlet processes, can naturally
434 arrive at an appropriate dimensionality through several algorithm iterations [43]. Hidden layer
435 dimensions of unsupervised neural networks are tunable hyperparameters defined by expected
436 input data complexity and performance. However, applied to gene expression data, these
437 metrics often provide conflicting results and unclear suggestions. In genomics applications, the
438 method Thresher uses a combination of outlier detection and PCA to identify the optimal
439 number of clusters [44]. Compression model stability can also be used to determine an optimal
440 latent dimensionality in gene expression data [45]. By considering only reproducible features,
441 ICA revealed 139 modules from nearly 100,000 publicly available gene expression profiles [46].
442 However, rather than using heuristics to select a biologically-appropriate latent dimensionality,

443 a researcher may instead elect to compress gene expression data into many different latent
444 space dimensionalities to generate many different feature representations.

445 There are many limitations in our evaluation of compressing gene expression data into
446 many latent dimensionalities across multiple methods. First, our approach takes a long time to
447 run. We are training many different algorithms across many different latent dimensionalities
448 and iterations, which requires a lot of compute time (**Additional File 1: Figure S14**). However,
449 because we are training many models independently, this task can be parallelized. Additionally,
450 we did not evaluate dimensionalities above k = 200. It is likely that many more signatures can
451 be learned, and possibly with even higher association strengths in higher dimensionalities for
452 certain biology. We also did not focus on detecting compressed features that represent
453 technical artifacts, which has already been widely explored [47,48]. Moreover, we did not
454 explore adding hidden layers in AE models. Many models trained on gene expression data have
455 benefited from using multiple hidden layers in neural network architectures [7,49]. Additional
456 methods, like DeepLift, can be used to reveal gene importance values in internal
457 representations of deep networks [50,51].

458 An additional challenge is interpreting the biological content of the compressed gene
459 expression features. Overrepresentation analysis (ORA) and gene set enrichment analysis
460 (GSEA) are commonly applied but have significant limitations [26,52]. ORA requires a user to
461 select a cutoff, typically based on standard deviation, to build representative gene sets from
462 each feature. ORA tests also do not consider the weights, or gene importance scores, in each
463 compression feature. Conversely, GSEA operates on ranked features, but often requires many
464 permutations to establish significance. Furthermore, ORA requires each tail of the compressed

465 feature distribution to be interpreted separately in algorithms that also learn negative weights.
466 The weight distribution is dependent on the specific compression algorithm, and the same
467 cutoff may not be appropriate for all algorithms and all compressed features. Instead, we
468 implemented a network projection based approach to quickly interpret compressed latent gene
469 expression features [53,54]. The approach is applied to the full and continuous distribution of
470 gene weights, operates independently of the algorithm feature distribution, does not require
471 arbitrary thresholds, and obviates the need to consider both tails of the distribution separately.
472 Nevertheless, additional downstream experimental validation is necessary to determine if the
473 constructed feature actually represents the biology it has been assigned.
474
475 **Conclusions:**
476 To enhance biological signature discovery, it is best to compress gene expression data
477 using several algorithms and many different latent space dimensionalities. These signatures
478 represent important biological signals, including various cell types, phenotypes, biomarkers,
479 and other sample characteristics. We showed, through several experiments tracking lower
480 dimensional gene expression representations, gene set coverage, and supervised learning
481 performance, that optimal biological features are learned using a variety of latent space
482 dimensionalities and different compression algorithms. As unsupervised machine learning
483 continues to be applied to derive insight from biomedical datasets, researchers should shift
484 focus away from optimizing a single model based on certain mathematical heuristics, and
485 instead towards learning good and reproducible biological representations that generalize to
486 alternative datasets regardless of compression algorithm and latent dimensionality.

487 **Methods:**

488 *Transcriptomic compendia acquisition and processing*

489 We downloaded transcriptomic datasets from publicly available resources. We
490 downloaded the batch-corrected TCGA PanCanAtlas RNAseq data from the National Cancer
491 Institute Genomic Data Commons (https://gdc.cancer.gov/about-
492 data/publications/pancanatlas). These data consisted of 11,069 samples with 20,531 measured
493 genes quantified with RSEM and normalized with log transformation. We converted Hugo
494 Symbol gene identifiers into Entrez gene identifiers and discarded non-protein coding genes
495 and genes that failed to map. We also removed tumors that were measured from multiple sites.
496 This resulted in a final TCGA PanCanAtlas gene expression matrix with 11,060 samples, which
497 included 33 different cancer-types, and 16,148 genes. The breakdown of TCGA samples by
498 cancer-type is provided in **Additional File 5**.

499 We downloaded the TPM normalized GTEx RNAseq data (version 7) from the GTEx data
500 portal (https://gtexportal.org/home/datasets). There were 11,688 samples and 56,202 genes in
501 this dataset. After selecting only protein-coding genes and converting Hugo Symbols to Entrez
502 gene identifiers, we considered 18,356 genes. There are 53 different detailed tissue-types in
503 this GTEx version. The tissues types included in these data are provided in **Additional File 5**.

504 Lastly, we retrieved the TARGET RNAseq gene expression data from the UCSC Xena data
505 portal [55]. The TARGET data was processed through the FPKM UCSC Toil RNA-seq pipeline and
506 was normalized with RSEM and log transformed [56]. The original matrix consists of 734
507 samples and 60,498 Ensembl gene identifiers. We converted the Ensembl gene identifiers to
508 Entrez gene names and retained only protein-coding genes. This procedure resulted in a total of

509 18,753 genes measured in TARGET. There are 7 cancer-types profiled in TARGET and the
510 specific breakdown is available in **Additional File 5**. All specific downloading and processing
511 steps can be viewed and reproduced at
512 https://github.com/greenelab/BioBombe/tree/master/0.expression-download.
513
514 *Training unsupervised neural networks*
515 Autoencoders (AE) are unsupervised neural networks that learn through minimizing the
516 reconstruction of input data after passing the data through one or several intermediate layers
517 [57]. Typically, these layers are of a lower dimensionality than the input, so the algorithms must
518 compress the input data. Denoising autoencoders (DAE) add noise to input layers during
519 training to regularize solutions and improve generalizability [58]. Variational autoencoders
520 (VAE) add regularization through an additional penalty term imposed on the objective function
521 [59,60]. In a VAE, the latent space dimensions (*k*) are penalized with a Kullback-Leibler (KL)
522 divergence penalty restricting the distribution of samples in the latent space to Gaussian
523 distributions. We independently optimized each AE model across a grid of hyperparameter
524 combinations including 6 representative latent dimensionalities (described in **Additional File 2**
525 and **Additional File 1: Figure S2**).
526
527 *Training compression algorithms across latent dimensionalities*
528 Independently for each dataset (TCGA, GTEx, and TARGET), we performed the following
529 procedure to train the compression algorithms. First, we randomly split data into 90% training
530 and 10% testing partitions. We balanced each partition by cancer type or tissue type, which

531 meant that each split contained relatively equal representation of tissues. Before input into the
532 compression algorithms, we transformed the gene expression values by gene to the [0, 1] range
533 by subtracting the minimum value and dividing by the range for each specific gene. We applied
534 this transform independently for the testing and training partitions. We selected this range
535 because it was compatible with all of the algorithms. We used the training set to train each
536 compression algorithm. We used the scikit-learn implementations of PCA, ICA, and NMF, and
537 the Tybalt implementations of VAE and DAE [8,61].

538 After learning optimized compression models with the training data, we transformed
539 the testing data using these models. We assessed performance metrics using both training and
540 testing data to reduce bias. In addition to training with real data, we also trained all models
541 with randomly permuted data. To permute the training data, we randomly shuffled the gene
542 expression values for all genes independently. We also transformed testing partition data with
543 models trained using randomly permuted data. Training with permuted data removes the
544 correlational structure in the data and can help set performance metric baselines.

545 One of our goals was to assess differences in performance and biological signal
546 detection across a range of latent dimensionalities (*k*). To this end, we trained all algorithms
547 with various *k* dimensionalities including *k* = 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 30,
548 35, 40, 45, 50, 60, 70, 80, 90, 100, 125, 150, and 200 for a total of 28 different dimensionalities.
549 All of these models were trained independently. Lastly, for each *k* dimensionality we trained
550 five different models initialized with five different random seeds. In total, considering the three
551 datasets, five algorithms, randomly permuted training data, all 28 *k* dimensionalities, and five

552 initializations, we trained 4,200 different compression models (**Additional File 2: Figure S1**).
553 Therefore, in total, we generated 185,100 different compression features.
554
555 *Evaluating compression algorithm performance*
556 We evaluated all compression algorithms on three main tasks: Reconstruction, sample
557 correlation, and weight matrix stability. First, we evaluated how well the input data is
558 reconstructed after passing through the bottleneck layer. Because the input data was
559 transformed to a distribution between 0 and 1, we used binary cross entropy to measure the
560 difference between algorithm input and output as a measure of reconstruction cost. The lower
561 the reconstruction cost, the higher fidelity reconstruction, and therefore the higher proportion
562 of signals captured in the latent space features. We also assessed the Pearson correlation of all
563 samples comparing input to reconstructed output. This value is similar to reconstruction and
564 can be quickly tracked at an individual sample level. Lastly, we used singular vector canonical
565 correlation analysis (SVCCA) to determine model stability within and model similarity between
566 algorithms and across latent dimensionalities [21]. The SVCCA method consisted of two distinct
567 steps. First, singular value decomposition (SVD) was performed on two input weight matrices.
568 The singular values that combined to reconstruct 98% of the signal in the data were retained.
569 Next, the SVD transformed weight matrix was input into a canonical correlation analysis (CCA).
570 CCA aligned different features in the weight matrix based on maximal correlation after learning
571 a series of linear transformations. Taken together, SVCCA outputs a single metric comparing
572 two input weight matrices that represents stability across model initializations and average
573 similarity of two different models. Because we used the weight matrices, the similarity

574 describes signature discovery. We use the distribution of SVCCA similarity measures across all
575 pairwise algorithm initializations and latent dimensionalities to indicate model stability [21].
576
577 *Assessing gene expression signatures present in BioBombe features*
578 We tested BioBombe sequentially compressed features to distinguish sample sex in
579 GTEx and TCGA data, and MYCN amplification in TARGET NBL data. We tested all compression
580 algorithms and latent space dimensionalities to determine the conditions in which these
581 features were best captured. First, we selected tissue-types and cancer-types in the GTEx and
582 TCGA sex analyses that were balanced by sex by selecting tissues with male to female ratios
583 between 0.5 and 1.5. We performed a two-tailed independent t-test assuming unequal
584 variance comparing male and female samples, and NBL samples with and without MYCN
585 amplification. We applied the t-test to all compression features identified across algorithms,
586 initializations, and dimensionalities. Shown in the figures are the top scoring feature per latent
587 space dimensionality and algorithm.
588 We applied the optimal MYCN signature learned in TARGET to an alternative dataset
589 consisting of a series of publicly available NBL cell lines [25]. The data were processed using
590 STAR, and we accessed the processed FPKM matrix from figshare [62]. We transformed the
591 dataset with the identified signatures using the following operation:
592 $$S_{g'}^{T} * D_{g' \text{ x } n} = D'_{s\ x\ n}$$
593 Where *D* represents the respective RNAseq data to transform, *S* represents the specific
594 signature, *g'* represents the overlapping genes measured in both datasets, *n* represents

595 samples, and D'_s represents the signature scores in the transformed dataset. Of the 8,000 genes
596 measured in TARGET data, 7,653 were also measured in external NBL cell line dataset (95.6%).
597 Using the sample activation scores for each of the top scoring features for sample sex in
598 TCGA and GTEx, and MYCN amplification in TARGET and the validation set, we performed two
599 tailed t-test with unequal variance comparing each group. For the TCGA and GTEx sex
600 comparison, our t-test compared male vs. female activation scores. For the TARGET and NBL
601 cell line analyses, our t-test compared MYCN amplified NBL samples vs. MYCN non-amplified
602 NBL samples. We add t-test statistics and p values in each sub figure.
603
604 *Gene network construction and processing*
605 We constructed networks using gene set collections compiled by version 6.2 of the
606 Molecular Signatures Database (MSigDB) and cell types derived from xCell [26–28]. These gene
607 sets represent a series of genes that are involved in specific biological processes and functions.
608 We integrated all openly licensed MSigDB collections which included hallmark gene sets (H),
609 positional gene sets (C1), curated gene sets (C2), motif gene sets (C3), computational gene sets
610 (C4), Gene Ontology (GO) terms (C5), oncogenic gene sets (C6) and immunologic gene sets (C7).
611 We omitted MSigDB gene sets that were not available under an open license (KEGG, BioCarta,
612 and AAAS/STKE). The C2 gene set database was split into chemical and genetic perturbations
613 (C2.CPG) and Reactome (C2.CP.Reactome). The C3 gene set was split into microRNA targets
614 (C3.MIR) and transcription factor targets (C3.TFT). The C4 gene set was split into cancer gene
615 neighborhoods (C4.CGN) and cancer modules (C4.CM). Lastly, the C5 gene set was split into GO
616 Biological Processes (C5.BP), GO Cellular Components (C5.CC), and GO molecular functions

617 (C5.MF). xCell represents a gene set compendia of 489 computationally derived gene signatures
618 from 64 different human cell types. The number of gene sets in each curation is provided in
619 **Additional File 6**. In BioBombe network projection, only a single collection is projected at a
620 time.

621 To build the gene set network, we used hetnetpy [63]. Briefly, hetnetpy builds networks
622 that include multiple node types and edge relationships. We used hetnetpy to build a single
623 network containing all MSigDB collections and xCell gene sets listed above. The network
624 consisted of 17,451 unique gene sets and 2,159,021 edges representing gene set membership
625 among 20,703 unique gene nodes (**Additional File 6**). In addition to generating a single network
626 using curated gene sets, we also used hetnetpy to generate 10 permuted networks. The
627 networks are permuted using the XSwap algorithm, which randomizes connections while
628 preserving node degree (i.e. the number of gene set relationships per gene) [64]. Therefore, the
629 permuted networks are used to control for biases induced by uneven gene degree. We
630 compared the observed score against the distribution of permuted network scores to interpret
631 the biological signatures in each compression feature.

632

633 *Rapid interpretation of compressed gene expression data*

634 Our goal was to quickly interpret the automatically generated compressed latent
635 features learned by each unsupervised algorithm. To this end, we constructed gene set
636 adjacency matrices with specific MSigDB or xCell gene set collections using hetnetpy software.
637 We then performed the following matrix multiplication against a given compressed weight
638 matrix to obtain a raw score for all gene sets for each latent feature.

639 $$H_{c \text{ x } n} * W_{n \text{ x } k} = G_{c \text{ x } k}$$

640 Where *H* represents the gene set adjacency matrix, *c* is the specific gene set collection, and *n*
641 represents genes. *W* represents the specific compression algorithm weight matrix, which
642 includes *n* genes and *k* latent space features. The output of this matrix multiplication, *G*, is
643 represented by *c* gene sets and *k* latent dimensions. Through a single matrix multiplication, the
644 matrix *G* tracks raw BioBombe scores.

645 Because certain hub genes are more likely to be implicated in gene sets and longer gene
646 sets will receive higher raw scores, we compared *G* to the distribution of permuted scores
647 against all 10 permuted networks.

648 $$H_{p_{c \text{ x } n}}^{1-10} * W_{n \text{ x } k} = G_p$$

649 $$G_{z-score} = \frac{G_{c \text{ x } k} - \overline{G_p}}{\sigma(G_p)}$$

650 Where $H_P^{1\text{-}10}$ represents the adjacency matrices for all 10 permuted networks and G_p represents
651 the distribution of scores for the same *k* features for all permutations. We calculated the z
652 score for all gene sets by latent features ($G_{z\text{-}score}$). This score represents the BioBombe Score.
653 Other network-based gene set methods consider gene set influence based on network
654 connectivity of gene set genes [53,54]. Instead, we used the latent feature weights derived
655 from unsupervised compression algorithms as input, and the compiled gene set networks to
656 assign biological function.

657 We also compared the BioBombe network projection approach to overrepresentation
658 analyses (ORA). We did not compare the approach to gene set enrichment analysis (GSEA)
659 because evaluating single latent features required many permutations and did not scale to the

660 many thousands of compressed features we examined. We implemented ORA analysis using a
661 Fisher's Exact test. The background genes used in the test included only the genes represented
662 in the specific gene set collection.
663
664 *Calculating gene set coverage across BioBombe features*
665 We were interested in determining the proportion of gene sets within gene set
666 collections that were captured by the features derived from various compression algorithms.
667 We considered a gene set "captured" by a compression feature if it had the highest positive or
668 highest negative BioBombe z score compared to all other gene sets in that collection. We
669 converted BioBombe z scores into p values using the pnorm() R function using a two-tailed test.
670 We removed gene sets from consideration if their p values were not lower than a Bonferroni
671 adjusted value determined by the total number of latent dimensionalities in the model.
672 We calculated coverage (C) by considering all unique top gene sets (U) identified by all
673 features in the compression model (w) and dividing by the total number of gene sets in the
674 collection (T_C).
675 $$C = \frac{U_w}{T_c}$$
676 We calculated the coverage metric for all models independently (C_i), for ensembles, or
677 individual algorithms across all five iterations (C_e), and for all models across k dimensions (C_k).
678 We also calculated the total coverage of all BioBombe features combined in a single model
679 (C_{all}). A larger coverage value indicated a model that captured a larger proportion of the
680 signatures present in the given gene set collection.
681

682 *Downloading and processing publicly available expression data for neutrophil GTEx analysis*

683 We used an external dataset to validate the neutrophil feature learned by compressing
684 GTEx gene expression data into three latent dimensionalities. We observed that this feature
685 contributed to improved reconstruction of blood tissue. To assess the performance of this
686 neutrophil signature, we downloaded data from the Gene Expression Omnibus (GEO) with
687 accession number GSE103706 [29]. RNA was captured in this dataset using Illumina NextSeq
688 500. The dataset measured the gene expression of several replicates of two neutrophil-like cell
689 lines, HL-60 and PLB-985, which were originally derived from acute myeloid leukemia (AML)
690 patients. The PLB-985 cell line was previously identified as a subclone of HL-60, so we expect
691 similar signature activity between the two lines [65]. Gene expression of the two cell lines was
692 measured with and without neutrophil differentiation treatments. Though DMSO is frequently
693 used to solubilize compounds and act as an experimental control, it has been used to create
694 neutrophil-like cells [66]. The validation dataset we used was generated to compare DMSO
695 activity with untreated cells and cells treated with DMSO plus Nutridoma [29]. We tested the
696 hypothesis that our neutrophil signature would distinguish the samples with and without
697 neutrophil differentiation treatment. We transformed external datasets with the following
698 operation:

699 $$W^T_{k \text{ x } g'} * D_{g' \text{ x } n} = D'_{k \text{ x } n}$$

700 Where *D* represents the processed RNAseq data from GSE103706. Of 8,000 genes measured in
701 *W*, 7,664 were also measured in *D* (95.8%). These 7,664 genes are represented by *g'*. All of the
702 "Neutrophils_HPCA_2" signature genes were measured in *W*. *D'* represents the GSE103706
703 data transformed along the specific compression feature. Each sample in *D'* is then considered

704 transformed by the specific signature captured in *k*. The specific genes representing
705 "Neutrophils_HPCA_2" is provided in **Additional File 3**.
706
707 *Downloading and processing publicly available expression data for monocyte GTEx analysis*
708 We used an additional external dataset to validate the identified monocyte signature.
709 We accessed processed data for the publicly available GEO dataset with accession number
710 GSE24759 [30]. The dataset was measured by Affymetrix HG-U133A (early access array) and
711 consisted of 211 samples representing 38 distinct and purified populations of cells, including
712 monocytes, undergoing various stages of hematopoiesis. The samples were purified from 4 to 7
713 independent donors each. Many xCell gene sets were computationally derived from this
714 dataset as well [27]. Not all genes in the weight matrices were measured in the GSE24759
715 dataset. For this application, 4,645 genes (58.06%) corresponded with the genes used in the
716 compression algorithms. Additionally, 168 out of 178 genes (94.38%) in the
717 "Monocyte_FANTOM_2" gene set were measured (**Additional File 3**). We investigated the
718 "Monocytes_FANTOM_2" signature because of its high enrichment in VAE $k = 3$ and low
719 enrichment in VAE $k = 2$.
720
721 *Machine learning classification of cancer types and gene alterations in TCGA*
722 We trained supervised learning classifiers using raw RNA-seq features and BioBombe-
723 derived features. In general, we trained supervised machine learning models to predict cancer
724 type from RNAseq features in TCGA PanCanAtlas RNAseq data. We implemented a logistic
725 regression classifier with an elastic net penalty. The classifiers were controlled for mutation

726 burden. More details about the specific implementation are described in Way et al. 2018 [67].
727 Here, we predicted all 33 cancer types using all 11,060 samples. These predictions were
728 independent per cancer type, which meant that we trained models with the same input gene
729 expression or BioBombe feature data, but used 33 different status matrices.

730 We also trained models to predict gene alteration status in the top 50 most mutated
731 genes in the PanCanAtlas. These models were controlled for cancer type and mutation burden.
732 We defined the status in this task using all non-silent mutations identified with a consensus
733 mutation caller [68]. We also considered large copy number amplifications for oncogenes and
734 deep copy number deletions for tumor suppressor genes as previously defined [69]. We used
735 the threshold GISTIC2.0 calls for large copy amplifications (score = 2) and deep copy deletions
736 (score = -2) in defining the status matrix [70]. For each gene alteration prediction, we removed
737 samples with a hypermutator phenotype, defined by having log10 mutation counts greater than
738 five standard deviations above the mean. For the mutation prediction task, we also did not
739 include certain cancer types in training. We omitted cancer types if they had less than 5% or
740 more than 95% representation of samples with the given gene alteration. The positive and
741 negative sets must have also included at least 15 samples. We filtered out cancer types in this
742 manner to prevent the classifiers from artificially detecting differences induced by unbalanced
743 training sets.

744 We trained models with raw RNAseq data subset by the top 8,000 most variably
745 expressed genes by median absolute deviation. The training data used was the same training
746 set used for the BioBombe procedure. We also trained models using all BioBombe compression
747 matrices for each latent dimension, and using real and permuted data. We combined

compressed features together to form three different types of ensemble models. The first type grouped all five iterations of VAE models per latent dimensionality to make predictions. The second type grouped features of five different algorithms (PCA, ICA, NMF, DAE, VAE) of a single iteration together to make predictions. The third ensemble aggregated all features learned by all algorithms, all initializations, and across all latent dimensionalities, which included a total of 30,850 features. In total, considering the 33 cancer types, 50 mutations, 28 latent dimensionalities, ensemble models, raw RNAseq features, real and permuted data, and 5 initializations per compression, we trained and evaluated 32,868 different supervised models.

We optimized all of the models independently using 5-fold cross validation (CV). We searched over a grid of elastic net mixing and alpha hyperparameters. The elastic net mixing parameter represents the tradeoff between l1 and l2 penalties (where mixing = 0 represents an l2 penalty) and controls the sparsity of solutions [71]. Alpha is a penalty that tunes the impact of regularization, with higher values inducing higher penalties on gene coefficients. We searched over a grid for both hyperparameters (alpha = 0.1, 0.13, 0.15, 0.2, 0.25, 0.3 and mixing = 0.15, 0.16, 0.2, 0.25, 0.3, 0.4) and selected the combination with the highest CV AUROC. For each model, we tested performance using the original held out testing set that was also used to assess compression model performance.

Evaluating model training time

We evaluated the execution time of training each compression algorithm for all three datasets across several latent dimensionalities. We used 8 representative latent

769 dimensionalities: k = 2, 4, 10, 16, 25, 50, 80, and 200. We conducted the time analysis using a
770 CPU machine with an Intel Core i3 dual core processer with 32 GB of DDR4 memory.
771
772 *Reproducible software*
773 All code to perform all analyses and generate all results and figures is provided with an open
774 source license at https://github.com/greenelab/biobombe [72]. All resources can be viewed
775 and downloaded from https://greenelab.github.io/BioBombe/.
776
777 **List of abbreviations:**
778 RNAseq = RNA sequencing; PCA = principal components analysis; ICA = independent
779 components analysis; NMF = non-negative matrix factorization; AE = autoencoder; DAE =
780 denoising autoencoder; VAE = variational autoencoder; TCGA = the cancer genome atlas; GTEx
781 = genome tissue expression project; TARGET = therapeutically applicable research to generate
782 effective treatments project; BRCA = breast invasive carcinoma; COAD = colon
783 adenocarcinoma; LGG = low grade glioma; PCPG = pheochromocytoma and paraganglioma;
784 LAML = acute myeloid leukemia; LUAD = lung adenocarcinoma; GEO = gene expression
785 omnibus; ROC = receiver operating characteristic; PR = precision recall; AUROC = area under the
786 receiver operating characteristic curve; AUPR = area under the precision recall curve; CV = cross
787 validation; ORA = overrepresentation analysis; GSEA = gene set enrichment analysis; SVD =
788 singular value decomposition; CCA = canonical correlation analysis; SVCCA = singular vector
789 canonical correlation analysis; TF = transcription factor; DMSO = dimethyl sulfoxide
790

791 **Declarations:**

792 *Ethics approval and consent to participate*

793 The TCGA, GTEx, and TARGET data used are publicly available and their use was
794 previously approved by their respective ethics committees.

795 *Consent for publication*

796 Not applicable.

797 *Availability of data and material*

798 All data used and results generated in this manuscript are publicly available. All results
799 of the analysis can be viewed and downloaded from
800 https://greenelab.github.io/BioBombe/. The analyzed data can be accessed in the
801 following locations: TCGA data can be accessed at https://gdc.cancer.gov/about-
802 data/publications/pancanatlas, the GTEx data can be accessed at
803 https://gtexportal.org/home/datasets, the TARGET data can be accessed at
804 https://toil.xenahubs.net/download/target_RSEM_gene_fpkm.gz, the neutrophil
805 validation data can be accessed using gene expression omnibus (GEO) accession number
806 GSE103706 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103706), the
807 monocyte validation data can be accessed using GEO accession number GSE24759
808 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24759). Software to
809 reproduce the analyses, and all results generated in this manuscript can be accessed at
810 https://github.com/greenelab/biobombe. The software has also been archived in an
811 additional publicly available repository at https://zenodo.org/record/3460539.
812 Additionally, all BioBombe results are available in a series of versioned archives (TCGA

BioBombe Results: https://zenodo.org/record/2110752); (GTEX BioBombe Results: https://zenodo.org/record/2300616); (TARGET BioBombe Results: https://zenodo.org/record/2222463); (Randomly permuted TCGA BioBombe Results: https://zenodo.org/record/2221216); (Randomly permuted GTEx BioBombe Results: https://zenodo.org/record/2386816); (Randomly Permuted TARGET BioBombe Results: https://zenodo.org/record/2222469). The full results of the TCGA classification analysis using BioBombe features is archived at https://zenodo.org/record/2535759.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded in party by The Gordon and Betty Moore Foundation under GBMF 4552 (CSG) and the National Institutes of Health's National Human Genome Research Institute under R01 HG010067 (CSG) and the National Institutes of Health under T32 HG000046 (GPW).

Authors' contributions

GPW performed the analysis, wrote the BioBombe software, generated the figures, and wrote the manuscript. GPW and CSG designed the study and interpreted the results. MZ and DSH developed the network software. VR developed the website. All authors read, revised, and approved the final manuscript.

Acknowledgements

We would like to thank Jaclyn Taroni, Yoson Park, and Alexandra Lee for insightful discussions and code review. We also thank Jo Lynne Rokita and John Maris for insightful discussions regarding the neuroblastoma analysis.

References:

1. Fehrmann RSN, Karjalainen JM, Krajewska M, Westra H-J, Maloney D, Simeonov A, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat Genet. 2015;47:115–25.

2. Engreitz JM, Daigle BJ, Marshall JJ, Altman RB. Independent component analysis: Mining microarray data for fundamental human gene expression modules. J Biomed Inform. 2010;43:932–44.

3. Kong W, Vanderburg CR, Gunshin H, Rogers JT, Huang X. A review of independent component analysis application to microarray gene expression data. BioTechniques. 2008;45:501–20.

4. Gaujoux R, Seoighe C. CellMix: a comprehensive toolbox for gene expression deconvolution. Bioinforma Oxf Engl. 2013;29:2211–2.

5. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol. 2013;25:571–8.

6. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, et al. Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks. Cell Syst. 2017;5:63–71.e6.

7. Chen L, Cai C, Chen V, Lu X. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. BMC Bioinformatics. 2016;17:S9.

8. Way GP, Greene CS. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pac Symp Biocomput Pac Symp Biocomput. 2018;23:80–91.

9. Rampasek L, Hidru D, Smirnov P, Haibe-Kains B, Goldenberg A. Dr.VAE: Drug Response Variational Autoencoder. ArXiv170608203 Stat [Internet]. 2017; Available from: http://arxiv.org/abs/1706.08203

10. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.

863 11. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression
864 residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat
865 Protoc. 2012;7:500–7.

866 12. Weinstein JN, Collisson EA, Mills GB, Shaw KM, Ozenberger BA, Ellrott K, et al. The Cancer
867 Genome Atlas Pan-Cancer Analysis Project. Nat Genet. 2013;45:1113–20.

868 13. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–
869 5.

870 14. Mullighan CG, Su X, Zhang J, Radtke I, Phillips LAA, Miller CB, et al. Deletion of IKZF1 and
871 prognosis in acute lymphoblastic leukemia. N Engl J Med. 2009;360:470–80.

872 15. Way G. TCGA BioBombe Results [Internet]. Zenodo; 2018 [cited 2019 Jan 20]. Available
873 from: https://zenodo.org/record/2110752

874 16. Way G. GTEX BioBombe Results [Internet]. Zenodo; 2018 [cited 2019 Jan 20]. Available
875 from: https://zenodo.org/record/2300616

876 17. Way G. TARGET BioBombe Results [Internet]. Zenodo; 2018 [cited 2019 Jan 20]. Available
877 from: https://zenodo.org/record/2222463

878 18. Way G. TCGA BioBombe Results - Randomly Permuted Data [Internet]. Zenodo; 2018 [cited
879 2019 Jan 20]. Available from: https://zenodo.org/record/2221216

880 19. Way G. GTEX BioBombe Results - Randomly Permuted Data [Internet]. Zenodo; 2018 [cited
881 2019 Jan 20]. Available from: https://zenodo.org/record/2386816

882 20. Way G. TARGET BioBombe Results - Randomly Permuted Data [Internet]. Zenodo; 2018
883 [cited 2019 Jan 20]. Available from: https://zenodo.org/record/2222469

884 21. Raghu M, Gilmer J, Yosinski J, Sohl-Dickstein J. SVCCA: Singular Vector Canonical Correlation
885 Analysis for Deep Learning Dynamics and Interpretability. Neural Inf Process Syst NeurIPS. 2017;

886 22. Clark B, Stein-O'Brien G, Shiau F, Cannon G, Davis E, Sherman T, et al. Comprehensive
887 analysis of retinal development at single cell resolution identifies NFI factors as essential for
888 mitotic exit and specification of late-born cells. bioRxiv [Internet]. 2018 [cited 2019 Feb
889 17];https://doi.org/10.1101/378950. Available from:
890 http://biorxiv.org/lookup/doi/10.1101/378950

891 23. Stein-O'Brien GL, Clark BS, Sherman T, Zibetti C, Hu Q, Sealfon R, et al. Decomposing Cell
892 Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species.
893 Cell Syst. 2019;8:395–411.e8.

894 24. Huang M, Weiss WA. Neuroblastoma and MYCN. Cold Spring Harb Perspect Med.
895 2013;3:a014415–a014415.

896 25. Harenza JL, Diamond MA, Adams RN, Song MM, Davidson HL, Hart LS, et al. Transcriptomic
897 profiling of 39 commonly-used neuroblastoma cell lines. Sci Data. 2017;4:170033.

898 26. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set
899 enrichment analysis: A knowledge-based approach for interpreting genome-wide expression
900 profiles. Proc Natl Acad Sci. 2005;102:15545–50.

901 27. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape.
902 Genome Biol [Internet]. 2017 [cited 2019 Jan 15];18. Available from:
903 https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1349-1

904 28. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular
905 Signatures Database Hallmark Gene Set Collection. Cell Syst. 2015;1:417–25.

906 29. Rincón E, Rocha-Gregg BL, Collins SR. A map of gene expression in neutrophil-like cell lines.
907 BMC Genomics. 2018;19:573.

908 30. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME, et al.
909 Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell.
910 2011;144:296–309.

911 31. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of
912 somatic mutation in human cancer genomes. Nature. 2007;446:153–8.

913 32. Dincer AB, Celik S, Hiranuma N, Lee S-I. DeepProfile: Deep learning of cancer molecular
914 profiles for precision medicine. bioRxiv [Internet]. 2018 [cited 2019 Sep 17]; Available from:
915 http://biorxiv.org/lookup/doi/10.1101/278739

916 33. Taroni JN, Grayson PC, Hu Q, Eddy S, Kretzler M, Merkel PA, et al. MultiPLIER: a transfer
917 learning framework reveals systemic features of rare autoimmune disease. 2018 [cited 2018
918 Sep 9]; Available from: http://biorxiv.org/lookup/doi/10.1101/395947

919 34. Wang J, Agarwal D, Huang M, Hu G, Zhou Z, Ye C, et al. Data denoising with transfer learning
920 in single-cell transcriptomics. Nat Methods. 2019;16:875–8.

921 35. Shi J, Luo Z. Nonlinear dimensionality reduction of gene expression data for visualization
922 and clustering analysis of cancer tissue samples. Comput Biol Med. 2010;40:723–32.

923 36. Bartenhagen C, Klein H-U, Ruckert C, Jiang X, Dugas M. Comparative study of unsupervised
924 dimension reduction techniques for the visualization of microarray gene expression data. BMC
925 Bioinformatics [Internet]. 2010 [cited 2019 Jan 26];11. Available from:
926 https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-567

927 37. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction
928 for visualizing single-cell data using UMAP. Nat Biotechnol. 2018;37:38–44.

929 38. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. bioRxiv [Internet].
930 2018 [cited 2019 Jan 26];http://biorxiv.org/lookup/doi/10.1101/453449. Available from:
931 http://biorxiv.org/lookup/doi/10.1101/453449

932 39. Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9:2579–
933 605.

934 40. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for
935 Dimension Reduction. arXiv:180203426 [Internet]. 2018; Available from:
936 https://arxiv.org/abs/1802.03426

937 41. Ben-Hur A, Elisseeff A, Guyon I. A stability based method for discovering structure in
938 clustered data. Pac Symp Biocomput Pac Symp Biocomput. 2002;6–17.

939 42. Wang J. Consistent selection of the number of clusters via crossvalidation. Biometrika.
940 2010;97:893–904.

941 43. Wang L, Wang X. Hierarchical Dirichlet process model for gene expression clustering.
942 EURASIP J Bioinforma Syst Biol. 2013;2013:5.

943 44. Wang M, Abrams ZB, Kornblau SM, Coombes KR. Thresher: determining the number of
944 clusters while removing outliers. BMC Bioinformatics. 2018;19:9.

945 45. Wu S, Joseph A, Hammonds AS, Celniker SE, Yu B, Frise E. Stability-driven nonnegative
946 matrix factorization to interpret spatial gene expression and build local gene networks. Proc
947 Natl Acad Sci. 2016;113:4290–5.

948 46. Zhou W, Altman RB. Data-driven human transcriptomic modules determined by
949 independent component analysis. BMC Bioinformatics [Internet]. 2018 [cited 2018 Dec 22];19.
950 Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-
951 2338-4

952 47. Leek JT, Storey JD. Capturing Heterogeneity in Gene Expression Studies by Surrogate
953 Variable Analysis. PLoS Genet. 2007;3:e161.

954 48. Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to
955 deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics.
956 2011;27:1496–505.

957 49. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell
958 transcriptomics. Nat Methods. 2018;15:1053–8.

959 50. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating
960 Activation Differences. ArXiv170402685 Cs [Internet]. 2017; Available from:
961 http://arxiv.org/abs/1704.02685

962 51. Lin C, Jain S, Kim H, Bar-Joseph Z. Using neural networks for reducing the dimensions of
963 single-cell RNA-Seq data. Nucleic Acids Res. 2017;45:e156–e156.

964 52. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive,
965 powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res.
966 2017;45:W130–7.

967 53. Fang Z, Tian W, Ji H. A network-based gene-weighting approach for pathway analysis. Cell
968 Res. 2012;22:565–80.

969 54. Dong X, Hao Y, Wang X, Tian W. LEGO: a novel method for gene set over-representation
970 analysis by incorporating network-based gene weights. Sci Rep [Internet]. 2016 [cited 2019 Jan
971 14];6. Available from: http://www.nature.com/articles/srep18871

972 55. Goldman M, Craft B, Kamath A, Brooks AN, Zhu J, Haussler D. The UCSC Xena Platform for
973 cancer genomics data visualization and interpretation. bioRxiv [Internet]. 2018 [cited 2019 Jan
974 21]; Available from: http://biorxiv.org/lookup/doi/10.1101/326470

975 56. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil enables
976 reproducible, open source, big biomedical data analyses. Nat Biotechnol. 2017;35:314–6.

977 57. Baldi P, Hornik K. Neural networks and principal component analysis: Learning from
978 examples without local minima. Neural Netw. 1989;2:53–8.

979 58. Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and Composing Robust Features
980 with Denoising Autoencoders. Proc 25th Int Conf Mach Learn [Internet]. New York, NY, USA:
981 ACM; 2008. p. 1096–1103. Available from: http://doi.acm.org/10.1145/1390156.1390294

982 59. Kingma DP, Welling M. Auto-Encoding Variational Bayes. ArXiv13126114 Cs Stat [Internet].
983 2013 [cited 2017 Mar 6]; Available from: http://arxiv.org/abs/1312.6114

984 60. Rezende DJ, Mohamed S, Wierstra D. Stochastic Backpropagation and Approximate
985 Inference in Deep Generative Models. ArXiv14014082 Cs Stat [Internet]. 2014 [cited 2017 May
986 10]; Available from: http://arxiv.org/abs/1401.4082

987 61. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn:
988 Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.

989 62. Harenza JL. Transcriptomic profiling of 39 commonly-used neuroblastoma cell lines.
990 2019;https://figshare.com/articles/STAR-reads/7613975/3.

991 63. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic
992 integration of biomedical knowledge prioritizes drugs for repurposing. eLife. 2017;6.

993 64. Hanhijärvi S, Garriga GC, Puolamäki K. Randomization Techniques for Graphs. Proc 2009
994 SIAM Int Conf Data Min. 2009;780–91.

995 65. Drexler HG, Dirks WG, Matsuo Y, MacLeod R a. F. False leukemia-lymphoma cell lines: an
996 update on over 500 cell lines. Leukemia. 2003;17:416–26.

997 66. Jacob C, Leport M, Szilagyi C, Allen JM, Bertrand C, Lagente V. DMSO-treated HL60 cells: a
998 model of neutrophil-like cells mainly expressing PDE4B subtype. Int Immunopharmacol.
999 2002;2:1647–56.

1000 67. Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, et al. Machine Learning
1001 Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Rep.
1002 2018;23:172–180.e3.

1003 68. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, et al. Scalable Open
1004 Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell
1005 Syst. 2018;6:271–281.e7.

1006 69. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome
1007 Landscapes. Science. 2013;339:1546–58.

1008 70. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates
1009 sensitive and confident localization of the targets of focal somatic copy-number alteration in
1010 human cancers. Genome Biol. 2011;12:R41.

1011 71. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B
1012 Stat Methodol. 2005;67:301–20.

1013 72. Way G. greenelab/BioBombe: BioBombe Analysis Version 1.2 [Internet]. Zenodo; 2019
1014 [cited 2019 Mar 9]. Available from: https://zenodo.org/record/3460539

1015