

1 Cox-nnet: an artificial neural network
2 method for prognosis prediction on high-
3 throughput omics data
4

5 Travers Ching*.†, Xun Zhu*.†, Lana X. Garmire*·†

6 *Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa,
7 Honolulu, HI 96822, USA

8 †Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA

9 **Running title**

10 Artificial neural network for prognosis

11 **Keywords**

12 Survival analysis, RNA-Seq, cancer, machine learning, bioinformatics

13 **Corresponding author**

14 Lana Garmire

15 701 Ilalo Street, Honolulu HI 96813

16 (808) 441-8193

17 lgarmire@cc.hawaii.edu

18

19 **Abstract**

20 Artificial neural networks (ANN) are computing architectures with massively parallel interconnections of
21 simple neurons and has been applied to biomedical fields such as imaging analysis and diagnosis. We
22 have developed a new ANN framework called Cox-nnet to predict patient prognosis from high throughput
23 transcriptomics data. In over 10 TCGA RNA-Seq data sets, Cox-nnet achieves a statistically significant
24 increase in predictive accuracy, compared to the other three methods including Cox-proportional hazards
25 (Cox-PH), Random Forests Survival and CoxBoost. Cox-nnet also reveals richer biological information,
26 from both pathway and gene levels. The outputs from the hidden layer node can provide a new approach
27 for survival-sensitive dimension reduction. In summary, we have developed a new method for more
28 accurate and efficient prognosis prediction on high throughput data, with functional biological insights.
29 The source code is freely available at <https://github.com/lanagarmire/cox-nnet>.

30

31 **Introduction**

32 Artificial Neural networks (ANNs) were developed in 1943 in order to model the activity of neurons¹. In
33 recent years, ANNs have caught renewed attention, thanks to the increased parallel computing power and
34 the promise of deep learning². The original ANN extension of Cox Regression was not designed to
35 handle high throughput input data³. Some recent attempts using ANNs to high dimensional survival data
36 simplified the regression problem as either a binary classification or fitting discrete variables of survival
37 time through binning , leading to loss of accuracy^{4,5}. Another study used time as an additional input in
38 order to predict patient survival or censoring status⁶, with the potential to overfit if the survival time and
39 censoring time are correlated. To avoid all these issues, we herein leverage the neural network extension
40 of Cox regression by a high-performance and easy-to-use package, particularly fit for high dimensional
41 data.

42 Besides Cox-nnet, some other modeling methods exist to predict patient survival. The standard method is
43 Cox proportional hazards (Cox-PH) regression, a semi-parametric and generalized linear model with an
44 exponential link function⁷. Another method, CoxBoost⁸, is an iterative “boosting” method modified from
45 the Cox-PH model. In each boosting iteration, it refits the parameters by maximizing the penalized
46 likelihood function. Rather than using L2 ridge regression common in Cox-PH, the number of boosting
47 iterations is used as the complexity parameter in CoxBoost and optimized via cross-validation (CV)⁸.
48 Random Forests Survival (RF-S) is another ensemble, non-linear method⁹. It combines many
49 bootstrapped decision trees in order to reduce the variance in the model, and then calculates a weighted
50 average of all the decision trees. Unlike Cox-PH and CoxBoost, RF-S does not use the log likelihood
51 function to determine the fitness of the model. Instead, it predicts estimated survival times, and uses
52 Harrel’s C-Index, a score that measures the correct ranking of individuals⁹.

53 The new software package we have developed here, named Cox-nnet, advances the ANN extension of
54 Cox regression for survival prediction on high-throughput data. The caliber of this package is manifested
55 in several aspects. First, it has improved technical performance in terms of both accuracy and speed. In

56 comparison with the other methods mentioned above (Cox-PH, RF-S and CoxBoost), Cox-nnet has better
57 overall predictive accuracy. It is also optimized on graphics processing unit (GPU) with at least an order
58 of computational speed-up over the central processing unit (CPU), making it a compelling new tool to
59 predict disease prognosis in the era of precision medicine. Second, Cox-nnet utilizes feature importance
60 scores based on the partial derivatives of gene features selected by the model, so that the relative
61 importance of the genes to prognosis outcome can be directly assessed. Thirdly, the hidden layer node
62 structure in ANN can be harnessed to reveal much richer information of featuring genes and biological
63 pathways, compared to the Cox-PH method. Overall, Cox-nnet is a desirable survival analysis method
64 with both excellent predictive power and usage to gain biological functions related to prognosis.

65 **Methods**

66 **The Cox model**

67 The Cox-PH model is a log-linear model that estimates individual hazard, i.e., an instantaneous measure
68 of the likelihood of an event, based on a set of features. The hazard is given by the equation:

$$h(t|X_i) = h_0(t) \exp \theta_i \quad (1)$$

$$\theta_i = X_i^T \beta \quad (2)$$

69 Where θ_i is the log hazard ratio for patient i . The partial likelihood is represented by the following
70 formula:

$$PL(\beta) = \prod_{C(i)=1} \frac{\exp \theta_i}{\sum_{t_j \geq t_i} \exp \theta_j} \quad (3)$$

71 Where $C(i)$ is the censoring status of a patient, and $C(i)=0$ if the patient was censored or 1 if the
72 patient died or had a recurrence event, etc. The partial log-likelihood is used as the cost function:

$$pl(\boldsymbol{\beta}) = \sum_{C(i)=1} \left(\theta_i - \log \sum_{t_j \geq t_i} \exp \theta_j \right) \quad (4)$$

73 In a Cox model with L2 ridge regression, a penalty term is added which is proportional to the L2 norm of
74 the coefficients. The cost function is minimized to find the best coefficients for the model:

$$Cost(\boldsymbol{\beta}) = -pl(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|^2 \quad (5)$$

75 where the tuning parameter λ is determined by maximizing CV.

76 The cross-validated performance metric may be Harrel's concordance index (C-index)¹⁰ or the "cross-
77 validated partial likelihood"¹¹. Since the contribution of each patient in the partial likelihood is
78 determined only in the context of all the other patients, the cross-validated partial likelihood is calculated
79 subtracting full partial likelihood from the training set in the CV. In the k-th iteration of a K-fold CV, the
80 optimal coefficients $\hat{\boldsymbol{\beta}}_{\lambda,k}$ are found by minimizing the cost function on the training sub-samples. If
81 $pl_k(\hat{\boldsymbol{\beta}}_{\lambda,k})$ is the partial likelihood of the training sub-samples, and $pl(\hat{\boldsymbol{\beta}}_{\lambda,k})$ is the partial likelihood of
82 the full dataset, then the cross-validated partial likelihood is the sum of differences:

$$cvpl(\lambda) = \sum_{k=1}^K pl(\hat{\boldsymbol{\beta}}_{\lambda,k}) - pl_k(\hat{\boldsymbol{\beta}}_{\lambda,k}) \quad (6)$$

83 **ANN extension of Cox regression**

84 The ANN extension of Cox regression (Cox-nnet) is a neural network whose output layer is replaced by a
85 Cox model. In a Cox-nnet model with one input layer of J input features and one hidden layer composed
86 of H hidden nodes, the linear predictor is replaced by the outputs of the hidden layer:

$$\theta_i = G(\mathbf{W}^T \mathbf{X}_i + \mathbf{b})^T \boldsymbol{\beta} \quad (7)$$

87 Where \mathbf{W} is the coefficient weight matrix between the input and hidden layer with the size $H \times J$, \mathbf{b} is
88 the bias term for each hidden node and G is the activation function (applied element-wise on a vector).
89 Subsequently, the ridge regression cost function is modified to:

$$Cost(\boldsymbol{\beta}, \mathbf{W}) = pl(\boldsymbol{\beta}, \mathbf{W}) + \lambda (\|\boldsymbol{\beta}\|^2 + \|\mathbf{W}\|^2) \quad (8)$$

90 In this manuscript, the tanh activation function is used, as it results in faster training time compared to the
91 sigmoid activation ¹². The tanh function is:

$$G(z) = \frac{\exp(z) - \exp(-z)}{\exp(z) + \exp(-z)} \quad (9)$$

92 In addition to ridge regularization, we also employ dropout regularization¹³. In this approach, nodes are
93 removed during each training iteration with probability 1-p. During evaluation, output from the nodes are
94 multiplied by p. The optimal dropout parameter, p, is determined through cross-validation on the training
95 set. Dropout regularization has been shown to reduce overfitting and improve performance over other
96 regularization schemes¹³.

97 The source code of cox-nnet can be found at: <https://github.com/lgormire/cox-nnet>, and can be installed
98 through the Python Package Index (PyPI). Documentation of package can be found at
99 <http://lgormire.github.io/cox-nnet/docs>.

100 Implementation in Python with Theano

101 We implement Cox-nnet using a feed forward, back propagation network with gradient descent. The
102 partial log likelihood is usually written as a double conditional sum (equation 4). To avoid the
103 computational inefficiency of calculating the partial log likelihood (equation 4) using two nested for
104 loops, we convert it into a formulation of matrix operations and basic sums. First we define an indicator
105 matrix \mathbf{R} with elements:

$$R_{ij} = \begin{cases} 1, & t_i \leq t_j \\ 0, & t_i > t_j \end{cases} \quad (10)$$

106 We also define an indicator vector \mathbf{C} with elements given by the censoring of each patient. An operation
107 using \mathbf{R} replaces the conditional sum over $t_j \geq t_i$, and an operation using \mathbf{C} replaces the conditional
108 sum over $\mathbf{C}(i)=1$ in equation 4. In Theano, the partial log likelihood is:

$$\text{pl}=\text{T.sum}((\theta - \text{T.log}(\text{T.sum}(\text{T.exp}(\theta) * \mathbf{R}, \text{axis}=1))) * \mathbf{C}) \quad (11)$$

109 **Model evaluation**

110 To evaluate the performance of all methods in comparison, we trained each model on 80% of the samples
111 for each dataset (chosen randomly) and evaluated the performance on the 20% holdout test set. The
112 output of Cox-PH, Cox-nnet and CoxBoost are the log hazard ratios (i.e., Prognosis Index, or PI) for each
113 patient. The hazard ratio describes the relative risk of a patient compared to a non-parametric baseline.
114 On the other hand, the output of RF-S is an estimation of the survival time for each patient.

115 We use C-index and log-ranked p-value based on dichotomization of the hold-out test data of the holdout
116 test data to measure the performance of each model. The C-index is a measure of how well the model
117 prediction corresponds to the ranking of the survival data¹⁴. It is calculated for censored survival data,
118 which evaluates a value between 0 and 1, with 0.5 equivalent to a random process. The C-index can be
119 computed as a summation over all events in the dataset, whereby patients with a higher survival time and
120 lower log hazard ratios (and conversely patients with a lower survival time but higher log hazard ratios)
121 are considered concordant. The C-index is a measure of concordance of the data with the model
122 prediction. To calculate the log-ranked p-value, a PI cutoff threshold is used to dichotomize the patients
123 in the data set into higher and lower risk groups, similar to our earlier report^{15,16}. A log-ranked p-value is
124 then computed to differentiate the Kaplan-Meier survival curves between the higher vs. lower risk groups.
125 In this report, we used the median log hazard ratio as the cutoff threshold.

126 Feature evaluation

127 For computing the importance of a feature in Cox-nnet, we use a method of partial derivatives (PaD)^{17,18}.
128 For each patient, we compute the partial derivatives of each input with respect to the linear output of the
129 model (e.g., the log hazard ratio). The average of the partial derivatives for each input across all patient
130 samples is calculated as the feature score.

131 Datasets

132 In order to evaluate the performance of Cox-nnet, we analyzed 10 TCGA datasets which were combined
133 into a pan-cancer dataset. The TCGA datasets included the following cancer types: Bladder Urothelial
134 Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Head and Neck squamous cell carcinoma
135 (HNSC), Kidney renal clear cell carcinoma (KIRC), Brain Lower Grade Glioma (LGG), Liver
136 hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma
137 (LUSC), Ovarian serous cystadenocarcinoma (OV) and Stomach adenocarcinoma (STAD). RNA-Seq
138 expression and clinical data were downloaded from the Broad Institute GDAC¹⁹. Overall survival time
139 and censoring information were extracted from the clinical follow-up data. Raw count data were
140 normalized using the DESeq2 R package²⁰ and then log-transformed. Datasets were selected from
141 TCGA based on the following criteria: > 300 samples with both RNASeq and survival data and > 50
142 survival events. In total, 5031 patient samples were used (see Table S1 for a patient tabulation by
143 individual dataset).

144 Results

145 Cox-nnet structure and optimization

146 Cox-nnet is the neural network extension of the Cox-PH model. We created a package suitable for high
147 dimensional datasets using the Theano math library in Python. The neural network model used in this
148 paper is shown in Figure 1 and an overview of modules in the Cox-nnet package is shown in Figure S1.
149 As a proof of concept, the current ANN architecture is composed of three layers: one input layer, one

150 fully connected hidden layer and an output “Cox regression” layer. The output layer of Cox-nnet replaces
151 the linear predictors in the standard Cox-PH model. Many other functions are implemented to improve
152 the usability of the package, including CVSearch, CVProfile, CrossValidation, and TrainCoxMlp.
153 CVSearch, CVProfile, CrossValidation are methods that perform CV to find the optimal regularization
154 parameter. TrainCoxMlp performs optimization of coefficients on the regularized partial likelihood
155 function. The optimization strategies include momentum gradient descent²¹, Nesterov accelerated
156 gradient²² and Ada Delta²³. A comparison of these descent methods is shown in Figure S2A, where
157 Nesterov accelerated gradient method achieved the best efficiency based on TCGA kidney renal clear cell
158 carcinoma (KIRC) data. Moreover, this package can be run on multiple threads or a Graphics Processing
159 Unit (GPU), and it achieves slightly faster training time compared to Random Forest and CoxBoost
160 (Figure S2B). Thus, Cox-nnet is a modern software implementation that can achieve efficient
161 computational time.

162 **Performance comparison of survival prediction methods**

163 We compared four methods, including Cox-nnet, Cox-PH, CoxBoost and RF-S, on 10 datasets from The
164 Cancer Genome Atlas (TCGA), which were selected based on having at least 50 death events (Table S1).
165 For each dataset, we trained the model on 80% of the randomly selected samples and determined the
166 regularization parameter using 5-fold CV on the training set. We used two types of regularizations, L2
167 ridge regularization (also known as weight decay) and dropout regularization. We evaluated the
168 performance on the remaining 20% holdout test set. Two metrics are used to evaluate the performance of
169 the model. The first one is Harrell’s concordance index (C-index) calculated for censored survival data
170^{10,24}. It evaluates the relative ordering of the samples and ranges between 0 and 1, with 0.5 equivalent to a
171 random process. The second metric is the log-ranked p-value from Kaplan-Meier survival curves of two
172 different survival risk groups. This is done by using the median threshold of Prognosis Index (PI), the
173 output of Cox-nnet, to dichotomize the patients into higher and lower risk groups, similar to our earlier

174 reports^{15,16,24}. A log-ranked p-value is then computed to differentiate the Kaplan-Meier survival curves
175 from these two groups.

176 The comparison of C-indices among the four methods over the 10 TCGA data is shown in Figure 2A.
177 Overall, Cox-nnet has higher predictive accuracy over the other three methods, regardless of the
178 regularization method. Cox-PH performs the second best, followed by CoxBoost and RF-S in descending
179 order (Figure 2B). The comparison of log-ranked p-values on the dichotomized survival risk groups is
180 shown in Figure S3. Generally, log-ranked p-values in the 10 TCGA datasets are more significant in Cox-
181 nnet, compared to other methods. However, the dichotomization of patients ignores the differences within
182 each dichotomized group, thus the resulting log-ranked p-values are less consistent than C-indices on the
183 same data.

184 **Biological relevance of hidden layer nodes of Cox-nnet**

185 To explore the biological relevance of the hidden nodes of Cox-nnet, we used the TCGA KIRC dataset as
186 an example. We first extracted the contribution of each hidden node to the PI score for each patient
187 (Figure 3A). The contribution was calculated as the output value of each hidden node weighted by the
188 corresponding coefficient at the Cox regression output layer. As expected, the value of the hidden nodes
189 strongly correlated to the PI score. However, there is still significant heterogeneity among the nodes,
190 suggesting that individual nodes may reflect different biological processes. We hypothesize that the top
191 nodes may serve as surrogate features to discriminate patient survival. To explore this idea, we selected
192 the top 20 nodes with the highest variances, and presented the patients PI scores using t-SNE, a popular
193 method to enhance the separation among samples²⁵. The nodes represent a dimension reduction of the
194 original data and clearly discriminate samples by their PI scores (Figure 3B). In contrast, the top 20
195 principle components obtained from principal component analysis (PCA) in combination with t-SNE fail
196 to separate the patient samples (Figure 3B). This drastic difference demonstrates that the nodes in Cox-
197 nnet effectively capture the survival information, and the top node PI scores can be used as features for
198 dimension reduction in survival analysis.

199 To further explore the biological relevance of the top 20 hidden nodes, we conducted Gene Set
200 Enrichment Analysis (GSEA)²⁶ using KEGG pathways²⁷. We calculated significantly enriched pathways
201 using gene correlation to the output score of each node (Figure 3C and Table S2), and compared these
202 enriched pathways to those from GSEA of the Cox-PH model (Table S3). To calculate statistical
203 significance of the pathways, we performed 10,000 permutations, followed by multiple hypothesis testing
204 with Benjamini Hochberg adjustment. A total of 110 (out of 187) significantly enriched pathways (Table
205 S2) were identified in at least one node, including seven pathways enriched in all 20 nodes that were not
206 found by the Cox-PH method (Table 1). In contrast, Cox-PH only identified 30 significantly enriched
207 pathways using the same significance threshold. Among the seven pathways, the P53 signaling pathway
208 stands out as an important biologically relevant pathway (Figure 4 and Figure S4), since it was shown to
209 be highly prognostic of patient survival in kidney cancer²⁸.

210 Next, we estimated the predicative accuracies of the leading edge genes (LEG) enriched in the KEGG
211 pathways from Cox-nnet vs. those enriched in Cox-PH model. We used the C-index of each LEG,
212 obtained from single-variable analysis (Figure 4). Collectively, LEGs from Cox-nnet have significantly
213 higher C-index scores ($p = 5.79e-05$) than those from Cox-PH, suggesting that Cox-nnet has selected
214 more informative features. In order to visualize these gene level and pathway level differences between
215 Cox-nnet and Cox-PH, we reconstructed a bipartite graph between LEGs for Cox-nnet or feature genes
216 (for Cox-PH) and their corresponding enriched pathways (Figure 5). Besides P53 pathway mentioned
217 earlier that is specific to Cox-nnet, several other pathways, such as insulin signaling pathway, endocytosis
218 and adherens junction, also have many more genes enriched in Cox-nnet. Among them, some have been
219 previously reported to relevant to renal carcinoma development and prognosis, such as CASP9²⁹,
220 TGFBR2³⁰, KDR (VEGFR)³¹. These results demonstrate that Cox-nnet model reveals richer biological
221 information than Cox-PH.

222 To further examine the importance of each gene relative to the survival outcome, we calculated the
223 averaged partial derivative (PaD) of each input gene feature over all patients, with respect to the linear

224 output of the model (e.g., the log hazard ratio). As demonstrated by the LEGs in seven common pathways
225 of all nodes in Cox-nnet, the feature importance scores produce stronger biological insight (Figure S4).
226 For example, the feature importance for the BAI1 gene in the P53 pathway is much higher in the Cox-
227 nnet model compared to the Cox-PH model. Corresponding to our finding, the BAI gene family was
228 found to be involved in several types of cancers including renal cancer^{32 33 34 35}. BAI1 acts as an inhibitor
229 to angiogenesis and is transcriptionally regulated by P53³⁶. Its expression level was significantly
230 decreased in tumor vs. normal kidney tissue, and was even lower in advanced stage renal carcinoma³⁵.
231 Mice kidney cancer models treated with BAI1 showed slower tumor growth and proliferation³⁷.
232 Additionally, the MAPK1 gene (also known as ERK2) has a much higher feature importance score in
233 Cox-nnet compared to Cox-PH, and is annotated in the Adherens Junction pathway as well as the Insulin
234 Signalling Pathway found by Cox-nnet. MAPK1 is one of the key kinases in intra-cellular transduction,
235 and was found constitutively activated in renal cell carcinoma³⁸. Drugs inhibiting the MAPK cascade
236 have been targeted for development³⁹.

237

238 **Discussion**

239 In this report, we have implemented Cox-nnet, a new non-linear ANN method, to predict patient survival
240 from high throughput omics data. Cox-nnet is an improved, modern alternative to the standard Cox-PH
241 regression, as demonstrated by increased performance for survival prediction and the capabilities to
242 explore more deeply the biological information.

243 First, through in-depth comparison of 10 TCGA RNA-Seq, Cox-nnet achieves overall statistically
244 significant improvements over Cox-PH on its predictive accuracy, as measured by C-indices.
245 Interestingly, the ensemble-based method RF-S consistently ranks worse than Cox-nnet and Cox-PH.
246 Because RF-S bootstraps both samples and features for individual trees, many uninformative features in
247 each tree may be chosen for node splitting in particularly high dimensional datasets, leading to a decrease

248 in overall accuracy⁴⁰. In contrast, the dropout and L2-regularization approach used by both Cox-nnet and
249 Cox-PH can prune out uninformative features.

250 Second, Cox-nnet can reveal a lot richer biological information than Cox-PH. This is manifested both at
251 the pathway and gene levels. The hidden nodes in the Cox-nnet model have distinct expression patterns,
252 and can serve as surrogate features for survival-sensitive dimension reduction. Many more significant
253 KEGG pathways are enriched which correlate with top nodes in Cox-nnet, as compared to those from the
254 Cox-PH model. A critical pathway for renal cancer development, P53 pathway, is only enriched by Cox-
255 nnet but not Cox-PH model in TCGA KIRC. Other pathways, including insulin signaling pathway,
256 endocytosis and adherens junction, have many more genes enriched by Cox-mmet. Moreover, leading
257 edge genes (LEGs) obtained from these KEGG pathways enriched by Cox-nnet (which are a fraction of
258 the gene features considered by the model) have collectively higher associations with survival.

259 As a promising new predictive method for prognosis, the current Cox-nnet implementation has some
260 limitations. Its architecture includes 3-layer ANN, and it is possible to incorporate other more
261 sophisticated architecture into the model, such as including more layers of neurons. A convolutional
262 neural network approach using convolutional and pooling layers could also be used, as those reported in
263 processing imaging or other types of positional data⁴¹. Additionally, it is possible to embed *a priori*
264 biological pathway information into the network architecture, e.g., by connecting genes in a pathway to a
265 common node in the next hidden layer of neurons. In the future, we plan to further analyze how different
266 neural network architectures affect the performance of Cox-nnet and compare the biological insights from
267 the various models.

268 **Author contributions**

269 LXG and TC envisioned the project and designed the work. TC coded the project and conducted the
270 analysis. XZ provided insight and discussion on neural networks. All authors have read, revised and
271 approved the final manuscript.

272 Competing financial interests

273 The author(s) declare no competing financial interests.

274 Acknowledgements

275 This research was supported by grants K01ES025434 awarded by NIEHS through funds provided by the
276 trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), P20 COBRE GM103457
277 awarded by NIH/NIGMS, R01 LM012373 awarded by NLM, R01 HD084633 awarded by NICHD, and
278 Hawaii Community Foundation Medical Research Grant 14ADVC-64566 to L.X. Garmire.

279 References

- 280 1 McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* **5**, 115-133 (1943).
- 281 2 Jones, N. (Nature Publishing Group MACMILLAN BUILDING, 4 CRINAN ST, LONDON N1 9XW, ENGLAND, 2014).
- 282 3 Faraggi, D. & Simon, R. A neural network model for survival data. *Statistics in medicine* **14**, 73-82 (1995).
- 283 4 Petalidis, L. P. *et al.* Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. *Molecular cancer therapeutics* **7**, 1013-1024 (2008).
- 284 5 Chi, C.-L., Street, W. N. & Wolberg, W. H. in *AMIA Annual Symposium Proceedings*. 130 (American Medical Informatics Association).
- 285 6 Joshi, R. & Reeves, C. in *Proceedings of the eighteenth international conference on systems engineering*. 179-184.
- 286 7 Therneau, T. M. & Grambsch, P. M. *Modeling survival data: extending the Cox model*. (Springer Science & Business Media, 2000).
- 287 8 Binder, H. CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks. *R package version 1* (2013).
- 288 9 Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The Annals of Applied Statistics*, 841-860 (2008).
- 289 10 Koziol, J. A. & Jia, Z. The concordance index C and the Mann–Whitney parameter $\Pr(X > Y)$ with randomly censored data. *Biometrical Journal* **51**, 467-474 (2009).
- 290 11 van Houwelingen, H. C., Bruinsma, T., Hart, A. A., van't Veer, L. J. & Wessels, L. F. Cross-validated Cox regression on microarray gene expression data. *Statistics in medicine* **25**, 3201-3216 (2006).
- 291 12 Haykin, S. & Network, N. A comprehensive foundation. *Neural Networks* **2** (2004).
- 292 13 Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**, 1929-1958 (2014).

- 307 14 Harrell, F. E., Lee, K. L. & Mark, D. B. Tutorial in biostatistics multivariable prognostic models:
308 issues in developing models, evaluating assumptions and adequacy, and measuring and reducing
309 errors. *Statistics in medicine* **15**, 361-387 (1996).
- 310 15 Huang, S., Yee, C., Ching, T., Yu, H. & Garmire, L. X. A Novel Model to Combine Clinical and
311 Pathway-Based Transcriptomic Information for the Prognosis Prediction of Breast Cancer. *PLoS*
312 *computational biology* **10**, e1003851 (2014).
- 313 16 Huang, S. *et al.* Novel personalized pathway-based metabolomics models reveal key metabolic
314 pathways for breast cancer diagnosis. *Genome medicine* **8**, 1 (2016).
- 315 17 Gevrey, M., Dimopoulos, I. & Lek, S. Review and comparison of methods to study the
316 contribution of variables in artificial neural network models. *Ecological modelling* **160**, 249-264
317 (2003).
- 318 18 Olden, J. D., Joy, M. K. & Death, R. G. An accurate comparison of methods for quantifying
319 variable importance in artificial neural networks using simulated data. *Ecological Modelling* **178**,
320 389-397 (2004).
- 321 19 Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July
322 2014. *Broad Institute of MIT and Harvard*, doi:10.7908/C1DN43V9 (2014).
- 323 20 Love, M., Anders, S. & Huber, W. Differential analysis of RNA-Seq data at the gene level using
324 the DESeq2 package. (2013).
- 325 21 Qian, N. On the momentum term in gradient descent learning algorithms. *Neural networks* **12**,
326 145-151 (1999).
- 327 22 Bengio, Y., Boulanger-Lewandowski, N. & Pascanu, R. in *Acoustics, Speech and Signal Processing*
328 (*ICASSP*), 2013 IEEE International Conference on. 8624-8628 (IEEE).
- 329 23 Battiti, R. Accelerated backpropagation learning: Two optimization methods. *Complex systems* **3**,
330 331-342 (1989).
- 331 24 Wei, R. *et al.* Meta-dimensional data integration identifies critical pathways for susceptibility,
332 tumorigenesis and progression of endometrial cancer. *Oncotarget* (2016).
- 333 25 Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*
334 **9**, 2579-2605 (2008).
- 335 26 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for
336 interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*
337 **102**, 15545-15550 (2005).
- 338 27 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*
339 **28**, 27-30 (2000).
- 340 28 Girgin, C. *et al.* P53 mutations and other prognostic factors of renal cell carcinoma. *Urologia*
341 *internationalis* **66**, 78-83 (2001).
- 342 29 Marques, I. *et al.* Influence of survivin (BIRC5) and caspase-9 (CASP9) functional polymorphisms
343 in renal cell carcinoma development: a study in a southern European population. *Molecular*
344 *biology reports* **40**, 4819-4826 (2013).
- 345 30 Akhurst, R. J. & Deryck, R. TGF- β signaling in cancer—a double-edged sword. *Trends in cell*
346 *biology* **11**, S44-S51 (2001).
- 347 31 Choueiri, T. K. *et al.* Phase II and biomarker study of the dual MET/VEGFR2 inhibitor foretinib in
348 patients with papillary renal cell carcinoma. *Journal of Clinical Oncology* **31**, 181-186 (2013).
- 349 32 Cork, S. M. & Van Meir, E. G. Emerging roles for the BAI1 protein family in the regulation of
350 phagocytosis, synaptogenesis, neurovasculature, and tumor development. *Journal of molecular*
351 *medicine* **89**, 743-752 (2011).
- 352 33 Fukushima, Y. *et al.* Brain-specific angiogenesis inhibitor 1 expression is inversely correlated with
353 vascularity and distant metastasis of colorectal cancer. *International journal of oncology* **13**, 967-
354 970 (1998).

355 34 Lee, J. *et al.* Comparative study of angiostatic and anti-invasive gene expressions as prognostic
356 factors in gastric cancer. *International journal of oncology* **18**, 355-362 (2001).
357 35 Izutsu, T., Konda, R., Sugimura, J., Iwasaki, K. & Fujioka, T. Brain-specific angiogenesis inhibitor 1
358 is a putative factor for inhibition of neovascular formation in renal cell carcinoma. *The Journal of*
359 *urology* **185**, 2353-2358 (2011).
360 36 Nishimori, H. *et al.* A novel brain-specific p53-target gene, BAI1, containing thrombospondin
361 type 1 repeats inhibits experimental angiogenesis. *Oncogene* **15**, 2145-2150 (1997).
362 37 Kudo, S. *et al.* Inhibition of tumor growth through suppression of angiogenesis by brain-specific
363 angiogenesis inhibitor 1 gene transfer in murine renal cell carcinoma. *Oncology reports* **18**, 785-
364 792 (2007).
365 38 Oka, H. *et al.* Constitutive activation of mitogen-activated protein (MAP) kinases in human renal
366 cell carcinoma. *Cancer research* **55**, 4182-4187 (1995).
367 39 Friday, B. B. & Adjei, A. A. Advances in targeting the Ras/Raf/MEK/Erk mitogen-activated protein
368 kinase cascade with MEK inhibitors for cancer therapy. *Clinical Cancer Research* **14**, 342-346
369 (2008).
370 40 Nguyen, T.-T., Huang, J. Z. & Nguyen, T. T. Unbiased Feature Selection in Learning Random
371 Forests for High-Dimensional Data. *The Scientific World Journal* **2015** (2015).
372 41 LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. *The*
373 *handbook of brain theory and neural networks* **3361**, 1995 (1995).

374

375

376

377 **Figures**

378 Figure 1. An overview of the neural network architecture used in this study.

379 Figure 2. A. Barplot of the C-index of the 10 TCGA datasets using four prognosis-predicting methods
380 (Cox-nnet, CoxBoost, Cox-PH and RF-S). Each dataset was randomly split into 80% training and 20%
381 testing sets. B. Heatmap of the performance rank of each dataset.

382 Figure 3. A. Hidden node output of the TCGA KIRC dataset. B. t-SNE plot of the top 20 hidden nodes
383 and the top 20 principal components in PCA analysis. C. Gene Set Enrichment Analysis: significantly
384 enriched KEGG pathways of the top 20 hidden nodes (adjusted p-value < 0.05).

385 Figure 4. Single variable C-index scores of the leading edge genes from Cox-nnet and Cox-PH. Cox-nnet
386 has significantly higher C-index scores ($p = 5.79e-5$).

387 Figure 5. Enriched pathway-gene bipartite network from the leading edge genes and significantly
388 enriched pathways. Significantly enriched pathways common to all 20 hidden nodes are labeled in green.
389 Leading edge genes found uniquely in Cox-nnet are labeled in orange, and genes found in both Cox-nnet
390 and Cox-PH are labeled in blue.

391 Table 1. Cox-nnet node-associated pathways. Significantly enriched pathways from common to all 20
392 hidden nodes that are not found in the Cox-PH Gene Set Enrichment Analysis (Adjusted p < 0.05).

Pathway	P.value	P.adjusted	Nodes
KEGG adherens junction	0.000	0.001	1-20
KEGG endocytosis	0.000	0.001	1-20
KEGG insulin signaling pathway	0.000	0.001	1-20
KEGG lysine degradation	0.000	0.003	1-20
KEGG p53 signaling pathway	0.000	0.003	1-20
KEGG pyruvate metabolism	0.000	0.001	1-20
KEGG sphingolipid metabolism	0.001	0.005	1-20

393

394 **Supplemental Figures**

395 Figure S1. An overview of the structure, methods and classes in Cox-nnet package.

396 Figure S2. A: comparison of descent methods on the TCGA KIRC dataset. The change in cost function
397 is evaluated over 100,000 iterations for three methods: gradient descent, momentum gradient descent and
398 the Nesterov accelerated gradient. B: Training time comparing CPU training time vs. GPU training time
399 on the same dataset.

400 Figure S3. Log-rank p-value bar plot of the 10 TCGA datasets.

401 Figure S4. Variable importance of the common leading edge genes of enriched KEGG pathways.

402

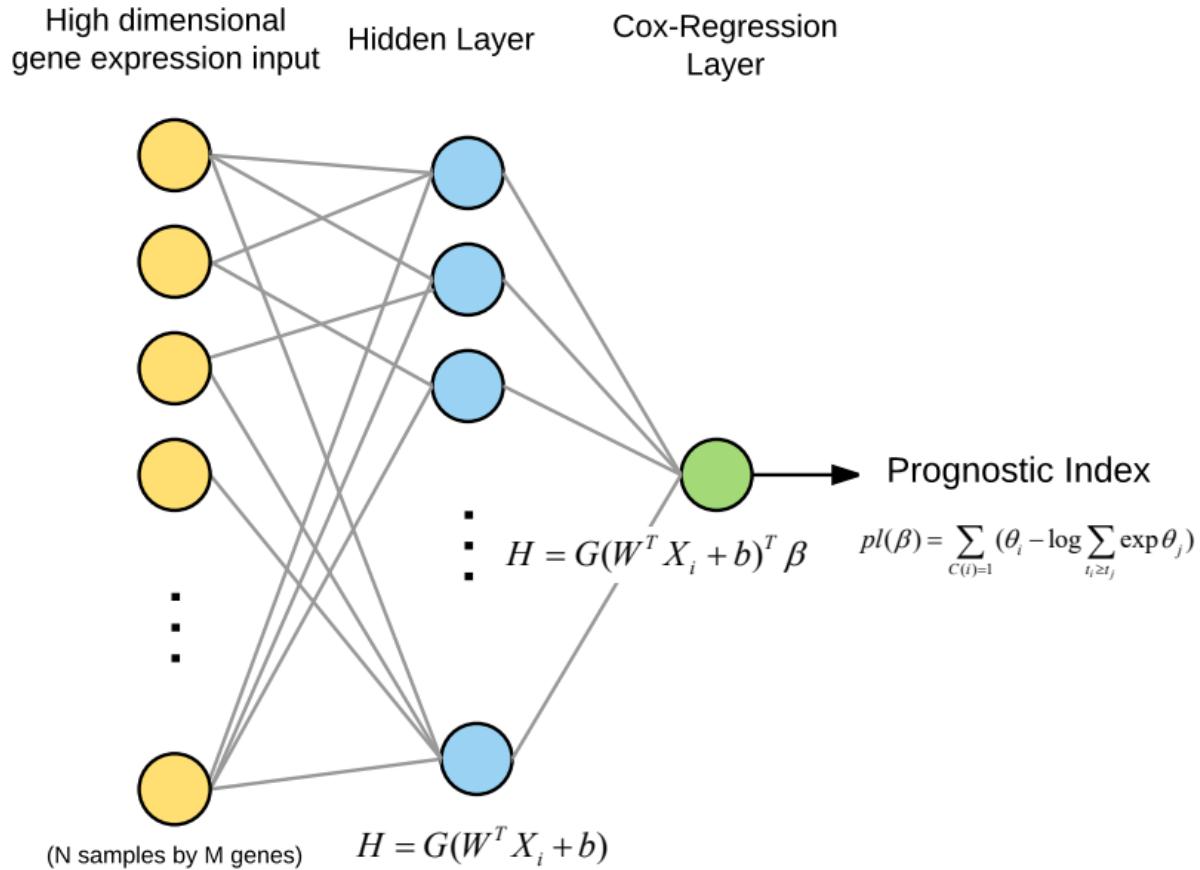
403 Table S1. Tabulation of TCGA patients by individual dataset.

404 Table S2. Significantly enriched pathways from the Cox-PH method ($p < 0.05$).

405 Table S3. Significantly enriched pathways from the Cox-nnet method ($p < 0.05$).

406

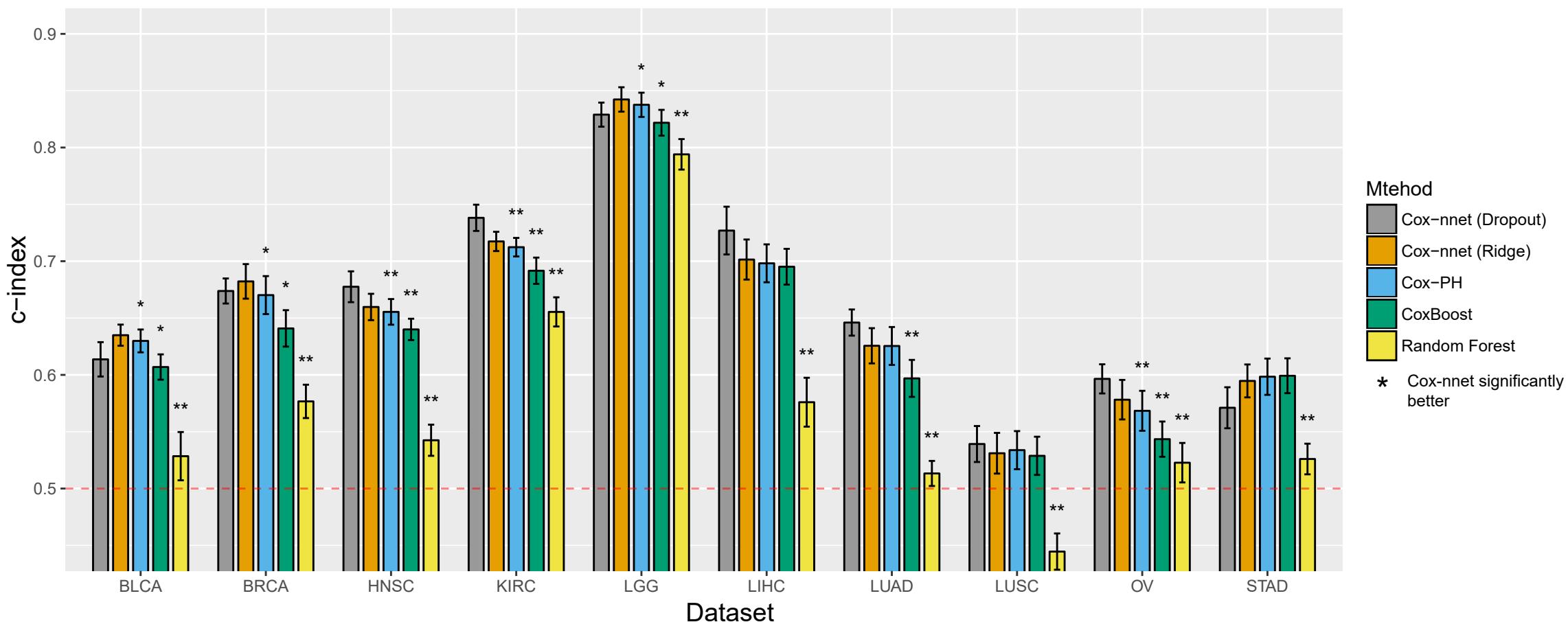
407



A

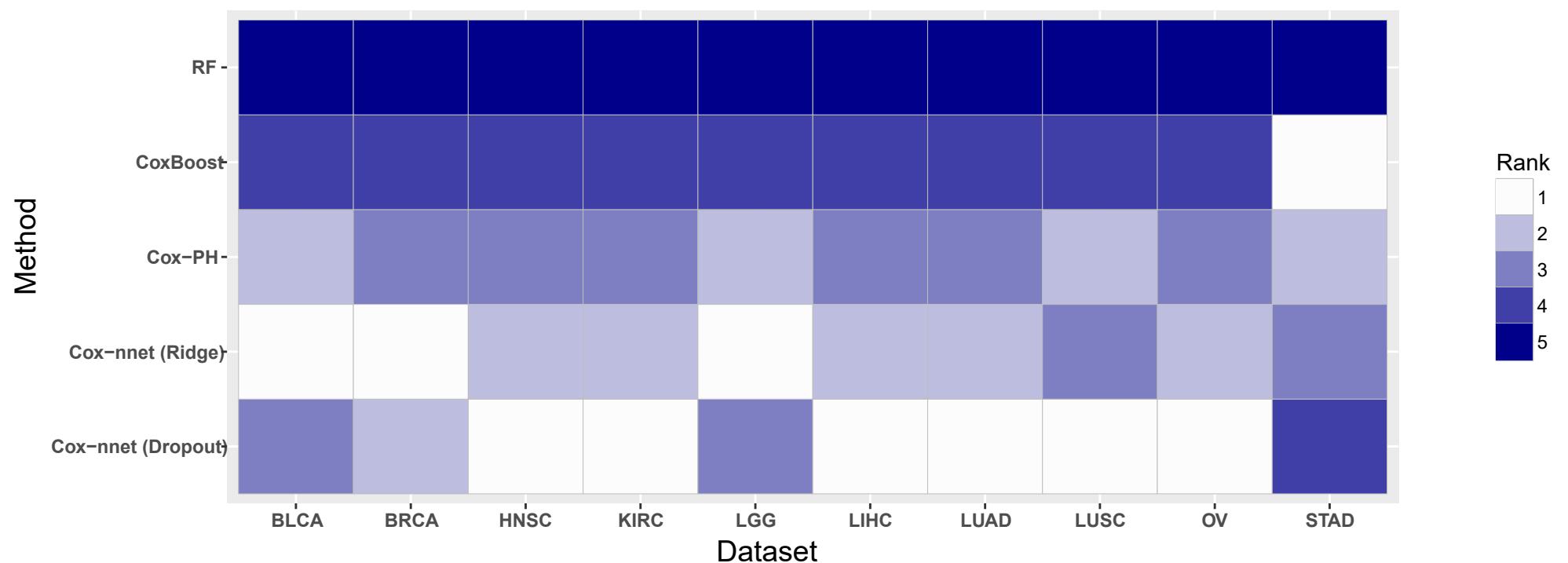
bioRxiv preprint doi: <https://doi.org/10.1101/093021>; this version posted December 11, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

Performance across TCGA datasets

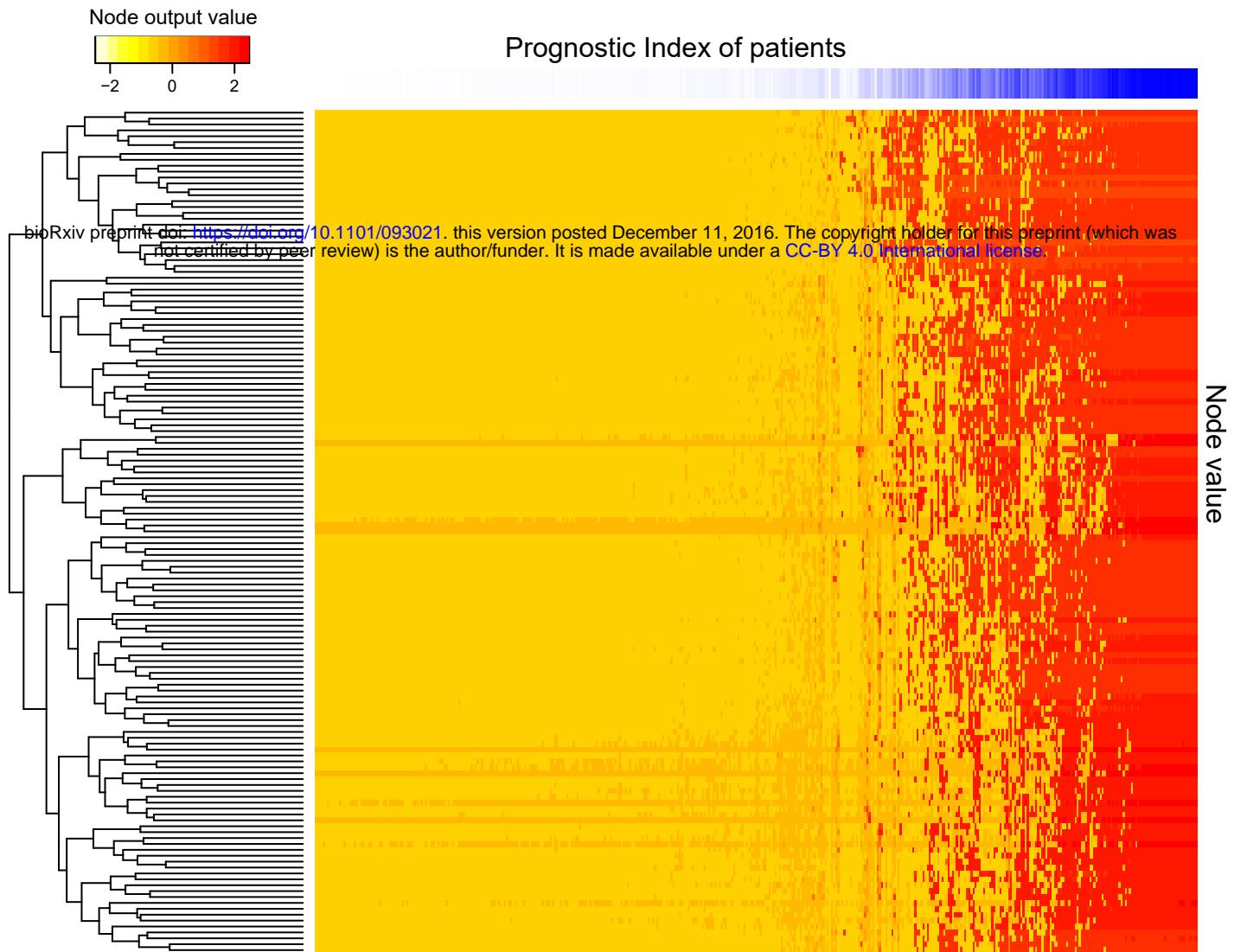


B

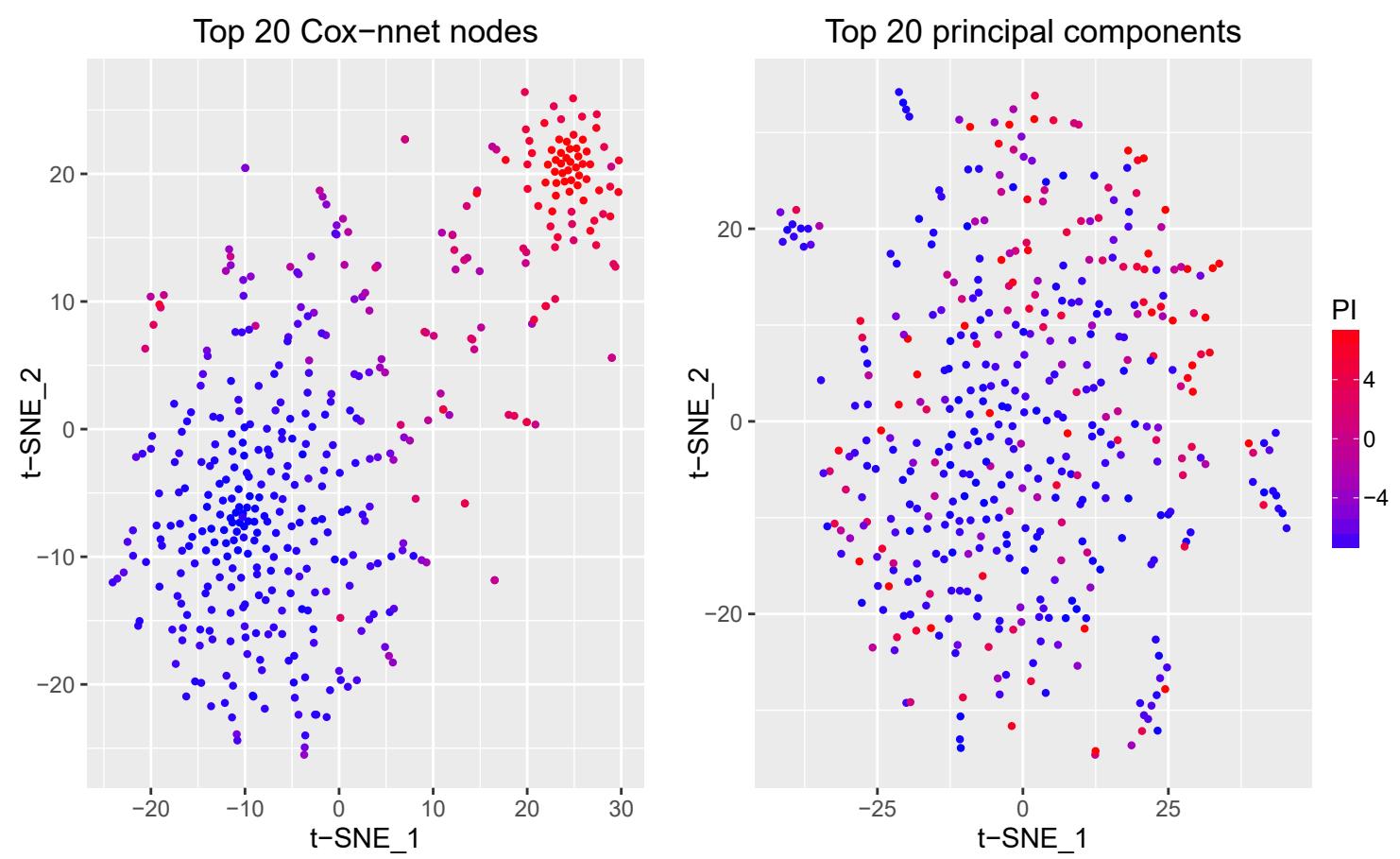
Performance Rank



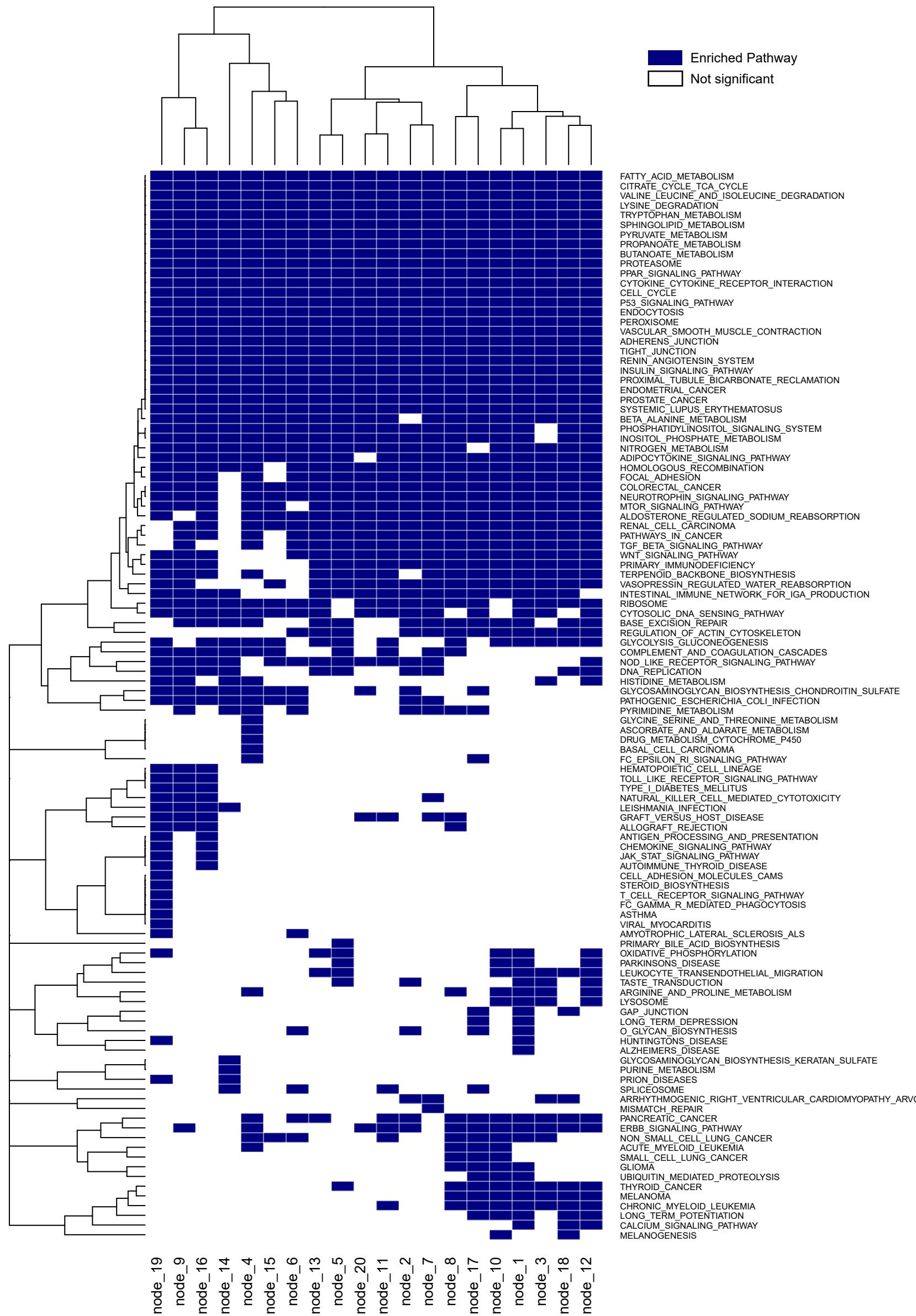
A Hidden layer node outputs of KIRC dataset



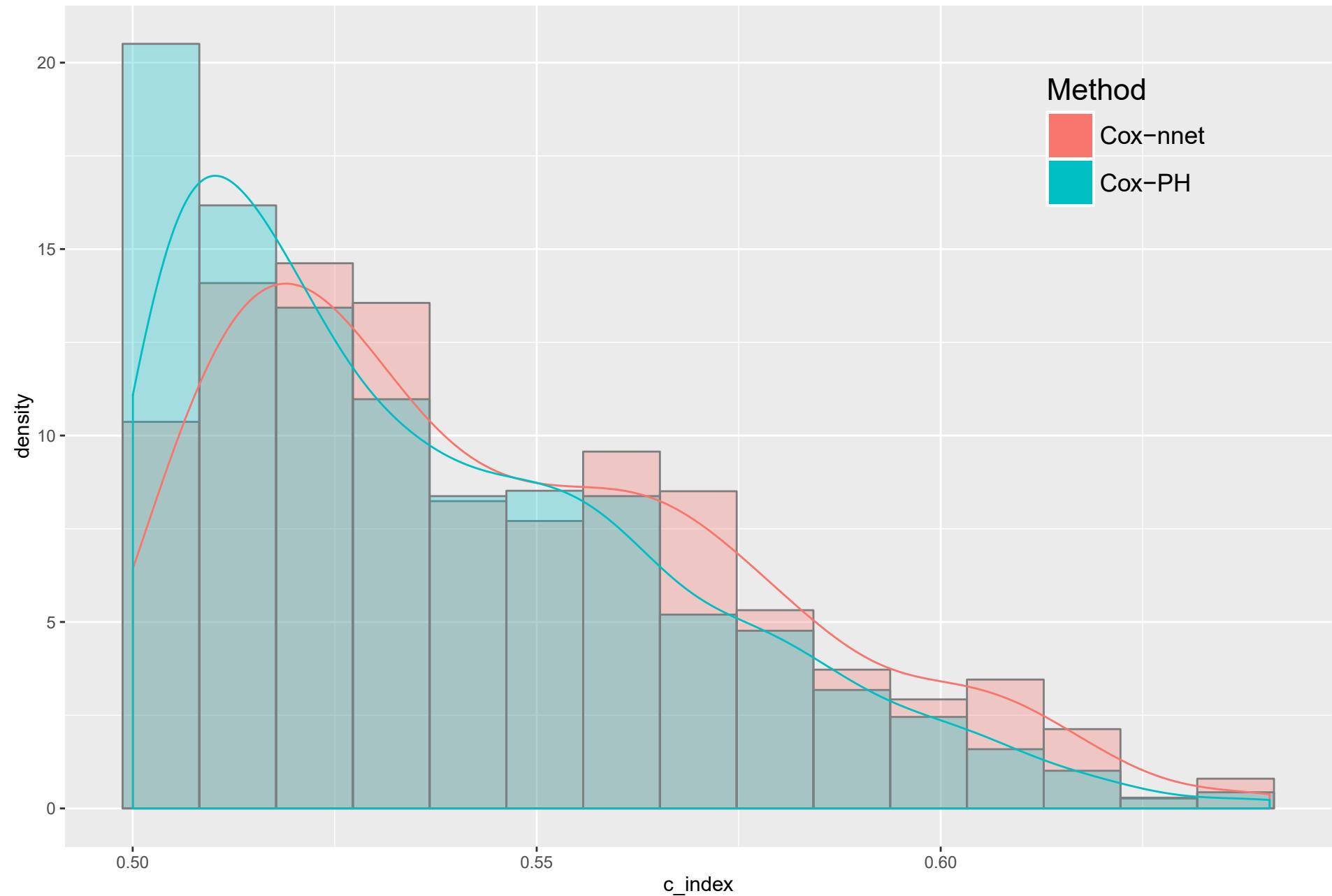
B t-SNE plot of top 20 Cox-nnet nodes



C Node clustering by pathway



Single variable c-index of leading edge genes



Enriched pathway-gene bipartite network

Enriched Pathway

- Cox-nnet specific leading edge genes
- Leading edge genes found by both Cox-nnet and Cox-PH models

