

PNAS



1

2 **Supporting Information for**

3 **Learning to Estimate Sample-specific Transcriptional Networks for 7000 Tumors**

4 **C.N. Ellington, B.J. Lengerich, T.B.K. Watkins, J. Yang, A. Adduri, S. Mahbub, H. Xiao, M. Kellis, and E.P. Xing**

5 **Corresponding Authors Eric P. Xing and Manolis Kellis.**

6 **E-mails: epxing@cs.cmu.edu, manoli@mit.edu**

7 **This PDF file includes:**

8 Supporting text

9 Figs. S1 to S6

10 Tables S1 to S3

11 Legends for Dataset S1 to S2

12 SI References

13 **Other supporting materials for this manuscript include the following:**

14 Datasets S1 to S2

15 Supporting Information Text

16 1. Resources and Reproducibility

- 17 • Contextualized documentation and installation instructions
18 <https://contextualized.ml/>
- 19 • Code for model training, plotting, and evaluation
20 <https://github.com/cnellington/CancerContextualized>
- 21 • Preprocessed data and contextualized networks for all 7997 patients with split labels
22 <https://zenodo.org/records/14885352>
- 23 • TCGA data download portal
24 <https://portal.gdc.cancer.gov/>
- 25 • TCGA reference subtypes
26 <https://bioconductor.org/packages/3.21/bioc/vignettes/TCGAbiolinks/inst/doc/subtypes.html>

27 2. Implementation

28 Our entire framework is implemented in PyTorch using the PyTorch Lightning framework within our open-source software
29 [Contextualized](#) (1). The context encoder, network archetypes, and contextualized network models are learned simultaneously
30 using end-to-end backpropagation of the network loss (defined in Methods).

31 **Training** The context data views (B) were concatenated sample-wise to create a single context feature vector encompassing all
32 views for each patient. Categorical features were one-hot encoded. Healthy samples were given ideal genomic profiles with
33 no SCNAs or SNVs. Other missing values were imputed with the feature mean. The full context vectors were compressed
34 to 200 features using a PCA learned on the training-validation split. We split our dataset into 80% training-validation and
35 20% testing. We created 30 bootstraps of the training-validation set and finally split into 80% training and 20% validation,
36 resulting in a 64-16-20 split for train-validation-test.

37 We bootstrapped our models, sampling the train-validation split with replacement to create 30 bootstraps. For each
38 bootstrap, we also initialized the models randomly. We trained the models using a batch-size of 10 and a learning rate of 1e-3.
39 We used early-stopping with a patience of 5 to end training when the minimum validation loss had not improved for 5 epochs.
40 We retain only the model with the minimum validation loss for each bootstrap.

41 Following this, we use the trained models to infer 30 networks (one from each bootstrap) for each sample in the original
42 non-bootstrapped training-validation set as well as the test set. To obtain the final predicted networks for each sample,
43 we average the network parameters over the bootstraps to get a single bootstrap-averaged network for each sample. The
44 mean-squared errors of the network relative to the sample expression is defined in the Methods (the mean of the linear residuals
45 in the network objectives). We report the mean error of these networks averaged over samples in the test set. Rather than
46 bootstrapping this entire procedure again to report the standard deviation of this result, we upper-bound the uncertainty by
47 reporting the standard deviation of the errors from the individual bootstrap models. This procedure is repeated for all of the
48 baseline methods. All plots can be reproduced using the error files uploaded to Zenodo with the pre-trained networks.

49 Additionally, we tested both gradient-based optimization and off-the-shelf SKLearn solvers for the baselines where possible,
50 applying early stopping with the gradient-based solvers to control overfitting across all methods. However, we found that the
51 SKLearn solvers were superior when applicable, and reported the best baseline results.

52 **Context Encoder** The context encoder is implemented as a multi-layer perceptron with 3 hidden layers, each 100 neurons wide
53 with ReLU activations. Model weights are initialized as `Uniform[-0.01, 0.01]`.

54 The context encoder is a highly flexible component of our framework and a driving force for future work. It can be used to
55 enforce assumptions about the relationships between contexts and models, between context features, and about the archetype
56 space. For instance, by using a multi-layer perceptron in this study, we naturally handle colinearities in a high-dimensional
57 feature space. Using a neural additive model (2, 3) instead of a multi-layer perceptron would provide context-feature-specific
58 archetype weights for interpretability. Similarly, the context encoder can be implemented as a convolutional neural network for
59 images (4) or a recurrent neural network for time series (5). At the context encoder head, we currently use an unconstrained
60 output, but applying a softmax activation would require all of the sample-specific models to lie within a polytope defined by
61 the archetypal networks.

62 **Hardware Limitations** 50 metagenes were chosen due to hardware limitations, as the largest number of metagenes we tried
63 which would fit in memory for training and inference on a MacBook Air 2020 with 16G RAM in 32-bit precision. This occupied
64 11.4G of memory while writing 30 bootstraps * 7997 samples * 50² sample-specific network parameters to disk.

65 3. Data

66 **A. Data sources.** The Cancer Genome Atlas* (TCGA) is a publicly-available pan-cancer datasource containing genomic,
67 transcriptomic, and clinical profiling of tumors from dozens of landmark studies. We queried TCGA for open access samples
68 with bulk RNA-sequencing and merged this dataset with two follow-up studies on an overlapping set of patients.

69 **Somatic copy number alterations (SCNAs)** SCNAs affect a larger fraction of the genome than do any other type of somatic
70 genetic alteration (6) and are a major driver of expression variation in cancer (7). We used copy number profiles derived from
71 TCGA samples using ASCAT (8) from a pan-cancer study of the role of allele-specific SCNAs in cancer (9).

72 **Driver single-nucleotide mutations (SNVs)** SNVs can be classified into "driver" mutations thought to provide selective growth
73 advantage and "passenger" mutations thought to have little role in promoting cancer development. We incorporated driver
74 SNVs from the TCGA-derived CHASMplus dataset (10)

75 B. Context Data Views.

76 **Clinical information** This data view incorporates sample tissue-of-origin, race, age at diagnosis, gender, year of birth, and days
77 to collection provided by TCGA.

78 **Biopsy Composition** This data view contains the sample's percent tumor cells, percent normal cells, percent tumor nuclei,
79 percent monocyte infiltration, percent lymphocyte infiltration, and percent neutrophil infiltration provided by TCGA. We also
80 incorporate expression-derived estimates of the fraction of a sample consisting of tumor cells from (9).

81 **Copy Number Alterations** From ASCAT (8), we gather whole genome doubling events as well as gain and loss events for
82 bp-specific regions of hg19 based on data from (11). We transform these gain and loss events into both arm-level and gene-level
83 events, where arm-level events affect 85% of an entire arm in the same event, while genes-level events affect a single gene. We
84 transform these into number of major and number of minor chromosome arms, and the number of major and minor alleles for
85 the set of 295 genes that overlap between COSMIC (12) and MSigDB (13). For both gene and arm-level events, we create a
86 separate indicator for loss of heterozygosity on each gene.

87 **Driver Mutations** From CHASMplus (10) we gather the mutations on all COSMIC (12) oncogenes/tumor suppressor genes and
88 binarize the presence or absence of a mutation in each gene.

89 **C. Baselines.** We are not aware of any other scalable meta-learning, deep learning, or varying-coefficient methods to produce
90 context-informed correlation, Markov, and neighborhood selection networks under a universal framework. State-of-the-art
91 gene regulatory network estimators are limited to population, cohort, and cluster-based approaches tailored to a single network
92 model class (14–16). As such, our baselines apply the network estimators in 2 under several well-known and general paradigms
93 for improving model personalization, broadly relating to subpopulation or cluster analysis. Our population baseline provides
94 no personalization, learning a single model for the entire population of training samples. Our context-clustered baseline takes
95 an unsupervised approach to personalization by first doing a k-means clustering with k=25 on the aggregated context views (3)
96 and then inferring cluster-specific networks. Our disease-clustered baseline uses a personalization oracle, grouping samples by
97 tumor type and then inferring disease-specific networks.

98 4. Extra Results

*www.cancer.gov/tcga

99 **A. Simulations.** We perform simulations to study the scaling properties of contextualized networks over network size (Fig. S1a,
 100 S2a) and context dimensionality (Fig. S1b, S2b).

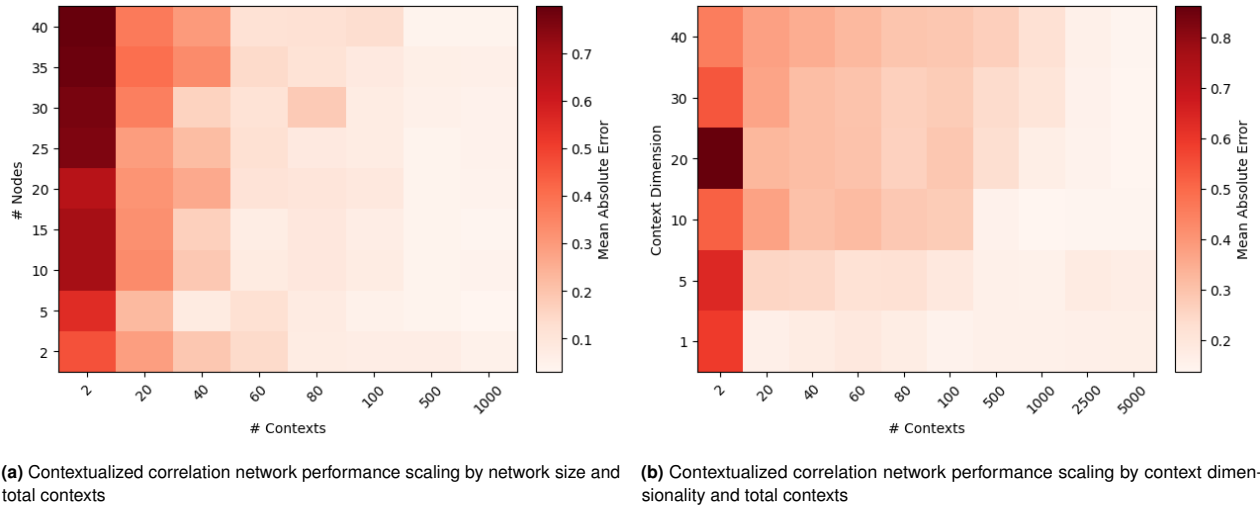


Fig. S1. Performance of estimators on held-out data, measured across (a) network size and (b) context dimensionality (i.e. complexity of context dependence) measured as the Mean-absolute Error (MAE) between predicted and ground truth correlation networks, averaged over three bootstrapped runs.

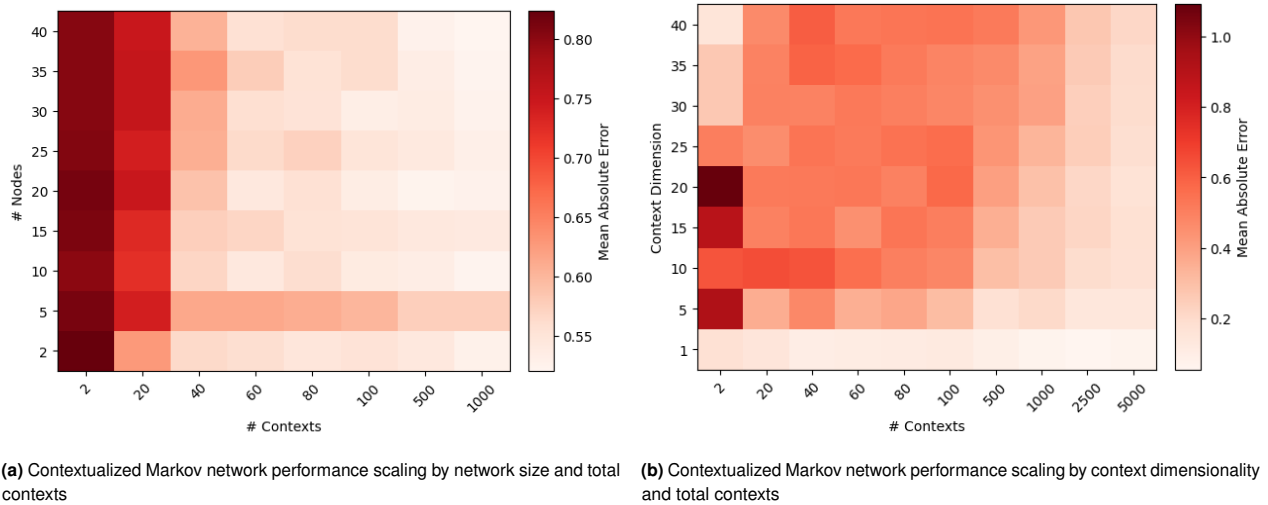


Fig. S2. Performance of estimators on held-out data, measured across (a) network size and (b) context dimensionality (i.e. complexity of context dependence) measured as the Mean-absolute Error (MAE) between predicted and ground truth Markov networks, averaged over three bootstrapped runs.

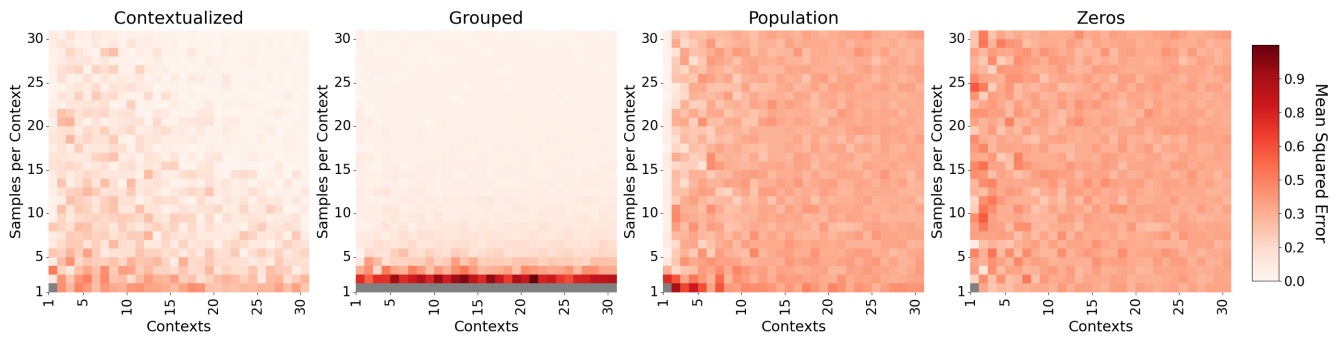


Fig. S3. Performance of correlation network estimators in terms of predicted versus true network parameters, measured over samples per context task and total contexts. Mean-squared error (MSE) is between the predicted and ground truth correlation matrices defining the correlation networks, averaged over five bootstrapped runs. Population estimates a single model for all contexts, grouped estimates a model for each context separately. Contextualization is the only method to scale both horizontally and vertically.

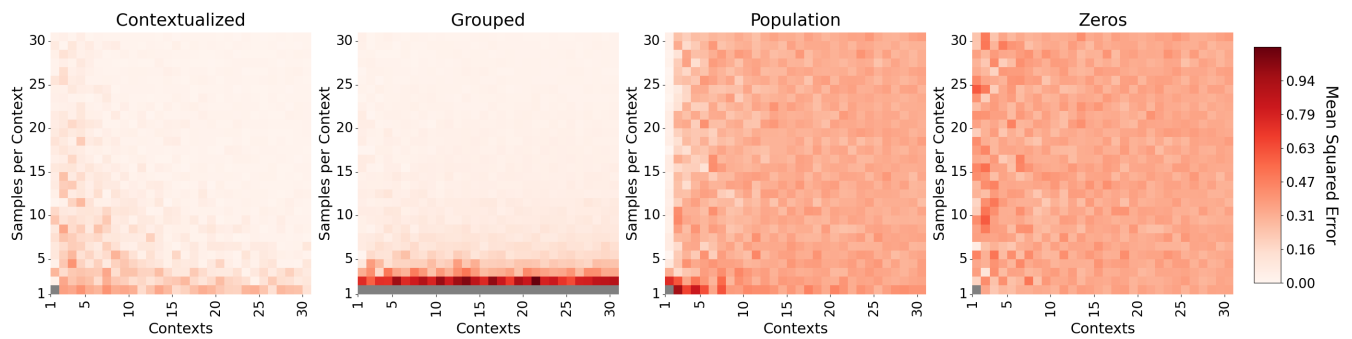


Fig. S4. Performance of Markov network estimators in terms of predicted versus true network parameters, measured over samples per context task and total contexts. Mean-squared error (MSE) is between the predicted and ground truth precision matrices defining the Markov networks, averaged over five bootstrapped runs. Population estimates a single model for all contexts, grouped estimates a model for each context separately. Contextualization is the only method to scale both horizontally and vertically.

Disease Type	Reference Subtypes	Expression Subtypes	Network Subtypes
Bladder Urothelial Carcinoma	0.411	0.151	0.713
Breast invasive carcinoma	1.484	0.616	1.558
Cholangiocarcinoma	–	–	–
Colon adenocarcinoma	0.016	0.014	1.219
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	–	–	–
Esophageal carcinoma	0.044	0.884	0.049
Glioblastoma multiforme	1.644	0.958	0.101
Head and neck squamous cell carcinoma	1.209	0.312	3.465
Kidney Chromophobe	0.715	13.802	0.211
Kidney renal clear cell carcinoma	5.042	6.109	13.741
Kidney renal papillary cell carcinoma	14.538	10.582	15.205
Brain Lower Grade Glioma	48.338	33.438	49.681
Liver hepatocellular carcinoma	0.009	0.427	0.827
Lung adenocarcinoma	0.687	1.172	0.507
Lung squamous cell carcinoma	0.123	0.105	0.249
Ovarian serous cystadenocarcinoma	0.704	0.684	0.05
Pancreatic adenocarcinoma	0.439	1.104	1.494
Prostate adenocarcinoma	–	–	–
Rectum adenocarcinoma	0.221	0.203	0.117
Skin cutaneous melanoma	–	–	–
Stomach adenocarcinoma	0.044	1.117	0.575
Thyroid adenocarcinoma	0.298	0.164	2.104
Uterine Corpus Endometrial Carcinoma	6.937	3.343	7.07
Uterine Carcinosarcoma	0.319	0.023	0.048
Uveal Melanoma	4.838	2.589	3.565

Table S1. Multivariate log-rank test comparison across different subtyping methods in terms of $-\log(p\text{-value})$. Only samples shared between all datasets are used to control for power. – indicates no samples are shared, or subtypes do not exist for TCGA.

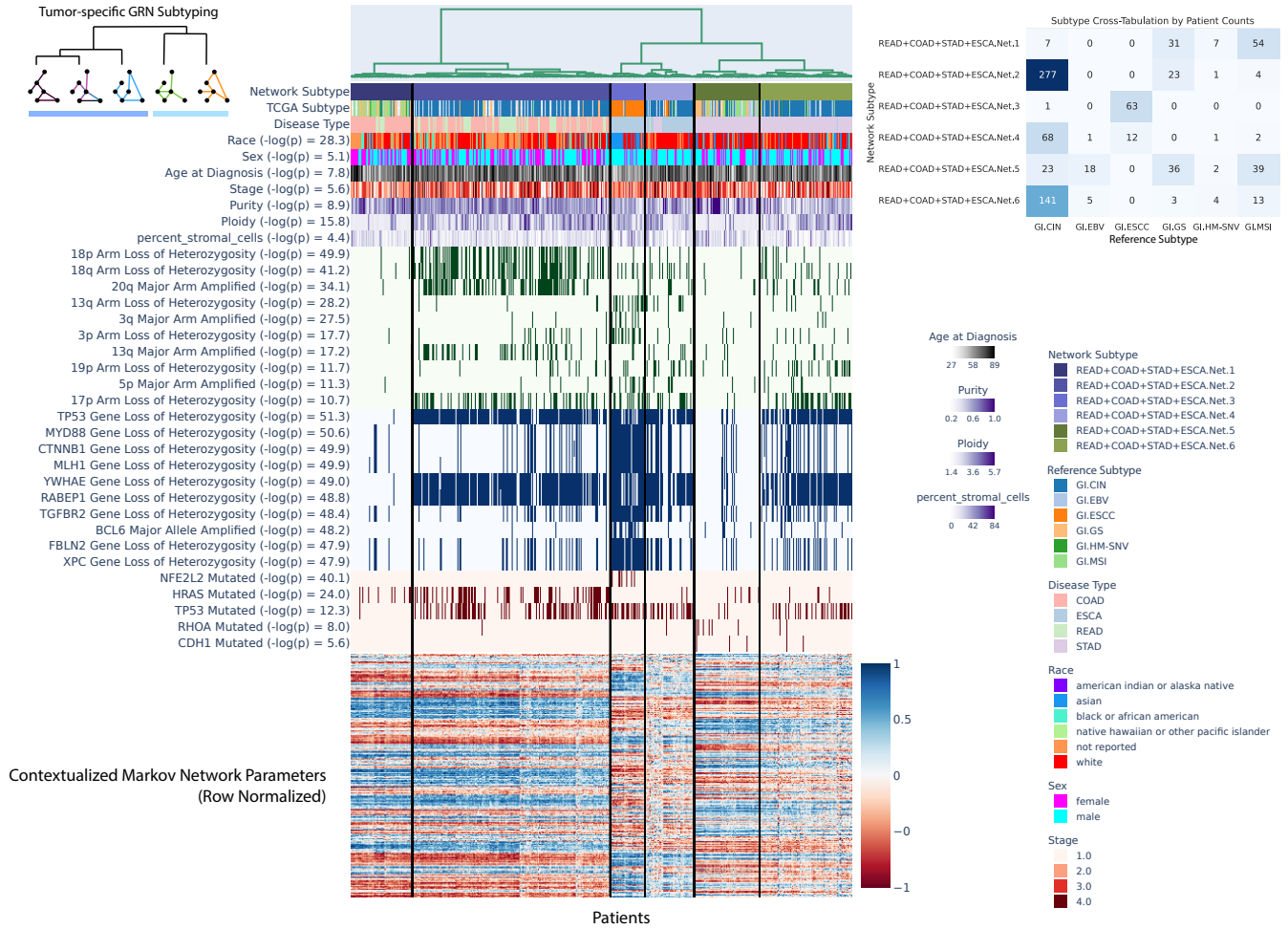
Disease Type	Reference Subtypes	Expression Subtypes	Network Subtypes
Bladder Urothelial Carcinoma	1.059	0.589	0.949
Breast invasive carcinoma	2.056	1.13	2.542
Cholangiocarcinoma	–	–	–
Colon adenocarcinoma	0.162	0.159	2.026
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	–	–	–
Esophageal carcinoma	0.323	1.598	0.284
Glioblastoma multiforme	1.631	1.509	0.63
Head and Neck squamous cell carcinoma	1.855	0.853	3.307
Kidney Chromophobe	0.715	13.802	0.211
Kidney renal clear cell carcinoma	5.61	5.247	14.82
Kidney renal papillary cell carcinoma	19.696	9.241	18.661
Brain Lower Grade Glioma	36.533	25.894	40.656
Liver hepatocellular carcinoma	0.105	0.619	1.263
Lung adenocarcinoma	1.67	2.29	1.198
Lung squamous cell carcinoma	0.614	0.417	0.625
Ovarian serous cystadenocarcinoma	1.414	1.325	0.287
Pancreatic adenocarcinoma	0.937	1.465	2.18
Prostate adenocarcinoma	–	–	–
Rectum adenocarcinoma	0.431	0.723	0.474
Skin Cutaneous Melanoma	–	–	–
Stomach adenocarcinoma	0.469	1.777	1.428
Thyroid carcinoma	0.837	0.831	3.242
Uterine Corpus Endometrial Carcinoma	5.555	3.319	7.42
Uterine Carcinosarcoma	0.319	0.023	0.048
Uveal Melanoma	5.076	2.61	4.536

Table S2. Minimum pairwise log-rank test comparison across different subtyping methods in terms of $-\log(p\text{-value})$. Only samples shared between all datasets are used to control for power. – indicates no samples are shared, or subtypes do not exist for TCGA.

	Expression	Reference	Network
Bladder	0.680	0.483	0.413
Brain	0.857	0.876	0.871
Breast	0.718	0.597	0.773
Colorectal	0.364	0.429	0.551
Esophagus	0.398	0.532	0.496
Eye	0.500	0.618	0.867
Head and Neck	0.558	0.587	0.572
Kidney	0.717	0.609	0.697
Liver	0.646	0.452	0.588
Lung	0.432	0.558	0.497
Ovary	0.412	0.591	0.536
Pancreas	0.468	0.634	0.542
Thyroid	0.475	0.413	0.580
Skin	0.424	0.436	0.333
Stomach	0.620	0.545	0.520
Uterus	0.767	0.742	0.632

Table S3. Concordance index of survival regressors for each tissue trained on different patient representations.

(a) Network Clusters Reveal Prognostic Biomarkers in Molecular Contexts Across Tissues



(b) Clusters of Patient-specific Transcriptomic Networks Reveal State-of-the-art Prognostic Cross-disease Subtypes

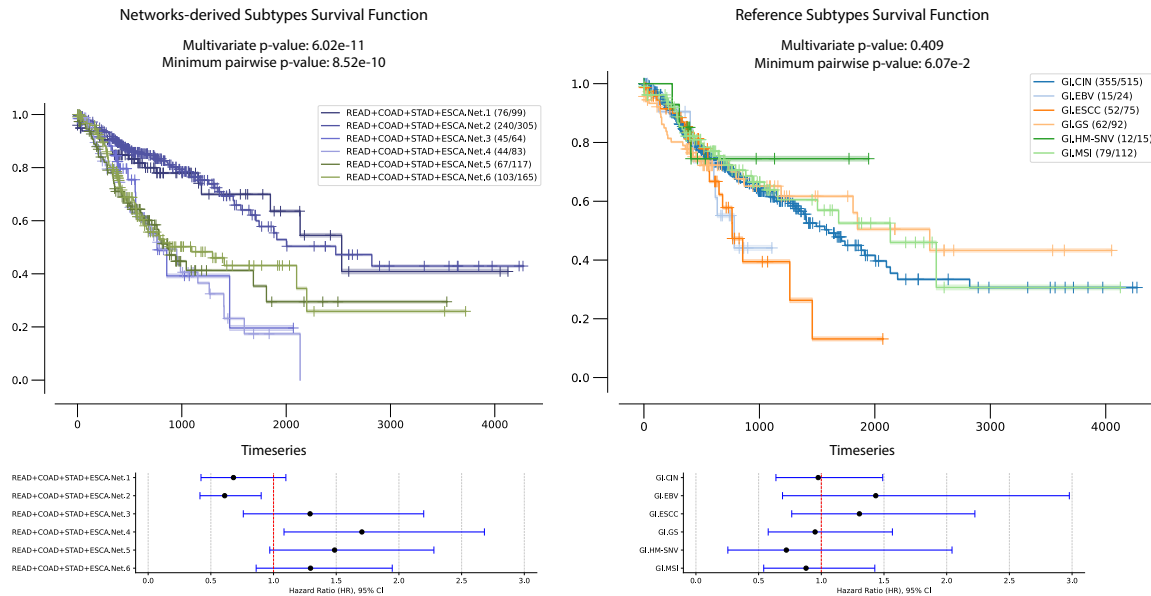
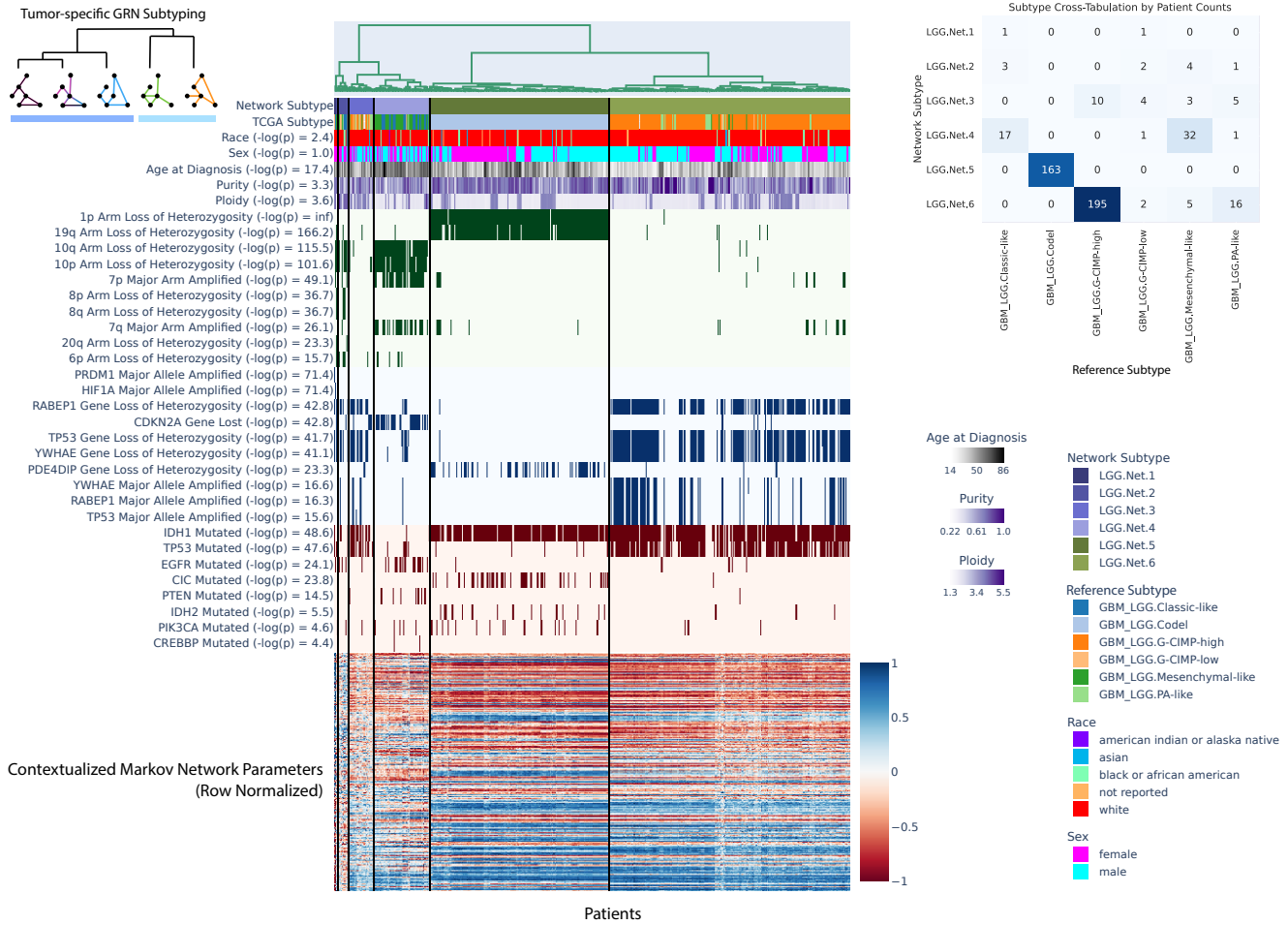


Fig. S5. Exploration of cross-disease network subtypes for cancers of the GI tract, including rectum adenocarcinoma, colon adenocarcinoma, stomach adenocarcinoma, and esophageal carcinoma, looking at correlated clinical information, arm-level copy alterations, gene-level copy alterations, and gene-level single nucleotide variations. Reference subtypes from (17).

(a) Network Clusters Reveal Prognostic Biomarkers in Molecular Contexts



(b) Clusters of Patient-specific Transcriptomic Networks Reveal State-of-the-art Prognostic Subtypes

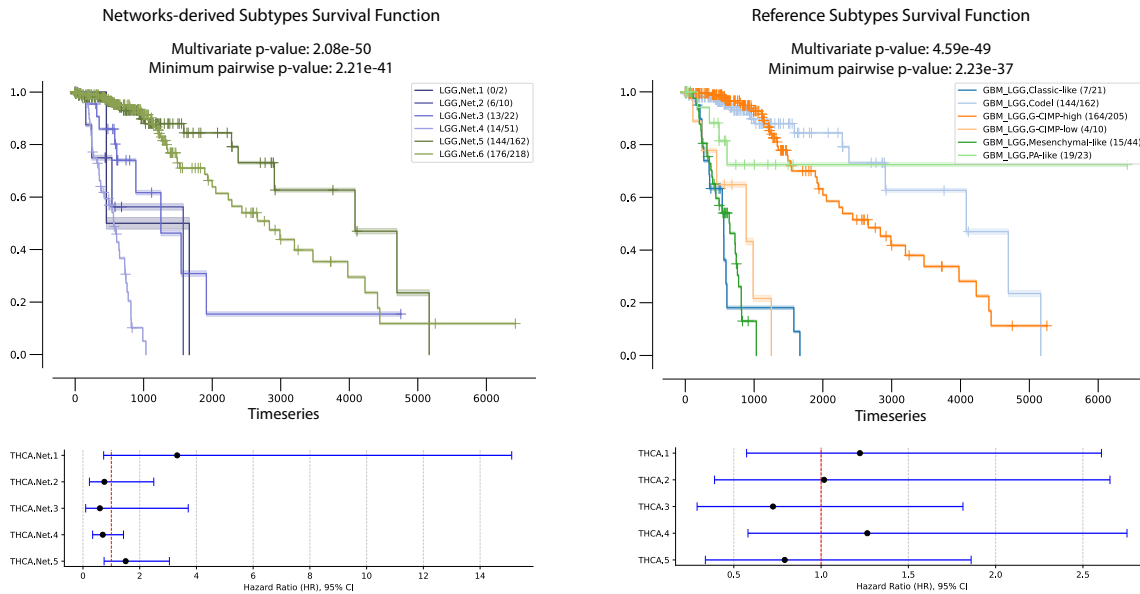


Fig. S6. Exploration of Brain Lower Grade Glioma, looking at correlated clinical information, arm-level copy alterations, gene-level copy alterations, and gene-level single nucleotide variations. Reference subtypes from (18).

101 **SI Dataset S1 (data.tar.gz)**

102 Pre-processed de-identified data from TCGA. <https://zenodo.org/records/14885352/files/data.tar.gz>

103 **SI Dataset S2 (results.tar.gz)**

104 Pre-trained contextualized networks (correlation, Markov, neighborhood) for all 7997 patients with train/test split labels.
105 Includes subtyping plots for all 25 disease types with contextualized Markov networks. [https://zenodo.org/records/14885352/files/](https://zenodo.org/records/14885352/files/results.tar.gz)
106 [results.tar.gz](https://zenodo.org/records/14885352/files/results.tar.gz)

107 **References**

- 108 1. CN Ellington, et al., Contextualized: Heterogeneous Modeling Toolbox. *J. Open Source Softw.* **9**, 6469 (2024).
- 109 2. R Agarwal, et al., Neural Additive Models: Interpretable Machine Learning with Neural Nets. (2021).
- 110 3. BJ Lengerich, ME Nunnally, Y Aphinyanaphongs, C Ellington, R Caruana, Automated Interpretable Discovery of
111 Heterogeneous Treatment Effectiveness: A COVID-19 Case Study. *J. Biomed. Inform.* p. 104086 (2022).
- 112 4. BJ Lengerich, et al., Discriminative Subtyping of Lung Cancers from Histopathology Images via Contextual Deep Learning.
113 *medRxiv* (2020) Publisher: Cold Spring Harbor Laboratory Press.
- 114 5. J Deuschel, et al., Contextualized Policy Recovery: Modeling and Interpreting Medical Decisions with Adaptive Imitation
115 Learning (2023) arXiv:2310.07918 [cs, stat].
- 116 6. TI Zack, et al., Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
- 117 7. C Calabrese, et al., Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020) Number: 7793 Publisher:
118 Nature Publishing Group.
- 119 8. P Van Loo, et al., Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107**, 16910–16915 (2010) Publisher:
120 Proceedings of the National Academy of Sciences.
- 121 9. RM Drews, et al., A pan-cancer compendium of chromosomal instability. *Nature* **606**, 976–983 (2022) Number: 7916
122 Publisher: Nature Publishing Group.
- 123 10. C Tokheim, R Karchin, CHASmplus Reveals the Scope of Somatic Missense Mutations Driving Human Cancers. *Cell*
124 *Syst.* **9**, 9–23.e8 (2019).
- 125 11. CD Steele, et al., Signatures of copy number alterations in human cancer. *Nature* **606**, 984–991 (2022) Number: 7916
126 Publisher: Nature Publishing Group.
- 127 12. Z Sondka, et al., The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev.*
128 *Cancer* **18**, 696–705 (2018) Number: 11 Publisher: Nature Publishing Group.
- 129 13. A Liberzon, et al., The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425
130 (2015).
- 131 14. E Kim, V Gheorge, T Hart, Dynamic rewiring of biological activity across genotype and lineage revealed by context-
132 dependent functional interactions. Publication Title: bioRxiv (2021).
- 133 15. S Mohammadi, J Davila-Velderrain, M Kellis, Reconstruction of Cell-type-Specific Interactomes at Single-Cell Resolution.
134 *Cell Syst* **9**, 559–568.e4 (2019).
- 135 16. M Stone, et al., Identifying strengths and weaknesses of methods for computational network inference from single cell
136 RNA-seq data. Publication Title: bioRxiv (2021).
- 137 17. Y Liu, et al., Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer cell* **33**, 721–735.e8 (2018).
- 138 18. M Ceccarelli, et al., Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse
139 Glioma. *Cell* **164**, 550–563 (2016).